Ontology Mapping for Geoinformation Integration

Koros Wesley Kipngénoh

February, 2010

Ontology Mapping for Geoinformation Integration

by

Koros Wesley Kipngénoh

Thesis submitted to the International Institute for Geo-information Science and Earth Observation in partial fulfilment of the requirements for the degree in Master of Science in *Geoinformatics*.

Degree Assessment Board

Thesis advisor	Dr. Ir. R. L. G. Lemmens Ms. Dr. I. Ivanova
Thesis examiners	Chair: Dr. Ir. R. A. de By External examiner: Dipl. Geoinf. C. Keßler

INTERNATIONAL INSTITUTE FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

ENSCHEDE, THE NETHERLANDS

Disclaimer

Acknowledgements

This thesis would have become impossible without the support of several people and organizations. I reciprocate my immeasurable gratitudes to all who contributed.

In particular, I thank my advisor Dr. Ir. R.L.G. Lemmens for his help and guidance during the last six months. I thank him for his patience and for sharing his experiences. He provided interesting and stimulating research environment that was critical for the progress of this thesis. I am grateful too to my second advisor, Ms. Dr. I. Ivánová for her valuable input, her patience and effort. Thanks also to Dr. Raul Zurita-Milla for his advise.

I want to express my thanks to Dr. Krzysztof Janowicz from Institute for Geoinformatics, University of Muenster, Germany, for his contribution and timely response to questions, advise towards the technical operation of SIM-DL, research papers and proceedings that provided the basis for chapters 2 and 4 of the thesis. Many thanks to Dipl. Geoinf. Carsten Keßler from the Institute for Geoinformatics, University of Muenster, Germany, for his input and confirmation of some ideas about the sensitivity of mappings and the similarity measurement using SIM-DL. His contribution formed part of chapter 4. I recognize Natasha Noy, who is a Senior research scientist, Stanford Center for Biomedical Informatics Research (BMIR), for her advise and response to questions and research papers. Her ideas formed the base for chapter 2. The pieces of advise about similarity matching tools by Parvel Shvaiko of University of Trento, Italy and Jérôme Euzenat of INRIA Grenoble Rhône-Alpes and LIG, France. Many gratitudes to Mr. Craig A. Smith, Software Licensing Manager, Sandia National Laboratories, for providing me with license for JESS, the Rule Engine which formed part of chapter 5. Appreciation to the entire team of ORNL DAAC,USA and Wageningen University for providing UMD and IGBP and LGN 4 datasets respectively. Credit to Racer Systems GmbH and Co.KG, Hamburg, Germany for allowing me to use RacerPro, one of the reasoners. I recognize also our IT department in particular Mr. Reinink for his help.

I am grateful too to all colleagues at the Department of Geoinformatics, in particular Mr. Dongpo Deng for fruitful discussions and encouragement. Thanks to my mom, brothers(Andrew and Christopher) and Ruth for their support and encouragement.

This thesis would have been impossible too without the support of the NUFFIC scholarship and the Dutch Government. It provided the financial support for the entire study, the thesis phase and contributed to the interesting and stimu-

lating research environment.

Abstract

Geoinformation integration provides a basis for manipulation of geospatial datasets across various sources of geodata. It is suitable for unlimited number of applications in the geoinformation domain. The use of conceptual representation of geospatial data and their respective relationships to establish the similarity of concepts and instances in heterogeneous sources, forms part of semantic geoinformation integration. The concept of ontology is the base for semantic data integration. Semantic data in the geoinformation domain is growing. The task of finding a concept in dataset A that correspond to a concept in dataset B, requires mappings between the concepts. This task is gaining importance. Given that semantic datasets are structured using different ontologies, (semi)-automatic ontology mapping techniques need to be utilized before geodata integration and retrieval.

Given two ontologies from two different semantic data sources, one must be in a position to tell whether they model the same real world phenomena and measure the degree of semantic similarity.

A similarity measurement reasoner is used in the establishment of the links between the compared concepts. The characteristics defining the compared concepts are used by the reasoner in establishing whether they correspond to the same entity. Mappings are defined for comparing two instances and the relations that hold between the compared instances and the compared concepts. These relations between compared instances and concepts are defined using very expressive rules. Similarity measure and the defined threshold are the platforms for defining the mapping relations. The realized mappings are used in an application such as geoinformation retrieval from the knowledge base. This application requires the use of datatype properties and very expressive rules. The whole process from the ontology development, similarity measurement, definition of the mapping relations and utilizing them in an application are discussed in this thesis research.

The ontology mapping task requires knowledge modelling using ontologies and reasoning techniques. It also depends on the application context. Thus ensuring deduction of knowledge from large and ambiguous domain specific semantic information sources. Similarity measurement technique can be used to determine which concept and instances represent similar notion. The datatype properties and very expressive rules can be used to integrate and retrieve the information from the knowledge base.

Keywords

Alignment, GIS, Integration, Land cover, Ontology, Ontology mapping, Ontology matching, Rules, Semantic, Similarity

Contents

List of Tables

List of Figures

Acronyms

API Application Programming Interface

CORINE Coordination of Information on the Environment

DIG Description Interface Group

DL Description Logics

DOLCE Descriptive Ontology for linguistic and Cognitive Engineering

GPL General Public License (GNU)

GIS Geographical Information System

HarmonISA Harmonization of Land-use Data

IGBP Global International-Biosphere Program

INSPIRE Infrastructure for Spatial Information in Europe

JESS Java Expert System Shell

KSL Knowledge Systems Laboratory

LGN National Land cover/ use Database of Netherlands

LISP List Processing (Programming language)

MIT Massachusetts Institute of Technology

MODIS Moderate Resolution Imaging Spectro-radiometer

OWL-DL Web Ontology Language-Description Logic

OWL Web Ontology Language

RACER Renamed ABox and Concept Expression Reasoner

RDF Resource Description Framework

SDI Spatial Data Infrastructure

SMI Stanford Medical Informatics

SPARQL Simple Protocol and (Resource Description Framework) Query Language

SPARQL-DL Simple Protocol and (Resource Description Framework) Query Language-DL

SQWRL Semantic Query Enhanced Web Rule Language

SUMO Suggested Upper Merged Ontology

SWRL Semantic Web Rule Language

UMD University of Maryland

UPM Universidad Politécnica de Madrid

W3C World Wide Web Consortium

Chapter 1

Introduction

The current advances in spatial information technology and increasing multiple sources of spatial data demand for spatial data integration. A dataset is produced based on some individual or community specifications. These specifications include the definition of spatial objects, multiple representations and varying scale. In addition, the temporal aspect in which the dataset is collected demands for this integration. This complicates spatial data integration, reuse and sharing. Heterogeneity depends on varying perceptions and conceptualizations of geographical phenomenon. This conceptualization and specification of heterogeneous concepts is called ontology [Gru93]. These ontologies are built for different application domains such as land use planning and land cover classification. Locating and retrieving desired geoinformation to meet the increasing demands of end-users is difficult. These are clearly apparent in the unfolding and distributive SDI environments. Existing SDI standards fail to resolve semantic problems that occur due to heterogeneous geodata content and diverse user communities [LSK+07]. Standards for semantics support unambiguous understanding and use of geoinformation. These geoinformation are collected by different organizations at different time periods and for different purposes [SLKB06].

1.1 Motivation

Many geospatial information communities have different views of the real-world spatial phenomena thus disparate perception of the contexts. For instance, given two ontologies from Overijssel datasets in Netherlands, find 'high density residential' land? But there could be only 'high intensity residential' land in Enschede dataset and 'residential' land in Hengelo datasets. This problem is due to semantic difference in the two datasets. This is because of different ways of the phenomenon interpretations,thus different definitions. Several geospatial ontologies have been developed by different individuals and communities. These geospatial ontologies could be different due to variation in methods of data collection, applications, scope and even time. The dispersive spatial datasets over similar domain results in different ontologies. The heterogeneities could be as follows:

- semantic heterogeneity,
- semantic constraint,
- difference in coverage mismatch in the part of the domain covered by the ontology,
- granularity:mismatch in the level of detail to which domain is modelled and
- perspectives:ontologies describing the same region and same level of detail, but different perspective;

These lead to no common understanding of concepts, thus causing difficulties in integration. Therefore, establishment of relations between the concepts and instances is necessary. Increasing applications demand the use of multiple spatial datasets from different sources. For instance, a community INSPIRE may intend to use both community standard and community-specific geospatial ontologies. An example is in the HarmonISA project by [Hal06] where land use data from border region of Austria, Slovenia and Italy have been developed based on different national and European CORINE land use classification systems [OZ08].

Motivational example

With two different geospatial datasets (A and B) of the same domain but different sources. How can the mappings between concepts be established? For instance, finding a concept definition in dataset A that corresponds to the concept definition in dataset B as in figure 1.1.

1.2 Problem definition

Applications such as determining land cover/use types with geospatial distribution can demand the use of incongruous geospatial datasets from different communities or from other diverse views on a specific locality. Further, geospatial ontology developers as third party may intend to use existing geospatial ontologies. These act as a base for the creation of new geospatial ontologies by combining knowledge from different smaller geospatial ontologies. Sharing of rich geospatial ontologies allow the community to utilize the already existing geoinformation. These culminate to saving resources, time and avoid duplication efforts in terms of related or common geoinformation. Harmonizing or integrating datasets from different classification systems requires a mapping between them [OZ08]. Due to intricate and compounded nature of geospatial ontologies, integrating them is difficult. For existence of a common understanding amongst communities, heterogeneous geospatial ontologies have to be brought to a 'mutual status'. This puzzle in literature is *ontology mapping*. This involves defining semantic relations between two concepts from two geospatial ontologies. This is meant for transforming instances from source to target geospatial

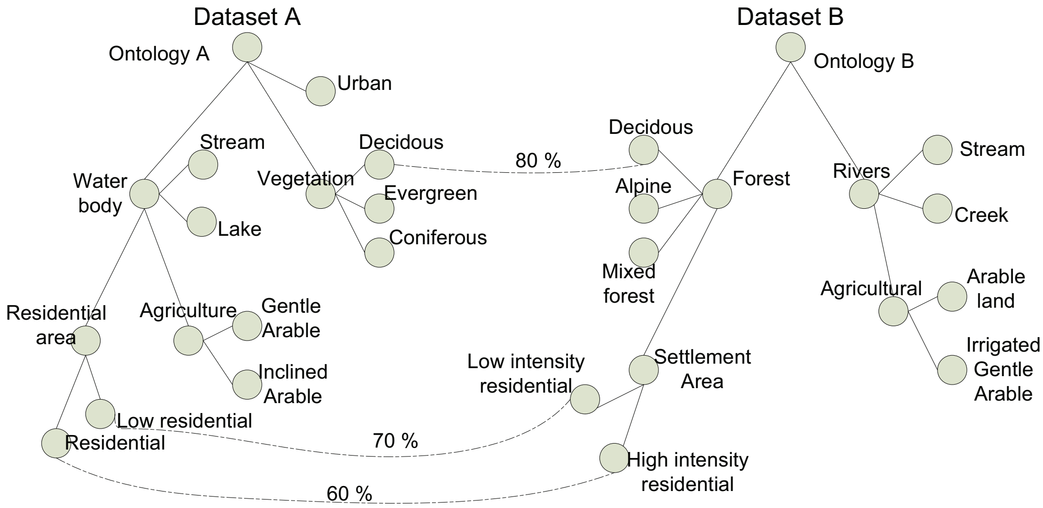

Figure 1.1: Motivational problem statement

ontology leading to integration of cross-over knowledge. The specific problem is how to realize these mappings. What exactly is it for? How useful can it be in an application context? This allows pragmatic interoperation across multiple local data sources and access to up-to-date data [ES07]. Integration of geo-datasets improve the dataset quality, enable interdisciplinary analysis, yields new knowledge, mutual enrichment and benefit and improvement of information sharing and reuse [Kie08].

Ontology mapping is classified into globally integrated and local ontology mappings. Also mapping of local ontologies and mapping that constitutes ontology merging, integration and alignment. The manner by which ontologies are built and maintained is an exceptional important controversy among three categories. For ontology integration, a single ontology in one subject is generated from two or more existing and different ontologies in different subjects [CSH06]. In *ontology merging*, a new ontology is created from two or more ontologies resulting in unification and replacement of original ontologies. For *ontology alignment*, it involves finding corresponding entity in one ontology which has the same intended meaning in the second ontology [Ehr07]. Mapping between different land cover classification systems are in two broad groups: *ontology-based approaches* where concepts define the classes in the classification systems to be mapped. In this approach, a general upper ontology is commonly agreed upon by developers of different applications. These developers extend this general ontology with concepts and properties specific to their applications. The other approach is *similarity-based approaches*. This involves the computation of numeric similarity values between two concepts thus expressing differences and commonality between them [AFCLL08].

1.3 Research Objectives

1.3.1 Main objective

The main target of this research study was to develop and evaluate methods of ontology mapping for geoinformation integration.

1.3.2 Specific objectives

1. To evaluate the methods of ontology mapping such as (similarity measurement and ontology-based) in knowledge structures for purposes of integrating geoinformation sources.

2. To exploit ontology mapping within a (web) application as a proof of concept for geoinformation integration and

3. To find how sensitive the mapping is in the event of manipulation(updates).

1.4 Research questions

1. Given two geospatial ontologies, how can similarities between them be established? How do we determine which concept and properties represent similar notions?

2. How can the mapping be defined for transforming source instances into the most similar target instances? How can the relation between instances be represented and used in practical application?

3. How do the similarity measurement methods specify the representation of mapping between the spatial entities?

4. How can the resulting mapping be put into application and maintained?

5. How can the ontology mapping be exploited to an [web] application e.g.,for posing queries across heterogeneous datasets.

1.5 Related work

A lot of research concerning semantic mapping has been done and is on-going both in non-geographical applications and few in geographical applications.

These non-geographical application disciplines include medicine and bioinformatics. These has resulted into development of ontology mapping approaches such as COMA++, MAFRA, PROMPT, GLUE to name a few [Ehr07]. For geo-applications, SIM-DL server has been developed and is still on-going [JRSK08]. Semantic data integration of geographical ontologies is well focused by [Kok06]. Classification of the existing integration approaches with emphasis on the semantics of the compared ontologies is also explained. In order to support automatic generalization processes, [SLKB06] gave a clear description of the semantic aspects that need to be considered. It discusses about semantic data integration in different scales and their maintenance in different environments. This includes links between multi-represented objects and the support of update propagation. A broad scope of ontology mapping tools, mapping categories and characteristics and survey on ontology mapping tools is depicted. In addition, different roles of ontology mapping categories are identified in [CSH06].

The analysis of spatial and geometrical characteristics of instances of datasets can be used to reveal semantic correspondences between classes of objects from different geo-ontologies. This is an approach for realizing the integration of data sets that are of different origins and with different resolution levels. The idea was to derive transformation rules with data mining methods which allow semantic connection between datasets [Kie08]. This approach was taken because of the absence of expert knowledge.

In [Sch08], it is postulated that similarity is essential for dealing with vague data queries, vague concepts. It is basis for semantic interoperability and integration. Existing similarity measures are classified as *geometric, feature, network, alignment and transformational models*. It is claimed that this was done according to knowledge representation and notion of similarity. In the book [OZ08], it is stated that many approaches to map different classification systems exist. These approaches can be classified in two groups: *ontology-based* and *similarity-based*. Similarity-based approach improves ontology mapping and geodata integration. It also supports geoinformation retrieval, browsing and knowledge acquisition. The preceding approach is considered by [LSK+07] and derived a taxonomy between concepts. In the latter approach, [Hal06] uses it to combine land use data. The core characteristics of semantic similarity and the roles of semantic similarity measurement for GIScience is well shown in [JRSK08]. In this research study, both approaches were considered for evaluation. This was for purposes of integrating the cross-over knowledge in application-specific ontologies..

1.6 Thesis Outline

This research study consists of six chapters. It starts by presenting an overview of geo-semantic modelling and mapping in chapter two. This chapter gives an understanding of ontology and geo-ontology. Languages that help in formalizing the concepts and retrieval of information from the knowledge model, ontology editing and mapping tools and Descriptive Logic reasoners are given too. Next is a presentation about the use case, semantic definition extraction and modelling of real-world contexts in an ontology. Chapter four, is about the semantic similarity measurement using SIM-DL and validation of the outcome of similarity measurements. In addition, representation of the mapping and sensitivity of the mappings is part of this chapter. Thereafter, is the integration and retrieving of information from the ontology using rules. An application example is presented. Finally, a discussion is made about the outcome of semantic similarity measurement and the validation and how this approach is useful in assessing the similarity of datasets for geoinformation integration.

1.7 Terminology

It is not surprising that some terms are used inconsistently about ontology mapping and even ontology itself. The terminology and definition of the terms used in this research study are clarified . This is to make some decisions about our understanding of the terminology. This is because there is not always an agreement on the exact meaning of the terms. In this research, the following definitions of the terms are adopted. A summary of these terms are in table A.1.

- *Ontology mapping*: This is the process where two ontologies are semantically related at conceptual and instance level. The source ontology instances are transformed into target ontology entities according to the semantic relations *see figure*1.2. The term 'mapping' is a function between the signatures. It entails finding corresponding concepts (mappings) between two ontologies. If any two concepts correspond, they have same meaning and refer to the same thing or closely related things. The mappings are expressed by some mapping rules that explain how the concepts correspond [DMQ05]. In addition, a change in ontology mapping for heterogeneous resolution is not explicit. However, ontology mapping is performed at run-time (during interoperability and communication). Resolving heterogeneity problems between ontologies allows interoperability.
- *Ontology alignment*: It is a set of correspondences between two or more (in case of multiple matching) ontologies. It is thus the output of the matching process. It involves finding corresponding entity in one ontology which has the same intended meaning in the second ontology [Ehr07, ES07].
- *Ontology matching*: This involves finding relationships or correspondences between entities of different ontologies [ES07]. Ontology matching is a promising solution to the semantic heterogeneity problem. It finds correspondences between semantically related entities of the ontologies. This task involves creation of mapping rules or alignments. The correspondences can be applicable for a number of tasks, such as merging ontologies, answering query, translation of data or for navigation on the semantic web. Matching ontologies makes it possible for the knowledge and data expressed in the matched ontologies to interoperate.
- *Ontology merging* Creation of a new ontology from two or more source ontologies possibly with overlapping concepts or definitions. The initial ontologies remain unaltered. The merged ontology contains the knowledge

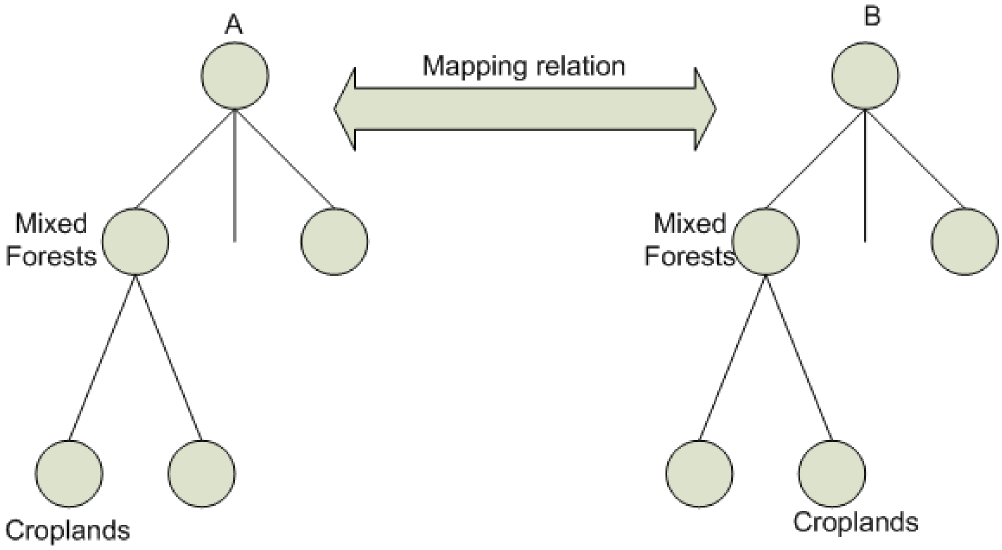

Figure 1.2: *Ontology mapping.* A set of mapping rules are produced to relate, for example, $ontologyA$ and $ontologyB$. The type of mapping rules produced determines the type of the algorithm (mapping, alignment or matching).

of the initial ontologies [ES07] *see figure* 1.3. Very similar to ontology integration. The only difference is in the domain of the sources. It is mainly applied when fusing knowledge from different sources in order to describe the domain more accurately. It is considered useful when a number of existing ontologies each partially describing the same domain. Heterogeneity resolution being a major part of the task of merging (but ontology merging is more than that), consists of obtaining a new ontology from the two matched ontologies. The matched entities in source and target ontologies are related as prescribed by the alignment. Expressing ontologies in the same language, merging often involves putting the ontologies together and bridge or articulation axioms are generated. The entities in the source ontology which have no corresponding entity in the target ontology remains unaltered in the merged ontology. Ontology merging is especially used when needs arise to carry out reasoning that involves several ontologies. It is also used when editing ontologies in order to create ontologies tailored for a particular application. It is stated in [ES07] that it is mostly followed by a phase of ontology re-engineering for instance suppressing unwanted parts from the obtained ontology.

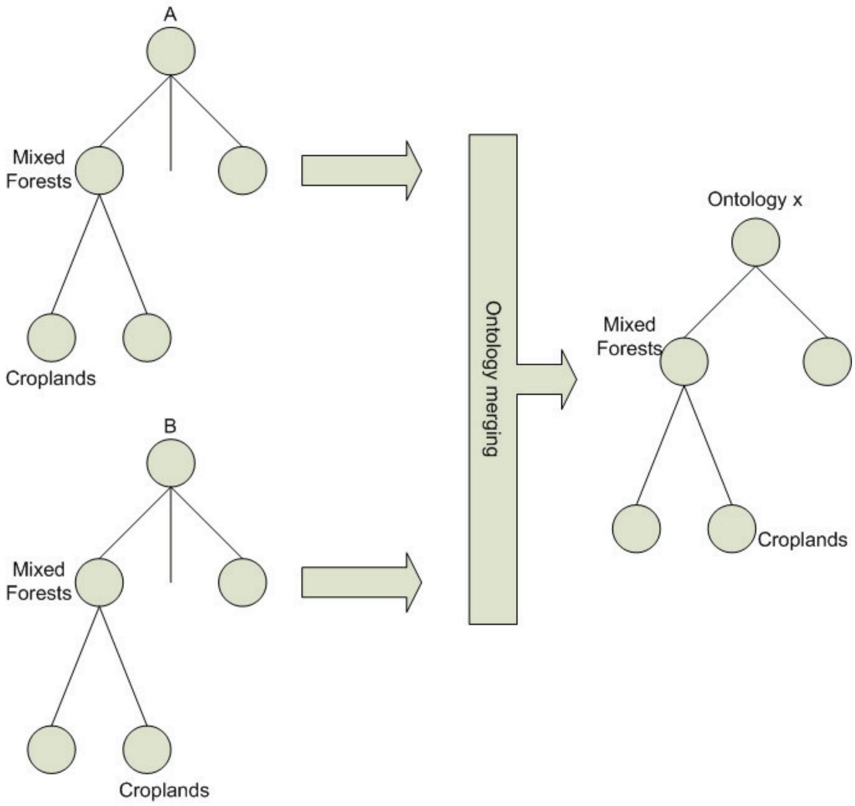

Figure 1.3: *Ontology merging.* $OntologyA$ is combined with $ontologyB$. Ontologies cover identical domains. Information from the source ontology is intermingled (not-identifiable) in the result $(ontologyX)$.

- *Ontology integration*: This is the inclusion in one ontology of another ontology. Assertions expressing the link between these ontologies usually are bridge axioms. The first ontology is unaltered while the second ontology is modified [ES07] *see figure* 1.4. It is very similar to ontology merging except in the domain of the sources (covering similar domains i.e loosely related domains). It is mainly applied when focus is on fusing knowledge from different sources in order to cover a broader domain. Ontology integration is very useful in ontology development (integrating independent sub-ontologies makes ontology design more efficient). Heterogeneity resolution is a major part of integration task (but ontology integration is more than that). The result contains each of the sources in loosely related (and easily identifiable) modules.

- *Bridge axioms*: These are formulas expressed in an ontology language. They express the alignments such that it is possible to carry out integration of the entities in an ontology together with another [ES07].

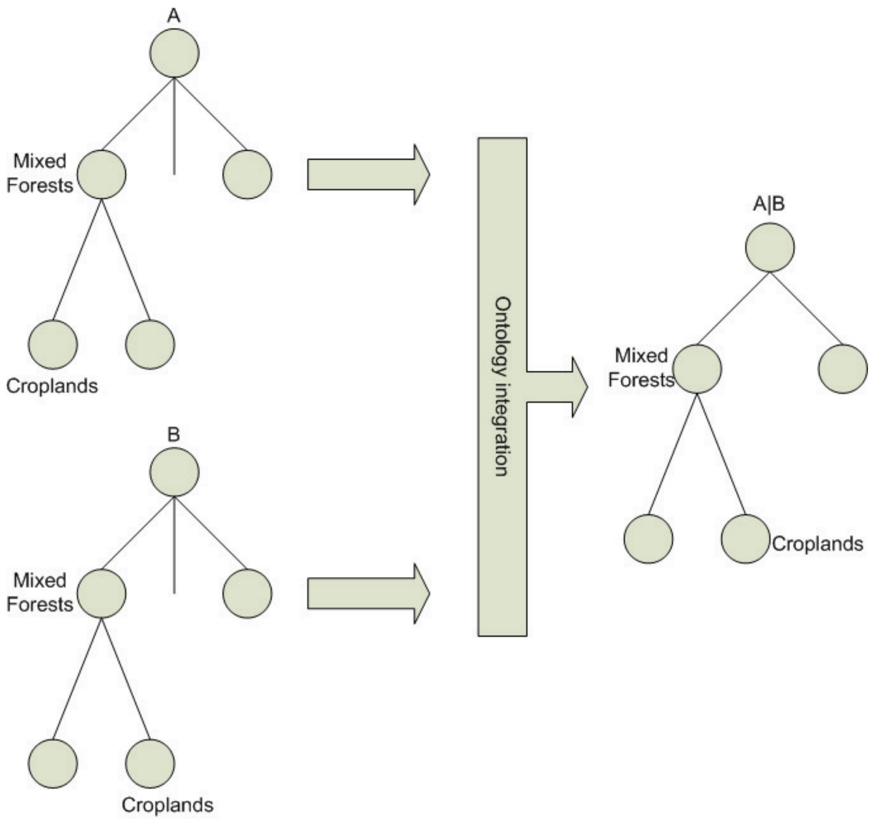

Figure 1.4: *Ontology integration.* $ontologyA$ is integrated with $ontologyB$. The ontologies cover similar domain. The result $(A|Bontology)$ contains each of the sources in loosely related (and easily identifiable) modules.

- *Ontology similarity* This refers to the comparison of whole ontologies or sub-elements thereof. This comparison returns a numerical value indicating whether the two ontologies have a high or low degree of similarity.

1.8 Summary

The research study has been introduced in this chapter. The motivating problem and approach together with the terms that are used throughout this research are presented. The following chapter gives the theory of the semantic modeling and mapping.

Chapter 2

Geo-semantic Modelling and Mapping

The theory of geo-semantic modelling is a fundamental step in geoinformation integration. It is a keystone for understanding the implicit meaning of the geodata. Geo-semantic modelling addresses the context of the meaning attached to data elements and how they relate to each other. Variety of the needs in geoinformation knowledge modelling, integration, knowledge management and knowledge re-use can be satisfied. Experts within geoinformation industry and academia, have been developing and deploying sharable and reusable models known as ontologies.

2.1 What is ontology?

As stated in chapter 1, ontology is defined as a 'formal, explicit specification of a shared conceptualization' [Gru93] . Ontologies represent a good number of things e.g.,trees, land, rivers in a given area of interest. These things are represented in ontology as *concepts* (classes) which are arranged typically in a hierarchy of classes and subclasses. A class is associated with various *properties* (slots or roles) that describe its features and attributes as well as restrictions on them (facets or role restrictions) [NM01]. [UG04] puts it that an ontology with a set of *instances* (individuals) makes a *knowledge base*. The manner in which the meaning of a concept is specified is what distinguishes different approaches to ontologies.

2.2 Understanding geo-ontologies

Time is ripe for the development of a geographic ontology that is comprehensive. This is for the domains of scientific research which includes or extending over geographic space. Much is now known about ontology in general and about its role in description, thought, language, geodata system interoperability and research. However, only a few decades ago, serious research on ontology of geographic phenomena has begun [SM98]. Further, this work in geographic ontology focuses on form. It either addresses specific kinds of geographic phenomena such as fields [PSB99]. It relates to naïve or common-sense geography [EM95] and to general principles. Geographic ontology of scientific research domains has received less explicit attention. An Ontology of the geospatial domain deals with the entirety of geospatial concepts, categories, relations, processes and with their inter-relations at different resolutions [MEH04].

2.2.1 Components of geographic ontologies

A geographic ontology defines geographic objects, fields, spatial relations, processes and their categories [MEH04]. It consists of the basic data models, concepts and representations for scientific computing about geographic phenomena. It includes also the ontological principles and structures to be supported in geographic space.

For geographic objects, as [SM98] have noted, there arises an issue of individuation. If topography is a continuous field of elevation, how are such elevation fields parsed into objects such as land cover, mountains, valleys, hills, and ridges? If a mountain is an object, then it is likely that it has indistinct or indeterminate boundaries. Any ontology of geospatial objects needs to deal with boundaries of this kind [BF96]. Other geographic objects have bona fide or genuine boundaries e.g., islands as a land cover in this scenario. Contrary, some geospatial objects have fiat boundaries. Fiat objects are all non-naturally and demarcated geographical entities. They are created by legislative acts or other decisions e.g.,land parcels or administrative. Objects may be considered to be bona fide or fiat, depending on the kind of their boundaries. Another aspect of the ontology of geographic objects involves categorizing them. Because individuation and classification are not always independent. Geographic object categories are more likely to depict individual, cultural or disciplinary differences than are table-top objects [SM01].

Not all geographic phenomena are conceptualized as objects. Fields may be defined simply as functions that map from position in space onto some measurement scale, including nominal scale. An ontology for common-sense geographic phenomena might be able to ignore geographic fields. However, fields are critical to scientific applications having geospatial dimensions. This is especially for fluid or soil or mineral geographic domains. A complete geographic ontology also provides definitions of spatial relations of dynamic aspects of geospatial phenomena (events, change, motion, etc.) and of more complex geographic processes.

2.2.2 What is special about geo-ontologies?

As stated by [FCM06], a geo-ontology has to provide a description of geographical entities, which can be conceived in two varied views of the world. These views are *field view* which considers spatial data to be a set of continuous distributions. For the *object view*, the world is conceived to be occupied by discrete, identifiable entities. Representation of geographic entities is complex. A geo-ontology is unique from other ontologies. This is because topology and part-whole relations play a major role in the geographic domain. Geographic objects

Table 2.1: Basic components of geo-ontology
(adapted from [FCM06])

Components of geo-ontology		
	Physical reality	**Social reality**
Bounded	Bona fide objects(e.g mountain)	Fiat objects (e.g parcel).
Continuous	Physical fields(e.g temperature)	Social distributions (e.g human development)

can be contiguous or connected, separated or scattered, closed or open. They are typically complex and have constituent parts [SM98]. Geographic entities are tied intrinsically to space and not just located in space [SM98]. It takes from space some of its structural characteristics, such as mereological, topological and geometrical properties.

Geo-ontology has two types of concepts: concepts corresponding to physical phenomena in real-world and concepts that correspond to features of the world that are created to represent social constructs. The first type of concepts is *physical concepts* and the other is *social concepts* [FCM06] *see table* 2.1.

In this research study, the focus is on mapping definitions of concepts between ontologies. Concepts are considered typically to be part of geospatial databases.

2.3 Why do we develop ontologies?

The novel focus of this research study is on semantics of geospatial information. More generally, it is on the relations between human minds and geoinformation systems about the geospatial phenomenon. The semantics of geospatial information is critical for the interoperability of geospatial data and software. It is also important that GIS software and technology be able to inter-operate with other softwares and spatial databases. Interoperability in geographical applications requires ontology for the geographic phenomena under consideration -any phenomena distributed over part or all of the Earths surface [MEH04].

The reasons why would one want to develop ontology are as given in [NM01] and they include:

- *For sharing a common understanding of the structural information among people or software agents*. For example, what if different Web sites contain geoinformation of a particular company. These Web sites share and publish the same underlying ontology of the geoinformation. The user can extract and aggregate these geoinformation from these sites. The aggregated geoinformation can then be used in answering respective queries. More so, use them as input geodata to other geographical applications.

- *For reuse of domain knowledge*. An organization dealing with geoinformation develops an ontology in detail of a particular domain. Other organizations or individuals interested in that particular domain can reuse the ontology. Besides, the ontology can be extended to describe the domain of interest in case a need for a larger ontology arises.

- *For making domain assumptions explicit*. Assumption within geographical application domains are explicitly made in case of changes of knowledge in that particular geographic domain. This can as well be useful to naïve users who should learn the meaning of concepts in the given geographical domain.
- *For separation of application specific knowledge from operational knowledge*. A task of configuring a geographical object from its components according to given specifications can be described. Implementation of a program that does this configuration independent of the geographical concept and components themselves. An ontology of geographical components and characteristics can be developed and the algorithm applied to carry out configuration.
- *For analyzing domain knowledge*. If the given declarative specification of concepts is available, the analysis of domain knowledge is possible. Analysis of concepts is valuable when reusing the existing ontologies and even in extending them.

2.4 Formalizing an ontology

Despite an ontology in principle being independent of a particular language, it is necessary to choose a language to describe it. In order to share, exchange and map ontologies, the language must be formal. Natural language alone is insufficient. This is because much of interpretation is left for the user, thereby leading to potentially missing significant aspects of an ontology. The phenomena behind geographic ontologies are their complexity in nature. Therefore, sophisticated knowledge representation methods are needed to abstract or represent them appropriately.

An ontology is formalized through definition of *classes*, *relations*, *functions* and *axioms* (e.g., see [UG96] for examples of the languages and tools that have been defined in computer science for developing ontologies). This basic fourfold structure defines an ontology and underlies all of this research study.

2.4.1 Species of Web Ontology Language (OWL)

Three different OWL sub-languages have been defined by the W3C's. Each sub-language is geared towards fulfilling different aspects of the above named incompatible full set of requirements. They include OWL Full, OWL-DL and OWL-Lite. In order to adopt the OWL language in the development of ontology, there is need to consider which sub-language best suits the intended application. The development of ontology is firmly driven by the intended usage. This is because the major difference between the OWL Full and OWL-DL is that OWL Full allows the use of classes as instances while OWL-DL does not. [GvH04] asserts that the choice of OWL Lite and OWL-DL depends on the extent at which users require the expressivity of constructs provided by OWL-DL and OWL Full. The choice between OWL-DL and OWL Full depends on the level to which users require the meta-modelling facilities of RDF Schema (e.g., defining categories of classes or attaching properties to classes).

2.4.2 Semantic Web Rule Language(SWRL) and Semantic Query Enhanced Web Rule Language(SQWRL)

SWRL is an expressive OWL-based rule language. SWRL provides more powerful deductive reasoning capabilities than OWL alone. It has formal semantics that extends OWL-DL and produces inferences that are informed by all OWL-DL constructs. SWRL has increased expressivity thus leading to undecidability of inference, though the inference made is formally sound [OCKT+05]. OWL-DL's inference is decidable, however the worst case decidability performance guarantee is pretty bad. For inference with SWRL could be theoretically undecidable [OCKT+05]. These facts make SWRL very expressive and built-in features of OWL become redundant in SWRL [OCKT+05]. To exploit the explicated information available or even on a common application development environment, the OWL, RDF and RDFs languages do not explain how semantic markup should be used. In the context of semantic web, rules are used as a knowledge representation language [SHD+03].

SWRL can be used to translate queries in natural language to queries on an OWL ontology. Some constraints and inferences cannot be expressed in OWL but can be expressed as SWRL rules or queries based on SWRL rules [OCSPA09]. However, some of the inferences require additional reasoning beyond what is supported by SWRL and OWL. The inference relation expressed as a SWRL rule can be asserted into the OWL knowledge base. SWRL rules are stored as OWL individuals. These individuals are described by OWL classes that are contained in SWRL ontology. However, it has the disadvantage that it involves incomplete inference [OCKT+05]. SWRL rules are full Horn like rules written in terms of OWL classes, properties, individuals, and data values [OCKT+05].

Example

In [OCKT+05] it is given that" The concept of person and male can be captured using an OWL concept (class) called *Person* with a subclass Man; the sibling and brother relationships are expressed using OWL concept (object) properties *hasSibling* and *hasBrother* with a domain and range of *Person*". The rule in SWRL would then be:

$Person\ (?p) \wedge hasSibling\ (?p, ?s) \wedge Man\ (?s) \rightarrow hasBrother\ (?p, ?s)$

Where p and s are variables.

Execution of this rule has the effect of adding the *hasBrother* property to all OWL concepts with one or more male siblings and assignment of its value to those siblings [OCKT+05].

SQWRL is an extension of SWRL. It supports querying of OWL ontologies. It is a built-in library implemented using standard SWRL built-in mechanism.

Besides, it is syntactically and semantically compatible with SWRL [OCSPA09].

2.4.3 Basic formalism for Description Logics

The formalism for Description Logics are the TBox and the ABox. In the TBox, the terminology of the knowledge base is defined while that of the ABox contains assertions concerning instances in the knowledge base. A TBox has a set of axioms that define how concepts (class) and roles (properties) are related.

In the following definitions, A and B signifies atomic concepts, R atomic roles, C and D concept definitions. The most general terminological axioms can be $C \sqsubseteq D$ and $C \equiv D$. The first axiom is an *inclusion* while the second is *equally*. Restriction of the left hand side to an atomic concept, the equality axiom becomes specialized.

2.5 Existing ontology mapping approaches and tools

Ontology mapping can take different embodiments: ontology-based approaches and automatic or semi-automatic approaches. Furthermore, there exist a number of tools for this operation.

2.5.1 Mapping approaches

There are two major approaches for mapping discovery between ontologies. They include the following:

Ontology-based approach

Since one of the role of ontologies is to facilitate knowledge sharing. Ontologies are developed with the explicit goal of providing the basis for further semantic integration. General upper ontology is agreed upon by developers of different applications. These developers can extend this general ontology with concepts and properties specific to their applications [Noy04]. This extension is performed in a consistent way with the definitions in the shared ontology. This enables finding correspondences between two extensions and is facilitated by this common 'grounding'. Examples of these upper ontologies include: SUMO, DOLCE, SENSUS and Cyc [KC07].

(Semi)-automatic machine learning approach

This second approach comprises of machine learning techniques that use various characteristics of ontologies to find mappings. These characteristics can either be the structure or definitions of concepts and instances of classes. This approach tends to rely more heavily on features of concept definitions or on explicit semantics of these definitions. Automated reasoning is used to identify the mappings.

Comparing the approaches

Researchers are hopeful that the domain and application specific ontologies reuse the foundational ontologies (SUMO, DOLCE, etc). This facilitates interoperability between these ontology-based applications. There are no concrete evidence of experience about these approaches claiming them to be success. However, there are reports on both the successes and challenges encountered for reusing them.

It is helpful to have ontologies that we need to match referring to the same upper ontology or conforming to the same reference ontology. However, this 'luxury' does not exist and the need to create mappings between ontologies that perhaps use the same specification language. Though they do not have any vocabulary beyond the specification language in common [Noy04]. People are reluctant to reuse because of usual problems with having standards. Nevertheless, there have been some successes (in domain-specific settings) and failures. It is due to these reasons that we consider the second approach which takes into account the definition of classes and instances in order to establish the mappings.

2.5.2 Ontology editing and mapping tools

[NM02] categorizes the tools into *tools for developing ontologies* and *tools for mapping, aligning, or merging ontologies*.

The *ontology development tools* are ontology editors. The mostly used ontology development tools are as shown in table B.1 in the *appendix B* [fBO00]. Ontology editors are important tools for supporting the elaboration of ontologies. They facilitate development and management of ontologies. In addition, they support the definition and modification of concepts, properties, axioms and restrictions, even some of them enable inspection and browsing of ontologies.

Tools for mapping, aligning, and merging ontologies are referred to as *mapping tools* in [NM02]. They help in finding similarities and differences between heterogeneous source of ontologies. These mapping tools can identify potential correspondences automatically. They can also provide the environment for the users to find and define these correspondences or both. Some examples of these tools are presented in table B.2 in *appendix B* . The following is an elaboration of Protégé and SIM-DL editing and mapping tools respectively. They are key tools in this research study.

Protégé

Protégé-OWL facilitates the browsing and editing of OWL ontologies. It offers a plug-in interface which allows visualization and other components to be built into its interface. It allows users to construct a domain ontology, to customize knowledge-acquisition forms and to enter domain knowledge. It is able to operate as a platform for extending access to other knowledge based systems and embedded applications as a library. This library can be used by other applications to access and visualize knowledge bases. Protégé offers a graphical user interface which allows ontology developers to focus on conceptual modelling without a need to know syntax of an output language such as RDFS or OIL [NSD+01]. The core of this environment is the ontology editor and the library of plug-ins it holds. These plug-ins add more functionality to the environment [CFLGP03]. Protégé has SWRLTab which is a SWRL editor plug-in and is for editing SWRL rules. SWRLTab supports an automatic completion of properties and class names. It also checks for syntax of the rules that have been entered.

SIM-DL

The SIM-DL reasoner and the plug-in are still under development. It is based on combined ideas from feature and network (distance) similarity models. The current beta version supports subsumption reasoning and similarity measurement up to $ALCHQ$ level of expressivity. Its syntax and semantics with respective names are shown in table D.1. This level of expressivity is defined as follows:

- AL: Allows concept intersection, full universal quantification, atomic negation and limited existential quantification (i.e. existential restrictions with fillers limited to OWL:thing)
- C: Complex negation (e.g not(A or B)). ALC allows disjunction and full existential quantification. This can be represented with conjunction and full negation and universal quantification and full negation respectively.
- H: Role (property) hierarchy.
- Q: Qualified number restrictions (qualified cardinality restriction).

At the moment, SIM-DL does not support for more expressive Description Logics beyond $ALCHQ$,(e.g., Functional roles(F) properties). SIM-DL is an open and Java-based semantic similarity reasoner for diverse description logics (from the ALC family up to $ALCHQ$). The reasoner can be used via DIG or a Protégé plug-in called SimCat. SIM-DL also supports classical subsumption reasoning and satisfiability checking. It compares a search concept with target concepts from two ontologies.

SIM-DL supports both a symmetric and asymmetric similarity measurement *see expressions* 2.1 and 2.2 respectively. By asymmetric, it means in one direction while for symmetric, it is in both. For instance, for asymmetric similarity measurement, a search concept can be taken as source or target concept as target in one direction. For symmetric, a search concept is taken as source or target concept as target in one direction. On the other hand, the same source or target concept in the other direction correspond to the former as in expression 2.1. These similarity measurements are used for information retrieval and alignment within an ontology or several ontologies. In SIM-DL, similarity between concepts in canonical form is measured by comparing their $ALCHQ$ definitions for overlap. A high level of overlap indicates high similarity and vice versa. In Description Logic (DL), concepts are specified based on primitive concepts and roles. This requires use of language constructors such as intersection, union and existential quantification. Similarity is defined as a polymorphous, binary and real-valued function X*Y $\rightarrow$ R[0,1] providing implementation for all language constructs offered by the used DL. It compares formal set restrictions [JKS+07].

$$sim(x, y) = sim(y, x)(symmetry) \tag{2.1}$$

$$sim(x, y) \neq sim(y, x)(assymetry) \tag{2.2}$$

Where x and y are the concepts to be compared to get the similarity $(sim(x, y)$. For instance, how similar is a Lake to a Canal both being water bodies? Taking x to be a search concept,C_s, e.g., Lake and comparing it to y, target concept, C_t, e.g.,Canal. Lake and Canal as water bodies are used as examples of concepts in this section.

The overall similarity between concepts is a sum of normalized (and weighted) single similarities calculated for all parts (i.e., superconcept) of the concept definitions. A similarity value of 1 indicates the compared concepts cannot be differentiated, whereas 0 implies completely distinctive.

The framework that SIM-DL applies to determine similarity between concepts specified using $ALCHQ$ level of expressivity is as follows. The framework consists of five steps and their implementation depends majorly on the semantic similarity and the underlying representation language [JKS+07].

1. *Query (search concept) and target concept selection*. Concepts that have to be compared are selected first. The search (query) concepts,C_s is part of the ontology being examined. Target concepts,$C_{t_1}, ..., C_{t_2}$ form the context of discourse,C_d (e.g., Water bodies). The context of discourse determines the set of concepts that are compared to the search concept C_s (Lake) and the target concepts, C_s (Canal). They are determined by specifying a context concept, C_c or selecting manually. Selection of context,C_c (Water bodies) concept determines which concepts are to be compared and it influences the similarity [Jan08, KRJ+08]. SIM-DL defines the set of target concepts which are subsumed by context concept C_c. It is defined in the expression below. The search concept,C_s is compared to each target concept,C_t from the set.

   $C_d = \{C_t | C_t \sqsubseteq C_c\}$.

   For instance,

   $waterbodies = \{Canal | Canal \sqsubseteq waterbodies\}$

2. *Concepts are transformed to canonical form*. Semantic similarity relies on contents about the concepts [JW09]. Given concept description of two ontologies specified in a given language and denoting same facts using different language elements. Normal form completes subsumption test for $ALCHQ$. These descriptions are transformed to a common form to eliminate unintended syntactic influence (i.e., to ensure that the syntactic dependency is reduced). This stage needs rewriting of rules to get a

canonical representation of compared concepts. For instance, the equivalent concepts are reduced to normal form. This is possible by means of rewriting rules that can preserve their equivalence as in expression 2.3. In addition, a map between equivalent expressions like $\forall R.\perp$ and $\leq 0R$. This ensures that specified concept descriptions are used only where they have an impact on cardinality of the sets regarded. For example this cardinality expression$(\geq 1R.C) \cap (\geq 2R.C)$ is mapped to $(\geq 2R.C)$ [JKS+07].

$$\forall R.C_s \sqcap \forall R.C_t \equiv \forall R.(C_s \sqcap C_t). \tag{2.3}$$

For example, assuming both the Lake and the Canal to be *navigable*, then

$\forall Navigable.Lake \sqcap \forall Navigable.Canal \equiv \forall Navigable.(Lake \sqcap Canal).$

Where $\forall R.C_s \sqcap \forall R.C_t$ has been rewritten to $\forall R.(C_s \sqcap C_t)$.
Rewriting of rules becomes increasingly complex with the expressivity of the used language.

3. *Definition of an alignment matrix for concept descriptors*: The former steps are for determining which concepts are selected for comparison. In this step, the alignment matrix give specification on which and what concept descriptors? Descriptors, D, for search concept are given as C_{D_s} and C_{D_t} for target concepts. For instance, super-concepts or subconcepts, dimensions and features are compared. The selection of comparable tuples of descriptors are in matrix form $C_{D_s} \times C_{D_t}$, which are the descriptors sets for the search concepts, C_s and target concepts, C_t. For SIM-DL, two compared concepts form an alignment matrix M_1 with all possible combinations of their parts. These parts are primitive, an existential quantification, value restriction or qualified number restriction [JKS+07].

4. *Application of constructor specific similarity functions to comparable pairs*: For each comparable concepts tuples with their respective descriptors, similarity is measured. Similarity function for each representation language is different and yields values between 0 and 1. SIM-DL offers a similarity function for each constructor [JKS+07].

5. *Determination of normalized overall similarity*: Overall similarity, equation D.1 for each selected tuples of the compared concepts is obtained by summing the single similarity values derived by application of similarity functions. For SIM-DL, each similarity function takes care of its normalization using the number of compared tuples. Each similarity function returns a value between $0 - 1$ to the function it was called by [JKS+07].

Similarity measurement process starts at the union level. Each concept on this level is formed by intersection. Similarity between concepts is measured by Sim_i function. Concepts of this intersection can be either primitive(Sim_p), an existential quantification(Sim_e), value restriction(Sim_f) or qualified number restriction $(Sim_{min}, Sim_{max}$ respectively) [JKS+07].

2.6 Description Logic reasoners

Description Logic reasoner performs several inference services. Some of the services include computing the inferred superclass of a class, determining whether class is or not consistent. It also helps in deciding whether or not, one class is subsumed by another, etc. Some of the popular Description Logic reasoners that are available are listed in table B.3. Another very important issue when creating applications on ontologies is the choice of a DL reasoning engine for the use cases that need reasoning. Various efficient reasoners are available for DL. They have different DL and their implementations are summarized in table B.3 of section B.2.

To determine which reasoners suits this research study, some tests were performed on the ontology. They include consistency checking, classification of ontology, satisfiability and checking for an instance support. Based on the tests and also from literature, Pellet and SIM-DL reasoners provide better support for this research study. Both are easily available. Pellet has much higher performance.

Sirin claims that Pellet supports the SPARQL-DL format for querying OWL-DL and OWL 2 ontologies. It can be accessed via three different APIs. The internal API is designed for efficiency but is missing features and has low usability. The Manchester OWL API has features for managing ontologies, though it does not support SPARQL. Jena lacks specific OWL 2 support. It is a Java framework used for building Semantic Web applications. However, it supports SPARQL and has a built-in SPARQL query engine. Therefore, the only way to retrieve knowledge from OWL 2 documents and stay be with the Simple Protocol and (Resource Description Framework) Query Language (SPARQL) standard is by utilization of Pellet's query engine through the Jena API [SP07].

2.7 Advantages and Limitations of ontology mapping

Ontology mapping is deemed to have the following benefits and challenges:

- Data transformation in order to make it conform to the other ontology.
- Used in answering the queries asked in a particular application. Here views are created.
- It helps in reasoning with the mappings like mapping composition.
- It enables extension of the ontologies to describe web services and to get the description of the services.
- Ontology mapping can cater for language differences during information integration.

Besides the above benefits that ontology mapping can support, it has the following limitations.

- There is no commonly agreed practical ontology mapping life cycle to prompt the emergence of a dynamic market for software solutions to support ontology mapping.
- Developers need to agree on a common way to specify the results of matching algorithms and mapping systems. Without this agreement, the use of mappings by users will quickly become limited.
- Lack of qualified cardinality restrictions, more expressive datatype reasoning and property chain inclusion axioms in the mapping tools. These can be a major issue for user community.

2.8 Summary

In this chapter, an insight on the theory that forms a basis for this research study has been presented. The understanding of the ontology and its relation in geographical perspective. The languages required for the explication and retrieval of the geospatial data have been looked at. The ontology editing and mapping tools as well as the DL reasoners have been explored. The next chapter involves the use of this theory in capturing the semantic descriptions of the land cover data, use of property relations and modelling of the knowledge.

Chapter 3

Use Case and Knowledge Modelling

Currently, technologies and devices are being created in order to capture a large amount of information about the land cover. Geographic surveys of the same phenomenon in a particular area are frequently carried out by various agencies or organizations. This may result in significantly different land cover information, even when methods employed are similar. These land cover data are analyzed and stored in Geoinformation Systems (GI) and made available via web. A search for geographic information on the Web returns several links representing different parts of our world.

3.1 Case study and data description

In this research study,MODIS-MCD12Q1 land cover type data were used. These land cover type are derived from observations spanning a years input of Terra and Aqua data. They are described by the classification schemes of UMD Global Land Cover Classifications and IGBP Global Land Cover. These schemes identified 17 land cover classes [Cen09]. The definitions from the schemes are obtained through a supervised decision-tree classification method. The features in the datasets and how it was collected are the same for both schemes in terms of scale, spatial extent and in temporal. However, difference occur in the definitions of concepts and naming of a few categories of land cover. This set of dataset cover the study area of Netherlands.

Another set of datasets to be used for this research study and in the same area is from CORINE and the national land cover/ use database of Netherlands (LGN 4 data). CORINE describes land cover (and partly land use) according to a nomenclature of 44 classes organized hierarchically in 3 levels. This database is based on national classification system and consists of 46 classes. It consists of land cover/ use classes [TW00]. It is organized hierarchically in two levels. This means that the two datasets have differences in the definitions of concepts, categories, classification, granularity,resolution and method of collection. CORINE dataset is in vector form while LGN 4 datasets is in raster form.

3.2 Use case scenario

An Environmental Agency (EA), for instance, would like to carry out interdisciplinary analysis of land cover dataset through integration and retrieval. Further, it may even want to improve the land cover information to enable sharing and reuse. These demands may encounter land cover datasets of the same feature. These datasets are divided between more than one system. As a rule, EA has either to choose one of them or carry out integration of these land cover datasets. A similar situation may arise when a given agency (Cadaster) applies some methodology (Cartographic) to carry out thematic survey. The EA however, might have to match this to other land cover dataset about another theme. This second thematic data was obtained using some other methodology, hence different data sources. As practice, EA officer queries from each data sources separately. The resulting answers may or may not be identical.

Supposing the EA gets all answers that agree, then it is inferred that no integration of these land cover datasets was needed. Other than supposed, there could be several discrepancies that could however be easily resolved. These land cover datasets are therefore heterogeneous in terms of representation and semantics. Semantics is the meaning that these different organizations attribute to their datasets according to their understanding of the world. This stands to reason however, that processes that involve categorization of answers demands for previous knowledge about the sources of the land cover datasets.

EA should have knowledge of present distribution and area of, such as agricultural, recreational and urban lands. In addition, information on their changing proportions. Failure to detect and resolve semantic discrepancies, the usage of integrated land cover dataset by the EA officer leads to invalid results. Even worse, the user(EA) do not know about semantic heterogeneity in the data they access and do not have a chance to realize the invalid results. With the occurrence of environmental and social sciences study phenomena over geographic space, a formal ontology of geospatial phenomena is essential for interoperable geospatial science.

Relying on common sense is critical source of semantic heterogeneity. Therefore, explicit definition of terms used in concept definitions is a solution to this problem [HG01]. Application of formal ontologies is a potential solution of semantic heterogeneity. These are guided by explicit and formal definition of concepts used in land cover classification systems. A formal ontology consists of *logical axioms* that convey a meaning of concepts for a particular community. An agreement on ontological definitions among members of an organization is an important discrepancy between ontologies and conceptual models. Figure 3.1 shows the process at which the needs of the EA can be achieved. Throughout this research study, figure 3.1 forms a basis towards achieving our research objectives.

This research study starts with formalization of land cover data sources in order to establish similarities between concepts from two different ontologies. This similarity measure is defined for both the concept (object) and instance (datatype) as in figure 3.1 for the concepts selected. These similarity measures give confidence to the EA officer to use the land cover datasets. These mea-

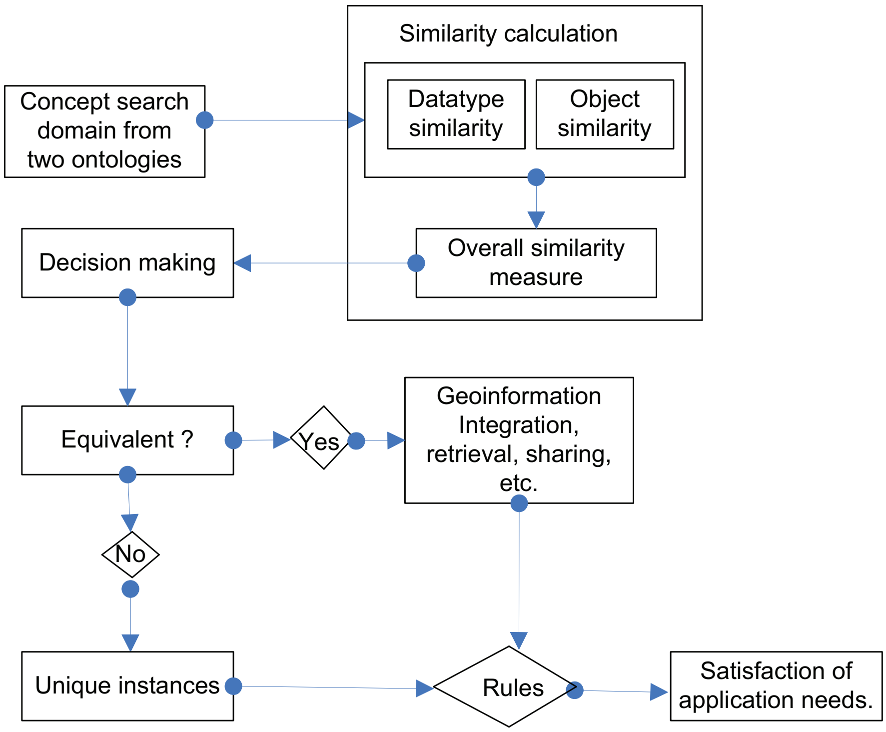

Figure 3.1: Ontology mapping and decision making process towards the intended applications.

sures depicts how similar or different the land cover datasets are, thus degree of corresponding concepts. A mapping of any two different concept definitions is established.

This EA can have the confidence of using the data if the outcome of similarity is consistent (equivalent) or inconsistent (disjoint). This level of confidence are related to the user (EA) and is determined by the intended application needs. The EA has to define the threshold value that two concepts are labeled as equivalent or not depending on the threshold *see figure* 3.1. This knowledge is then used by the EA to resolve land cover semantic variation by performing integration, retrieval and even for improving their quality, thus enable sharing and reuse.

To evaluate how well the concept semantic similarities calculation using SIM-DL meet the similarity of feature representation, a validation is conducted. In validating the proposed method for semantic similarity measurements, the raster datasets for our study area are compared. The representation of features are in pixel form. Therefore with regard to the datasets, we have *per-object* and *per-pixel* scenarios. In per-pixel case, individual pixel is treated as basic unit while in per-object it consists of multiple-pixel regions (objects). The more features that match between two raster or vector datasets the higher is the similarity. In this validation, attention is paid to assessing the similarity measures using the pixel-by-pixel and object-by-object approaches.

3.3 Comparison of land cover classification categories

The categories at the lowest hierarchical level were examined for CORINE and LGN 4, UMD and IGBP. For simplicity and clarity, this research study was restricted only to a small, but representative set of land cover categories from the two classification systems. They were properly selected to account for a range of heterogeneities encountered between land cover categorizations. The selected categories are:

- CORINE categories 1 (Artificial surfaces) and 2 (Agricultural areas)
- LGN 4 categories Urban area and Agricultural area.
  While for UMD and IGBP, a few of the classes were considered as shown in table C.3

The tables C.1 and C.2 show the categories of the land cover/use used in this research study. The term ‘category type’ refers to categories found in various land cover classification systems under the same term (name of the category). However, this exhibits differences in their definitions or the contexts under which they are used.

3.4 Determination of semantic relations and properties

Enrichment of land cover classes with semantic properties and relations for revealing similarities and difference was performed. Definitions of categories in classification systems are important source of semantics. They are the only available sources that can be relied upon especially in existing land cover data collections [KKT05]. These were meant for disambiguation of the given geographic categories. Examples of these properties and relations are shown in table 3.1.

Example

In UMD land cover classification, *Evergreen Broadleaf Forests* is defined to be ‘lands covered by trees with a percent canopy cover ≥ 60 % and height exceeding 5m’. The semantic properties and relations such as *MaterialCover* of value ‘Tree canopy’ and *hasHeight* of value ‘$\geq$ 5m’ and *hasCoverOf* with value ‘≥ 60 %’ were determined.

The same land cover category defined in IGBP classification to be ‘Lands covered by trees with percent canopy cover ≥ 60 % and height exceeding 2 meteres’. Similar semantic properties and relations determined above applies

Table 3.1: Some examples of semantic properties and relations

Semantic properties
Purpose
Cause
Location
MaterialCover
Size
Semantic relations
coveredBy
hasCoverOf
contains
isWithin
AssociatedWith

in this case. It is evident that the two different land cover classification systems define the land cover 'Evergreen Broadleaf Forests in different ways *see table* 3.2. It was important to specify semantic relations and properties used

Table 3.2: Determination of semantic information for category 'Evergreen Broadleaf Forest' abbreviated as *E. Broadleaf* from UMD and IGBP.

DN code	Land cover	MaterialCover	hasHeight	hasCoverOf
2	E. Broadleaf (UMD)	tree canopy	$\geq$ 5m	$\geq$ 60 %
2	E. Broadleaf (IGBP)	tree canopy	$\geq$ 2m	$\geq$ 60 %

in defining land cover categories. These were used for decomposing definitions of land cover categories. The semantic information was used to disambiguate similar categories. This disambiguation is by modelling the knowledge through ontologies *see section* 3.5. The definitions after modelling form the basis for semantic similarity calculation.

3.5 Ontology modelling and principles

Formalization of the knowledge extracted in section 3.4 is an iterative process *see figure* 3.2. Modelling of ontologies has no defined methodology. However, they include set of stages occurring in ontology modelling. Besides, there are guidelines and principles *(see subsection 3.5)* that assist in all stages. From literature, there are two kinds of complementary methodologies. These are *stage-based* e.g [UG96]) and *iteration evolving prototype*e.g., (MethOntology [P9́4]). These methodologies have two stages namely:

- *Informal Stage*: In this stage, the ontology is sketched out using either natural language description or some diagram techniques.
- *Formal stage*: This stage in particular involves encoding of the ontology in a formal knowledge representation language such asOWL that is machine computable.

The informal representation in this sense helps the former(the ontology designer) and the formal representation helps the latter (the user). A methodology followed in this research study is as in [MDH05] and is summarized in figure 3.2.

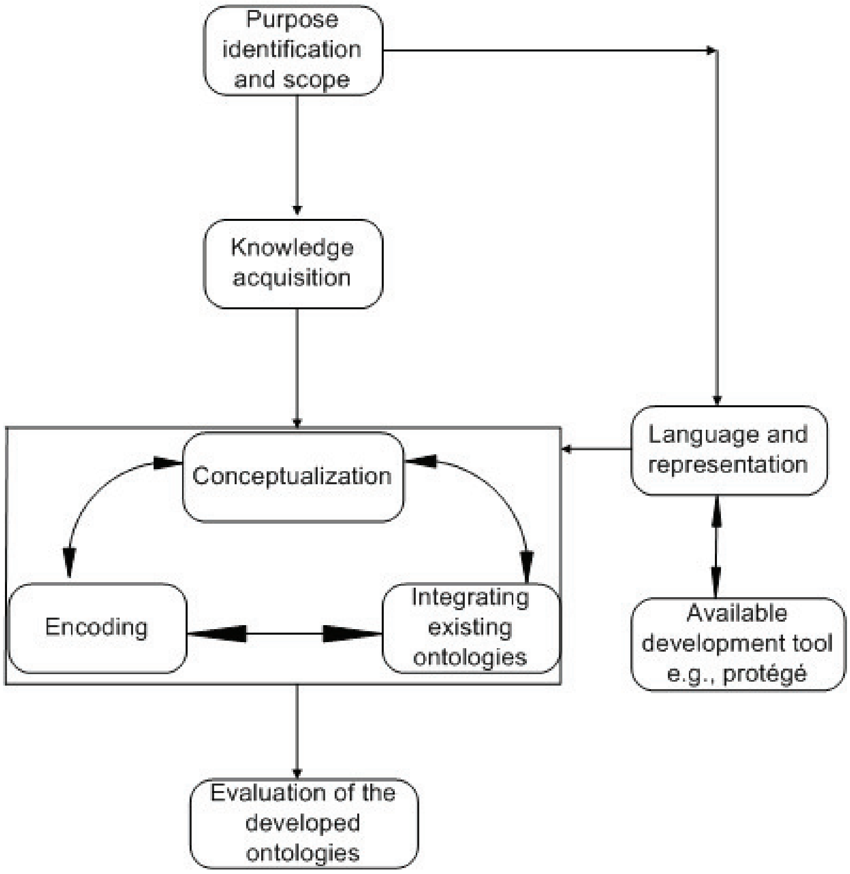

Figure 3.2: ontology building process

Ontology modelling principles

As pointed in section 3.5 above, principles arising from the language structure are vital. A peril is on incorrect use of $\exists$ and $\forall$ constructs. These constructs differ in their formalOWL interpretation from the natural language interpretation. $\forall$ construct in natural language means *each and every* while in DL it

Deciduous Broadleaf Forests $\equiv$ Land $\wedge$ $\exists$ hasCover.Broadleaf

Deciduous Needleleaf Forests $\equiv$ Land $\wedge$ $\exists$ hasCover.Needleleaf

Figure 3.3: Example of how concepts are constructed

means *each and every if there is one*. $\exists$ construct is equivalent to the informal language construct meaning *at least one* thus posing less problems. However, care must be observed. This is because informal language assumption may not hold true in DL. The fact that OWL uses open-world reasoning, the variation in construct interpretation cause problems when modelling ontologies. The problem surfaces during subsumption hierarchy calculation. It is quite vital to ensure concepts keep the meanings when translated intoDL definitions [Hal06].

It is difficult for reasoners to understand implicit knowledge in land cover concept and restriction names. This knowledge should be made explicit. However, these intentions influence each other. For instance, addition of knowledge to create a smart model decreases its level of understanding. Increasing readability of model may make it impossible to describe some aspects of the domain. In addition, the language used and its capabilities may also influence the structure of the model. Avoiding these troubles require use of necessary and sufficient conditions *see figure* C.1. This restricts the properties to certain values thus not having the defined concepts arranged in hierarchy. Definition of concepts on necessary and sufficient conditions make reasoner infer the defined concepts. Example in expression 3.3 shows how concepts are formally defined from informal definition. The tables 3.3 and 3.4 are extracts of two ontologies with the formal definition of concepts and roles.

This means that Deciduous Broadleaf Forests is equivalent to vegetation and at least one of its cover must be broadleaf. For Deciduous Needleleaf Forests is vegetation and at least one of its cover must be needleleaf.

Table 3.3: Definition of concepts and roles for IGBP ontology

Some concepts within IGBP ontology
Deciduous Broadleaf forests $\equiv$ $\exists$ contains (seasonalBroadleaf) $\sqcap$ ($\geq$ 60% (cover)) $\sqcap$ ($\geq$ 2m (high))
Deciduous Needleleaf forests $\equiv$ $\exists$ contains (seasonalNeedleleaf) $\sqcap$ ($\geq$ 60 % (cover)) $\sqcap$ ($\geq$ 2m (high))
Evergreen Broadleaf forests $\equiv$ $\exists$ contains (Evergreen trees) $\sqcap$ ($\geq$ 60 % (cover)) $\sqcap$ ($\geq$ 2m (high)) $\sqcap$ $\forall$ hasfoliage (Green)
Evergreen Needleleaf forests $\equiv$ $\exists$ contains (Evergreen trees) $\sqcap$ ($\geq$ 60 % (cover)) $\sqcap$ ($\geq$ 2m (high)) $\sqcap$ $\forall$ hasfoliage (Green)
Grasslands $\equiv$ $\exists$ coveredBy (Herbacious trees) $\sqcap$ ($\leq$ 10 % (trees)) $\sqcap$ ($\leq$ 10 % (shrubs))
Water bodies $\equiv$ $\exists$ contains (Lakes $\sqcup$ Oceans $\sqcup$Seas $\sqcup$ Reservoir)
Mixed forests $\equiv$ $\forall$ 4 (forest types) $\sqcap$ ($\geq$ 60% (cover)) $\sqcap$ ($\geq$ 2 m (height))

Table 3.4: Definition of concepts and roles for UMD ontology

Some concepts within UMD ontology
Deciduous Broadleaf forests $\equiv \exists$ contains (seasonalBroadleaf) $\sqcap$ ($\geq$ 60 % (cover)) $\sqcap$ ($\geq$ 5m (high))
Deciduous Needleleaf forests $\equiv \exists$ contains (seasonalNeedleleaf) $\sqcap$ ($\geq$ 60 % (cover)) $\sqcap$ ($\geq$ 5m (high))
Evergreen Broadleaf forests $\equiv \exists$ contains(Evergreen trees) $\sqcap$ ($\geq$ 60 % (cover)) $\sqcap$ ($\geq$ 5m (high)) $\sqcap \forall$ hasfoliage (Green)
Evergreen Needleleaf forests $\equiv \exists$ contains (Evergreen trees) $\sqcap$ ($\geq$ 60 % (cover)) $\sqcap$ ($\geq$ 5m (high)) $\sqcap \forall$ hasfoliage(Green)
Grasslands $\equiv \exists$ coveredBy (Herbacious trees) $\sqcap$ ($\leq$10 % (trees)) $\sqcap$ ($\leq$ 10% (shrubs))
Water bodies $\equiv \exists$ contains (Lakes $\sqcup$ Oceans $\sqcup$ Seas $\sqcup$ Reservoir)
Mixed forests $\equiv \exists$ coveredBy(Decidous $\sqcap$ Needleleaf) $\sqcap$ ($\geq$ 60 % (cover)) $\sqcap$ ($\geq$ 5m (height))

3.6 Summary

In this chapter, the use case was introduced in a land cover application context. The semantics from the natural definitions of land cover categories were extracted and formally defined. An overview of various stages of concept modelling in ontologies was observed. Guiding principles in constructing the concepts from natural to formal interpretations were noted as well. As a result, the respective ontologies for land cover classification systems were constructed as shown in Protégé environment figure C.1. These modelled knowledge form a basis for semantic similarity measurement between concepts in chapter 4.

Chapter 4

Geo-semantic Similarity Measurements for Mapping

This chapter pertains to the measurement of the concept semantic similarity between two different ontologies. Semantic similarity measurement plays a vital role in spatial data integration for the use case in chapter 3. Semantic similarity measurement method using SIM-DL is considered. This leads to establishment of necessary links used as the basis for integration of geoinformation.

4.1 Semantic similarity assessment

Similarity has been set to determine why and how entities are categorized and why some categories can be similar to each other while others are not [GS04]. The mapping between concept definitions is the initial stage towards spatial data integration, retrieval and improvement of these datasets. This mapping between concepts semantics is a global problem. Measuring semantic similarity among the land cover concepts is a core method for assessing the degree of semantic interoperability within and between ontologies. It is essential for dealing with unclear data queries, unclear concepts or natural language. In addition, it acts as basis for semantic information integration and retrieval.

A number of approaches that define the mapping apply lexical relations method. This mapping cannot be accomplished solely by lexical comparison. This is because names (like tags) can be abbreviations, acronyms, phrases in different languages. More so, they could be misspelled or used unexpectedly in jargon specific ways. Other methods compare the structure of ontologies, contrary to semantics. The challenge of semantic similarity measurement is the comparison of meanings as opposed to purely structural. The semantics during the mapping have to be preserved.

Semantic similarity measurement in this research study is established while keeping the ontologies independent from each other. These semantics were defined and formalized in chapter 3. A language *see section* 2.5.2 is specified to express the nature of entities and functions. It is needed to determine how (conceptually) close the compared entities are. Entities can be expressed in terms of attributes. These representations of entity types are more complex. Entity types are specified either as sets of features, dimensions in multidimensional space or formal restrictions specified on sets using various description logics. These are dependent on the expressivity of representation language. Similarity is measured between entity types which are representation of concepts in human minds. This depends on what is said in terms of computational representation about these entity types. This is also dictated by the representation language, leading to the fact that most measures of similarity cannot be compared [JKS$^+$07]. Context is a challenge for semantic similarity measurement. It influences semantic similarity measurement. Semantic similarity cannot be determined without defining in relation to what is similarity measured [KRJ$^+$08].

4.1.1 Computation of the semantic similarity

The land cover concepts are nodes linked by *is-a* and *partOf* edges in a semantic network having a hierarchical structure. The main approach in which SIM-DL computes the concept similarity is intensional-based approaches. Concepts in our case are in $ALCHQ$ and below normal form. Similarity measurement using SIM-DL server is a process as depicted in figure 4.1. The search concept and and the target concept are selected. Figure 4.2 shows SIM-DL interface as a plug-in in Protégé. With reference to figure 4.1 an explanation of how this measurement is obtained is elaborated as follows.

This similarity is calculated as a number of superconcept, the target concepts, C_t shares with the search concept, C_s *(see section 2.5.2part 1)*, divided by the number of superconcept of C_s for standardization [KRJ$^+$08]. It also measures similarity formed by existential values, number restrictions or quantifications. More over, it considers network distance measure for roles and also determines the statistical co-occurrence of primitives. The overall similarity D.1 between concepts is a normalized and weighted sum of single similarities calculated for all parts(i.e. superconcept) of the concept definitions.

The weighting is emphasized on the disjunction level contrary to intersection. This is because every individual member of a concept formed by disjunction can either be a member of all its single concepts or only some of them. This weighting acts as adjustable factor for relative importance of C_s and C_t and is always 1 [JKS$^+$07]. Normalization ensures that derived inter-role similarity is ranging between 0 and 1. It is integrated as part for similarity measure introduced for restrictions and quantifications. During canonization *(see section 2.5.2 part 2)*, the equivalent concepts are reduced to normal form by means of rewriting rules that can preserve their equivalence. For instance in equation 2.3.

The similarity measure is based at a scale of $0-1$ but expressed between $0-100\%$. The computation process in which the SIM-DL server query the reasoner is shown in figure 4.3. Eventually, the output, overall similarity in figure 4.1, figure 3.1 and figure D.1 is a mapping as shown in table 4.1 and table D.2 for IGBP and UMD. For CORINE and LGN 4 dataset, the similarity results are shown in tables 4.2 and D.3. The context in this case is land cover and

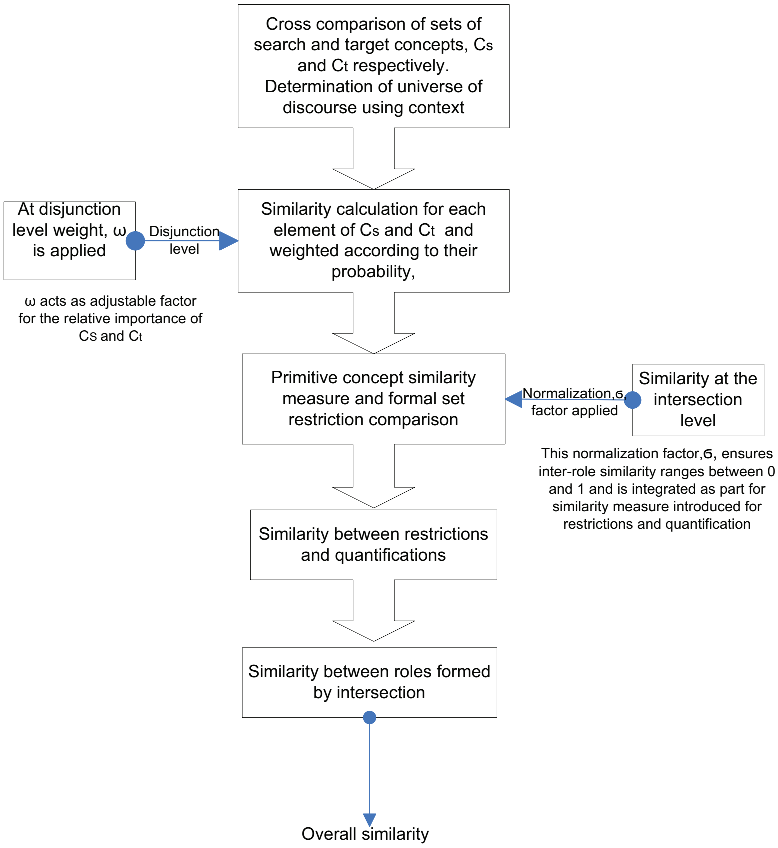

Figure 4.1: Schematic diagram for semantic similarity measurement process in SIM-DL

[KRJ+08] explains how context influences similarity measurement.

4.1.2 Discussion of semantic similarity results

Similarity measurement of land cover concept definitions using SIM-DL gives rise to semantic similarity measures between concepts. The outcome in table 4.1 shows for example, that 'Evergreen needleleaf forests' in IGBP ontology is more similar to 'Evergreen needleleaf forests'(DN=1) in UMD ontology. They both share the same superconcepts which reflects the expert's conceptualization, because the definitions given is similar, however, there is difference in height value as defined in tables 3.3 and 3.4 described in chapter 3 *section* 3.5.

The measure of SIM-DL reasoner reflects differences in concept definitions. This means the two compared concepts have difference in value restriction, despite having common superconcept and all other definitions being similar.

.

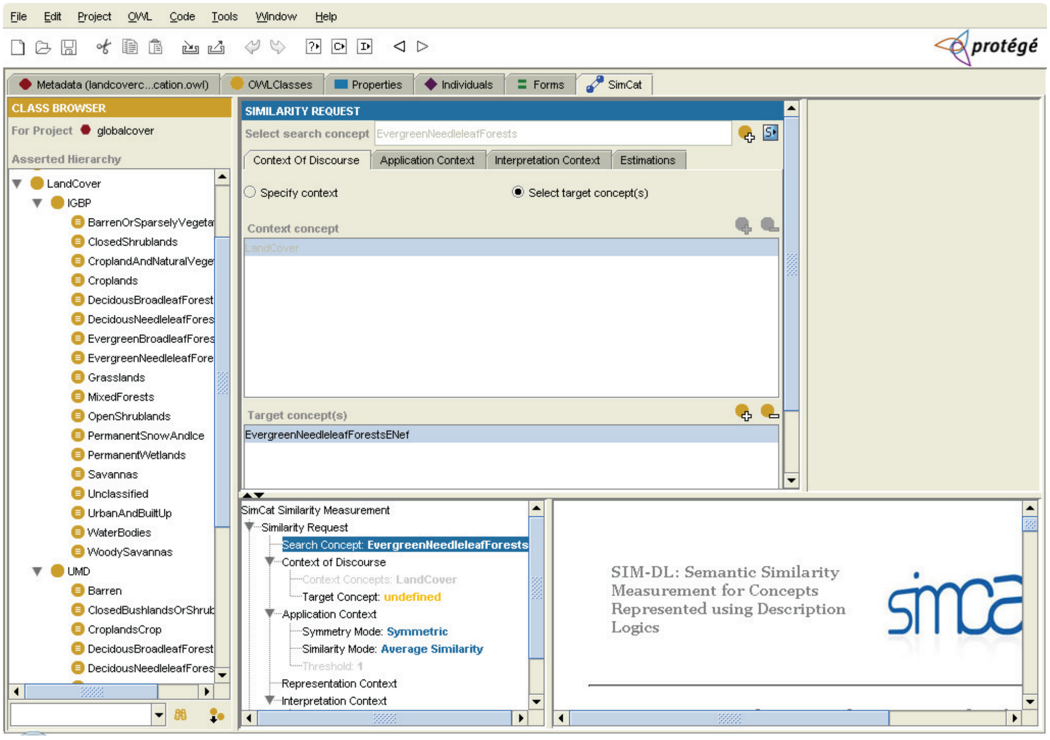

Figure 4.2: SIM-DL plugin in Protégé:Selecting search concept and target concept

The other concepts with DN=6 and DN=7 have similarity values being low. This is because they only have a common superconcept (canopy) which is a least common subsumer(lcs). The semantics of the compared concepts do not match in all properties defining them. In this semantic similarity measure, not only is the shared suprconcept considerd, but also the network distance. This is measured for roles and the statistical co-occurrence of primitives and is also determined in all situations.

For comparison of CORINE and LGN 4 land cover definitions in table 4.2, these results depict a difference in the compared concept definitions. This is as a result of variation in the understanding of land cover concepts. This leads to granularity mismatch in both land cover categories. The CORINE land cover definitions do not match that of LGN4. This is because of the difference in model coverage. The parts of the domain in CORINE are not covered in LGN 4. This is shown with a 'No class' label and (–) in the table 4.2. In addition, the level of details are different, therefore, the similarity is 0 %e.g., comparison of industrial/ commercial units for CORINE to Orchards in LGN4. This is because there is no common definition of the concepts. However, from table 4.2 SIM-DL still gives low similarity values which is very insignificant. This could be due

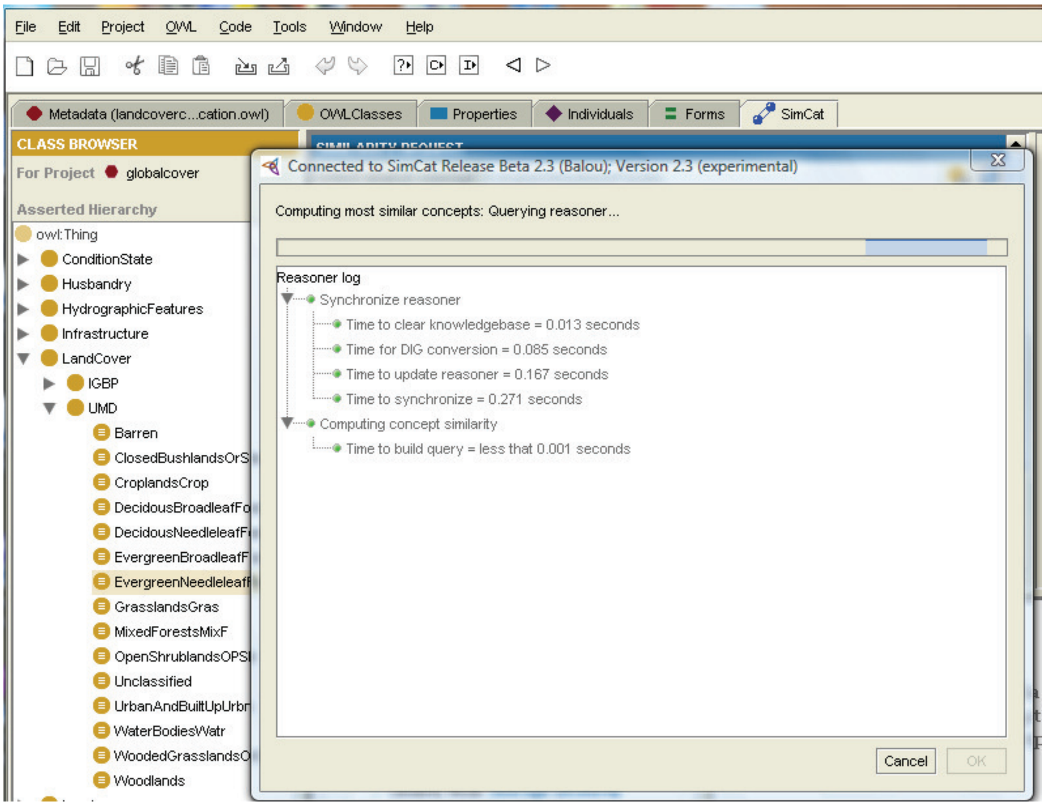

Figure 4.3: SIM-DL plug-in in Protégé:Querying the reasoner. This computes the most similar concepts.

to the imposition of rigid restriction on the output with the aim of improving its precision. Therefore, it is evident that granularity mismatch cannot permit using ontological constraints for land cover concepts (classes)in LGN4 at the lower levels of the hierarchy if CORINE ontology does not distinguish between these land cover classes.

4.2 Validation of SIM-DL similarity measures and discussion

SIM-DL semantic similarity measure in section 4.1.1 rendered the results discussed in subsection 4.1.2. We validate these results from SIM-DL by applying the validation framework defined below.

4.2.1 How to find a matching set

To determine the matching candidate between two datasets of similar granularity requires a simple overlay as shown in figure 4.4. Dataset A is overlaid

Table 4.1: Extract of SIM-DL semantic similarity measures of land cover concepts for UMD and IGBP dataset

Class code	Search concept, C_s	Target concept, C_t	Similarity, $S(\%)$
0	Water bodies	Water bodies	75
1	Evergreen Needleleaf forest	Evergreen Needleleaf forest	80
2	Evergreen Broadleaf forest	Evergreen Broadleaf forest	80
3	Deciduous Needleleaf forest	Deciduous Needleleaf forest	77
4	Deciduous Broadleaf forest	Deciduous Broadleaf forest	77
5	Mixed forests	Mixed forests	49

Table 4.2: Extract of SIM-DL semantic similarity measures of land cover concepts for CORINE and LGN 4 dataset

Class code CORINE	Class code LGN4	Search concept (C_s)	Target concept (C_t)	Similarity(%)
112	2	Discontinuous Urban Fabric	Maize	2
211	0	Non-Irrigated Arable Land	No class	-
242	5	Complex cultivation Patterns	Cereals	1
243	0	Land Principally Occupied By Agriculture	No class	-
231	8	Pastures	Greenhouses	3
121	9	Industrial/ Commercial Units	Orchards	0

to dataset B. However, if datasets are differently modelled such that one contains more objects than the other dataset, a simple 1:1 relation may or may not return the matching candidate as in figure 4.5.

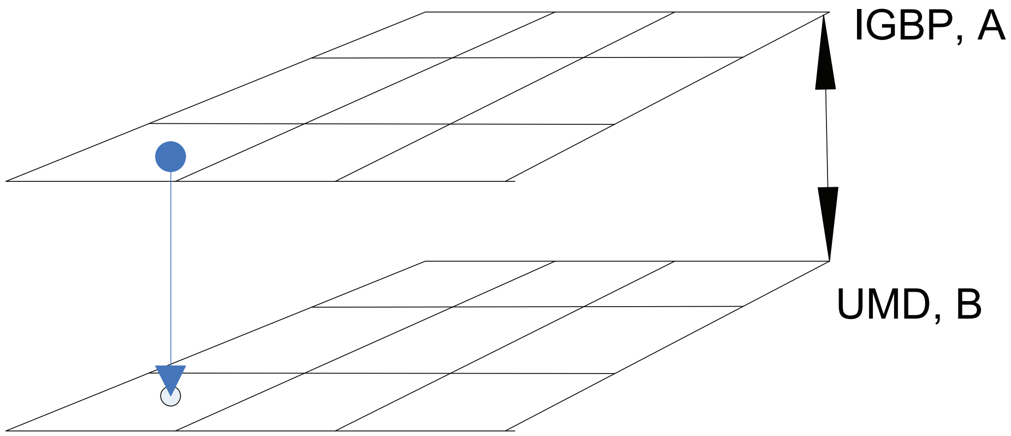

Figure 4.4: Finding a matching set in a dataset with similar granularity.

Finding a matching pair in vector dataset of polygon dimension

Identifying a matching set in two vector polygons, any two set of points that belong to the source i.e.,vector dataset, A and a target vector dataset, B,. Mapping A_i to B_j has to be found such that A_i and B_j are similar. In this regard, we propose the approach shown in figure 4.6. This is a matrix frequency in which finding a matching set process finishes on level which contains the matching candidate thus becomes a 1:1. as in figure 4.6. This approach is open for research in the future.

4.2.2 Introduction of the validation framework

Feature Contrast Model (FCM)

Tversky challenged the dimensional and metric assumption that underlies the geometric similarity models and developed an alternative feature contrast model approach. In this approach, similarity is determined by matching features of compared entities and integrating these features by the formula below. Let A, B, be feature sets of objects a, b, respectively and $S(a, b)$ be similarity measure between objects a and b as shown in figure 4.7. Feature contrast model

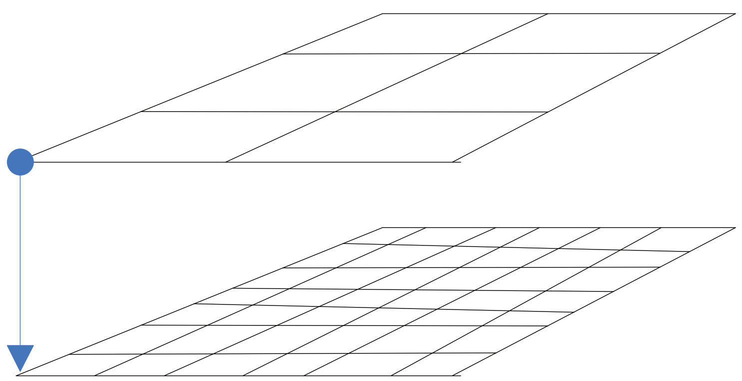

Figure 4.5: Finding a matching set in a dataset with different granularity.

(FCM) represents form of feature matching functions which satisfy Tversky's assumptions of feature matching process. This validation approach is based on matching assumption, amongst monotonicity and independence, solvability and invariance [Amo77]. The following expressions are according to [Amo77].

$$S(a,b) = F(A \cap B, A - B, B - A) \tag{4.1}$$

This implies that similarity measure could be expressed as function of three parameters: common features shared by two objects $(A \cap B)$ and distinctive features belonging to only one of the two objects (i.e, $A - B$ and $B - A$).

A simple form of matching function $F(x)$, the FCM is given by

$$\begin{aligned} S(a,b) &= F(A \cap B, A - B, B - A) \\ &= \theta f(A \cap B) - \alpha f((A - B) - \beta f(B - A) \end{aligned} \tag{4.2}$$

where $f(x)$ is a non-negative function of feature x and θ, α and β are three non-negative constants. In addition, function $f(x)$ is assumed to satisfy feature additivity as below

$$f(A \cup B) = f(A) + f(B) \tag{4.3}$$

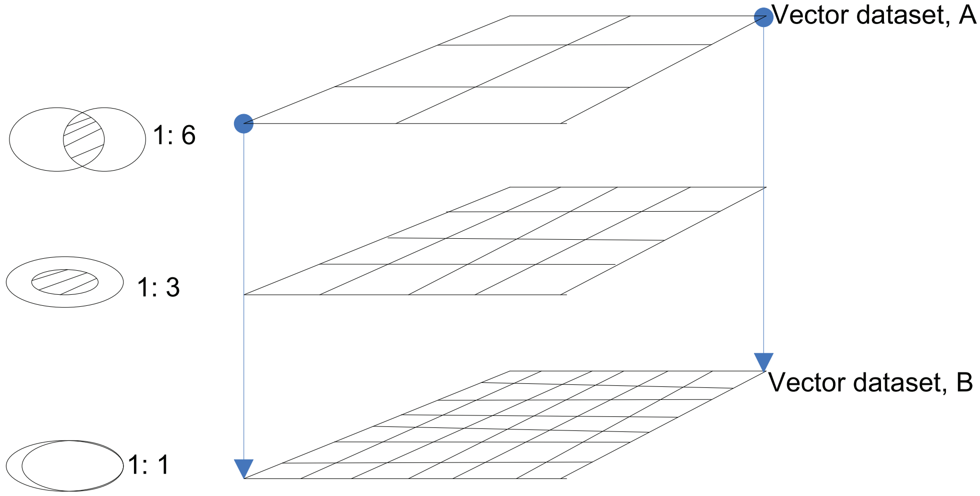

Figure 4.6: Process of finding a matching set of candidate in vector data of polygon dimension.

4.2.3 Validation Framework and similarity measures

To develop a framework for this validation process, we consider the use and extension of Tversky's Feature Contrast Model (FCM) introduced above. This is applied in measuring the degree of similarity between the outcome of two datasets from different origins. We take this model to be suitable for both per-object and per-pixel cases introduced in section 3.2 of chapter 3.

Letting IGBP dataset be A and UMD dataset be B, similarly to CORINE and LGN 4 land cover datasets. Figure 4.7 is referred to for the concept similarity of A and B. Then equation follows

$$Similarity(A, B) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)} \tag{4.4}$$

Similarity of A with respect to B based on features is expressed as a function (f) of three parameters:

- $f(A \cap B)$: Denotes features common to A and B i.e.,(A, B) in figure 4.7
- $f(A - B)$: This denotes features belonging to A and missing in B i.e.,$(A, \neg B)$ in figure 4.7 and

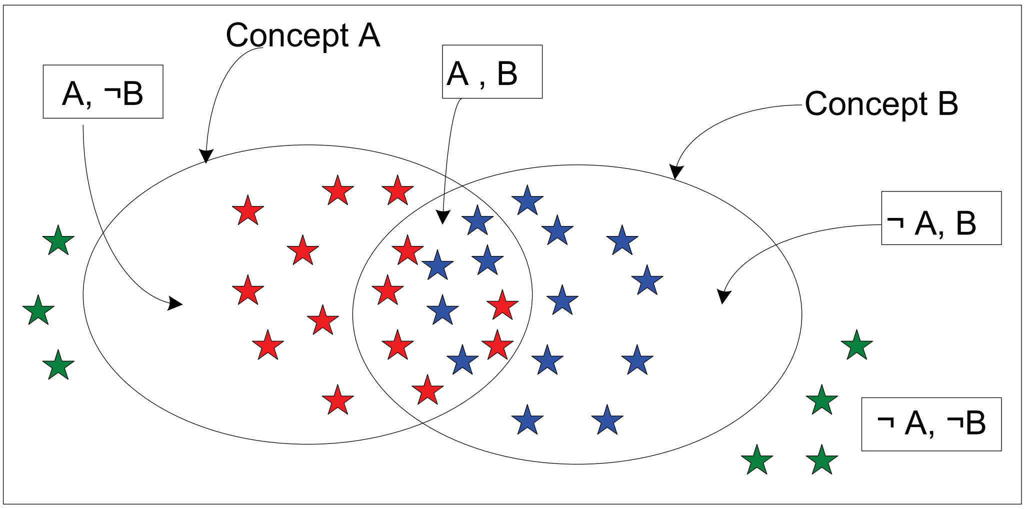

Figure 4.7: Concept similarity for concept A and concept B

- $f(B-A)$ denotes features that belong to B and missing in A i.e.,$(\neg A, B)$. $(A-B)$ and $(B-A)$ represent distinctive features in both datasets.
- α and β denotes the weight for $f(A-B)$ and $f(B-A)$ respectively, which are regarded as weights for two aspects of mismatch.

In this research study, similarity measure is symmetric which means $\alpha = \beta$ and we assume these constants to be 1. This setting of constants cannot be validated. This is because constants can only regulate common feature and distinctive features globally. In this validation, there is no direct use of semantic features but regard class labels(class encoding) of land cover as semantic features. Regions in one class is related to some semantic features, area being one of these semantic features.

For a single land cover assessment, the overall similarity $(sim(A, B))$ was introduced as in expression 4.5. This can be understood as the percentage of matched land cover among the total number of land cover in both dataset A and B.

$$Sim(A,B)_i = \frac{f(A_i \cap B_i)}{f(A_i \cap B_i) + \alpha f(A_i - B_i) + \beta f(B_i - A_i)} \tag{4.5}$$

From expression 4.5 i is the designated land cover and we consider $\alpha = \beta = 1$. In a cross table the diagonal cells are regarded as matched features i.e., $f(A_i \cap B_i)$

while the off-diagonal cells of the cross table are the mismatched features, i.e., $f(A_i - B_i)$ and $f(B_i - A_i)$.

Therefore, the overall similarity of per-pixel is designated as in expression 4.6 below

$$\begin{aligned} Sim(A,B)_i &= \tfrac{N(A_i \cap B_i)}{N(A_i \cap B_i)+N(A_i - B_i)+N(B_i - A_i)} \\ &= \tfrac{\mathbf{n}_{ii}}{\mathbf{n}_{ii}+(\mathbf{n}_{i+}-\mathbf{n}_{ii})+(\mathbf{n}_{+i}-\mathbf{n}_{ii})} \end{aligned} \tag{4.6}$$

where,

- N is a function of number of objects for objects and for the number of pixels.
- n is the actual number of objects.
- i is the land cover in question.

How many pixels for a land cover i, in dataset A agrees with those in dataset B? This is achieved by using the equation 4.7:

$$Sim(A,B)_i = \frac{N(A_i \cap B_i)}{N(A_i \cap B_i) + N(A_i - B_i)} \tag{4.7}$$

The overall similarity assessment obtained using the approach above are in terms of location or spatial extent. This is because computed results are from pixels with random locations. Applying similarity assessment procedure above, counting the number of pixels per land cover and using the size of the pixels to get the area. The overall similarity measure for the object (land cover) can be obtained by applying the formula above.

In respect to the comparison of two datasets (A and B), matrix function was used in ERDAS suite. This function outputs the coincidence values of the input data sets each of land cover classes. It produces a set of pixels for every coincidence of land cover in two datasets i.e., a long the diagonal. For the validation purposes, CORINE dataset, which is in vector form was converted to raster in order to be compared to LGN 4 dataset.

With the application of expression 4.6, using the number of pixels from ERDAS suite, the similarity of an extract of land cover classes designated by class codes (DN values) are shown in table 4.3 for IGBP and UMD dataset and table 4.4 for CORINE and LGN 4 dataset. Tables D.4 and D.5 in the appendix show full set of the results.

The outcome of the validation framework shown in tables 4.3 and 4.4 were obtained as a result of the common pixel values that was taken into consideration. The results obtained using this framework are considered to be the measure of matching features in terms of spatial extent or location. This is because pixels with random locations are used to compute the respective results. In this framework, the per-pixel measures of similarity for single land cover classes was integrated. These outcome are based on the information at per-pixel level.

Table 4.3: Extract of similarity judgement by validation framework for UMD and IGBP dataset

Class DN code(IGBP)	Class DN code (UMD)	similarity judgement(%)
0	0	98
1	1	62
2	2	47
3	3	47
4	4	59
5	5	78

Table 4.4: Extract of similarity judgement by validation framework for CORINE and LGN 4 dataset

Class code CORINE	Class code LGN4	Search concept (C_s)	Target concept (C_t)	Validation(%)
112	2	Discontinuous Urban Fabric	Maize	2
211	0	Non-Irrigated Arable Land	No class	-
242	5	Complex cultivation Patterns	Cereals	40
243	0	Land Principally Occupied By Agriculture	No class	0
231	8	Pastures	Greenhouses	36
121	9	Industrial/ Commercial Units	Orchards	25

4.3 A comparison of SIM-DL and validation framework similarity measures

The comparison of the semantic similarity from the SIM-DL and the validation framework is shown in tables 4.5 and D.6 for UMD and IGBP datasets. Tables 4.6 and D.7 depicts comparison of semantic similarity from SIM-DL and validation framework for CORINE and LGN 4 dataset. A complete list of

Table 4.5: Comparison of an extract of similarity results from SIM-DL and validation framework for UMD and IGBP dataset

Class DN code	SIM-DL(%)	validation(%)
0	75	98
1	80	62
2	80	47
3	77	47
4	77	59
5	49	78

Table 4.6: Comparison of an extract of similarity results from SIM-DL and validation framework for CORINE and LGN4 dataset

Class DN code CORINE	Class DN code LGN 4	SIM-DL(%)	validation(%)
112	2	2	2
211	0	-	-
242	5	1	40
243	0	-	-
231	8	3	36
121	9	0	25

the compared similarity results between SIM-DL and validation framework is shown in section D.5 of the appendix. The graphical representation of these similarity results are in section D.6i.e. figure D.2 for CORINE and LGN4 and figure D.3 for IGBP and UMD land cover datasets. From these results we can observe that there is no match in the similarity measures from using SIM-DL and our validation framework.

Due to different approach of operation for SIM-DL and validation framework, comparing them is a difficult role. SIM-DL is based on different representation language($ALCHQ$)and is based on semantic descriptions of concepts only. The validation framework is based on pixel values, where features (in terms of location) of real objects are considered. This makes it become more specific. Further, because of the differences in the granularity in CORINE and LGN 4 land cover classifications, few corresponding objects in terms of pixels can be observed. This is because, LGN 4 datasets has detailed land cover classification, than CORINE which has less detail. Because of the granularity difference, this causes the difference in the definitions, thus SIM-DL could not find the matching definition. The land cover classes in border area of the features in CORINE dataset may belong to a number of land cover classes in LGN 4. However, they may not fall within the corresponding land cover class in CORINE.

The land cover classification quality is of concern. The similarity measure about the limited area represented by single pixel might have been different from a similarity measure of a large area. In this large area, the pixels form only a part. This pixel-based approach is inadequate for assessing the similarity of land cover. This is because, spatial unit changes from an individual pixel to an individual land cover

4.4 Mapping representation and interpretation

The mapping elements are defined in [ZS06] as a 5-uple $\langle id, C_s, C_t, n, R\rangle$ where

- id is a unique identifier of the given mapping element (for the correspondence).
- C_s and C_t are the respective classes and their properties of the source,C_s and target,C_t concepts respectively, in two ontologies.
- n is a similarity measure(confidence measure)which depicts the strength that the correspondence under consideration holds. This measure belongs to an ordered set in the range of $(0, 1)$. This holds for the correspondence between the concepts C_s and C_t.
- R is a mapping relation between the entities C_s and C_t e.g., equivalence, subsumption that holds between C_s and C_t.

The relation R is defined based on the confidence measure and we categorize as in table 4.7

Table 4.7: Semantic similarity measure ($C_s$and C_t) and interpretation

Similarity measure	Similarity interpretation category
$sim(C_s,\ C_t)$ =100	Equivalent ($\equiv$)
$sim(C_s,\ C_t)$ =0	Not similar(disjoint to) ($\perp$)
$sim(C_s,\ C_t)\geq$ x (threshold)	subsume ($\supseteq$)

Referring to table 4.1 and table 4.7, according to SIM-DL, the similarity measure for the fact that equivalent relation holds between *Evergreen Needleleaf forest* in IGBP ontology and *Evergreen Needleleaf forest* in UMD ontology is 80%. Each class in an ontology has a unique name identifier, in this case we use id_{C_s} and id_{C_t} for search and target concept respectively. The mapping element is thus represented as follows:

$\langle id_{C_s,C_t}, EvergreenNeedleleafforest, C_s, EvergreenNeedleleafforest, C_t, 80, \equiv\rangle$

Taking semantic differences, we consider the semantics of land cover concepts to match, when the similarity measure is of certain value $(x\%)$ or better. The value x is the threshold. In the case of a $1:1$ or $1:many$ relationship, the following condition can hold: $sim(C_s, C_t) \geq x\%$. The value $x\%$ is influenced by the needs intended by the user and the application in question. This value is adapted with respect to the user's goals. Any category of mapping depend on a threshold value.

In an application context, such as the one for EA, similarity of two entities exceeding a certain specific threshold(x) are considered to be equivalent, thus consistent. Disjoint relation between the compared land cover classes can be used by EA as evidence for considering the concepts to be different. Disjointness relation between land cover concepts are used for knowledge base updating because it provides evidence of inconsistency. Equivalence and subsumption allows relevant data structures in the source ontologies to be traced.

4.5 Sensitivity of the mappings due to revision of concepts

The semantics of Land cover ontologies can change when the intention is to correct for errors in the semantics. It can also be for accommodation of current information about concept meaning and making an adjustment to representation of given domain. Can a change in the concept definition have an impact on the mappings? In this section we consider in an ontology revision of semantics of concepts as a change in components of an ontology. This revision can be addition or removal of categories, properties and axioms that make up a meaning of concept.

The impact that semantic revision has on the compiled land cover knowledge can vary. To start with we look at the effect of the modification of concepts definitions. In the rest of this section, C represents concepts before change and C' concept after the modification. The following are possible ways in which the meaning of concept C may relate to new concept C'.

1. meaning of concept does not change: $C \equiv C'$. This could be due to change being in the other part of the ontology.
2. meaning of land cover concept changes to becomes more general $C \subseteq C'$.
3. meaning of land cover concept becomes more specific $C' \subseteq C$
4. meaning of land cover concept changes such that there is no subsumption relationship between C and C'

Similar observation is made for a change in the relation before, R and after, R'. How these different changes influence the concept interpretation, can it also influence mappings? We need to find out this.

Taking c to be the set of all concepts and r the set of all relations. Further, $C \in c$ and $R \in r$, changing the meaning of concept C has similar impact on the interpretation of the mapping between two concepts definitions. Such that: $C \subseteq C' \Rightarrow sim(C_s, C_t)$. This is because a change in relation R has the same effect on the mapping. Therefore, a change that adds or removes the concepts or relation constructs may restrict or loosen the semantics that were originally present, hence change in mappings.

Revision of semantics of the land cover concepts in the ontology can be defined to be set of rules (R_c). This allows the modification of the meaning of concepts selected via the concepts that specify the domain of the application. A rule consists of a condition specifying the circumstance under which the rule is ignited [KRJ$^+$08]. It modifies semantics of concepts and the affected concepts in which the modification applies. Each revision as shown above either adds by intersection of superconcept to the influenced concepts or deletes it. These rules are represented as given in [KRJ+08].

$$R_c : condition \ \rightarrow \ (\pm \ modifying \ concepts, \ affected \ concepts). \qquad (4.8)$$

A change towards modification of meaning of search concept,C_s and the target concept,C_t can be quantified in terms of how many of the superconcept are affected as a result of this modification. If the modification of the semantics of concepts make the search and target concepts similar, the absolute value is positive if it becomes less similar it is negative. This impact as in the [KRJ+08]is measured from a range of 0 to 1. At 0, there is no change on meaning of concepts considered while at 1, there is change.

For instance in our case in table 3.3, three concepts are extracted from the IGBP land cover and examined the impact of modification with respect to their similarity. We introduced a set of rule applied to different concepts following expression 4.8.
$R_1 : Siltation \rightarrow \{waterbodies\}, (\neg Reservoir, \{Oceans, Seas, Lakes\})$
$R_2 : Wildfire \rightarrow \{DeciduousBroadleafforests\}, \{+Baresoil\}$
$R_3 : Storm \rightarrow \{Grasslands\} (+Woody, \{Herbaceous, TreeCanopy, Shrubs\}$

Since similarity is calculated based on representations, the representation of these semantics of concepts consists of the preceding rules. R_1 removes the class *reservoir* i.e., $\neg Reservoir$. Use of $-reservoir$ removes only class *reservoir* from the class definitions. However, it does not explicitly state that water bodies no longer consists of reservoir. Addition of $\neg Reservoir$ overwrites existing definitions and does not lead to a contradiction(i.e., unsatisfiable concept) [Jan08]. As a result of siltation, R_2 adds the class 'Baresoil' thus change in the meaning of 'Deciduous Broadleaf forests':

- Baresoil (Deciduous Broadleaf forests $\exists$ contains.Baresoil) and
- Woody (Grasslands $\exists$ contains.Woody) due to wildfire and storm respectively.

Table 4.8: Semantic similarity measure (C_s and C_t) before and after modification of the concept definition.

Sim(C$_s$, C_t)	**Similarity before**	**after**
(Waterbodies, Waterbodies)	75	60
(Deci. Broadleaf forests, Deci. Broadleaf forests)	77	86
(Grasslands, Grasslands)	80	89

Table 4.8 depicts the computed similarities using SIM-DL for the respective meaning of concepts before and after modification of definitions of the land cover concepts. These results are represented graphically the figure D.4 in section D.6. From these outcome, we can observe that similarity value changes with deletion and addition of the concept definition. This change affects the

shared superconcepts and the respective concept descriptors, thus change in the mapping itself. Therefore the question we asked above can be answered.

4.6 Reasoning with mappings between classes and properties

Ontology matching as defined in section 1.7, the mappings are expressed by some mapping rules which explain how the compared concepts correspond. Considering correspondences, data integration requires both correspondences between class correspondences (concepts) and correspondences between instances. In figure 3.1 it is the datatype similarity measurement. Instance mappings are needed in order to retrieve properties relevant for heterogeneity resolution in the knowledge base.

Based on the confidence measure levels obtained in section 4.4, an arbitrary threshold (x) determined by EA, a rule can be set up to express that two concepts are equivalent or different. In particular SWRL language *see section* 2.4.2 is used to develop these rules. The property *sameAs* is used to declare that two individuals are identical. Also *differentFrom* property gives an opposite of effect of *sameAs* [OCKT^{+}05].

By applying the semantic rules, land cover for instance of type *Evergreen Needleleaf forests* in table 4.1 is furnished with additional datatype properties e.g., *hasLocation*, *hasLatitude*, *hasLongitude*, *hasSimilarityValue* e.t.c. The similarity value is adapted to be part of the properties of the concept in question. Having the class definitions on hand, a rule can be set. Using the similarity measure value, it can be inferred that two compared concepts, for example, a and b are the same individuals, thus are equal. A rule is then written to reflect $sameAs\ (a, b)$ and if they differ a rule is written too for *differentFrom (a, b)*.

To assert a rule to show the semantic relation categories between any two compared concepts, then the following rules qualify:

- Equivalent

$LandCover(?x, ?w) \wedge comparedTo(?x, ?w) \wedge hasSimilarity(?x, ?a) \wedge$
$swrlb : equal(?a, value) \rightarrow Equivalent(?x, ?w)$

- Not Similar (disjoint)

$LandCover(?x, ?w) \wedge comparedTo(?x, ?w) \wedge hasSimilarity(?x, ?a) \wedge$
$swrlb : notEqual(?a, value) \rightarrow DifferentFrom(?x, ?w)$

- Subsumes relation

The following SWRL rule verifies that a land cover (evergreen needle leaf forest) has similarity value greater or equal to a threshold value (x) that suits the needs of the user.

$LandCover(?x, ?w) \wedge comparedTo(?x, ?w) \wedge hasSimilarity(?x, ?a) \wedge$
$swrlb : greaterThanOrEqual(?a, value) \rightarrow someRelation(?x, ?w)$.

Further, SWRL rules can be used to delegate the comparison of any two instances according to similarity measures from SIM-DL. The code snippet below, gives an additional example for the comparison of two instances of type 'Forest' in rule syntax.

$(?F1, typeforest : forest) \wedge (?F2, typeforest : forest) \wedge similar(?F1, ?F2) \rightarrow$
$sameAs(?F1, ?F2)$

The mapping relation categories shown above, SWRL rules can be used to determines if the compared land cover instances are qualified for the respective relation i.e., equivalent, disjoint, or the subsumption relation.

The semantic similarity between two instances is calculated based on the semantic similarity of object properties i.e. concept *see section* 4.1.1

4.7 Summary

Similarity measurement using SIM-DL, the validation of the similarity results using feature contrast model(FCM) and discussion of the results is given. The representation of the mappings and the sensitivity of the mappings due to manipulation of the semantics of concept is also explored. The next chapter is about how to use the mappings obtained so far in retrieving geoinformation from the knowledge base. This is with assumption that the semantic similarity measures satisfy the requirements of the EA based on the decision made as shown in figure 3.1.

Chapter 5

Using ontology mappings: Retrieving Geoinformation

In the preceding chapters, an OWL-DL encoded ontology was designed to formalize relationships between land cover concepts. Similarity between land cover concepts from two different sources was measured. This enables the mapping of semantically heterogeneous land cover datasets. More so, rules were as well used to represent the mappings and express similarity between instances based on the SWRL built-ins.

Nevertheless, it is impossible to make explication of all the knowledge as ontologies. For example, spatial relations among land cover locations and distribution can be difficult. The implicit spatial information from the existing knowledge base can be deduced by rules. These rules are created for the retrieval and making inference to land cover information.

5.1 Distinction of approaches to information retrieval from knowledge base

The approaches to knowledge retrieval are many but varied in level of operations. The following is a summary of distinctive knowledge retrieval approaches mentioned in sections 2.4.2 and 2.6

1. SPARQL: via Protégé's plug-in or Jena's ARQ, one can pass SQL like queries over an OWL ontology(OWL-DL) and have the respective result passed back based on pattern matching. ARQ is a query engine for Jena that supports the SPARQL RDF Query language [SP07].

2. OWL-API or Jena API: Provides a way to get a class for example and list its subclasses, restrictions, individuals and so forth.

3. Pellet (or FACT for Example): Provides 'Logical Query Capabilities' in the sense that 'Equivalent' classes or 'Complete' classes can be classified into an existing asserted model.

4. Using SWRL rules and SQWRL(Semantic Query Enhanced Web Rule language): Which is claimed in literature to be quite expressive. SWRL re-

strict value ranges and can be used to reason about the data in order to populate the ontology. SWRL has very close connection with OWL-DL and the fact that it has well defined semantics.

In this research study, based on the facts above, we found SWRL being attractive and thus we opted for it as a rule language and SQWRL as query language. These are discussed in section 2.4.2.

5.2 Querying geoinformation using semantic rules

Conceptual search applies translations and semantic rules to convert the query with special handle of ontologies. Consequently, EA has to integrate information from ontologies from various sources to achieve their objective. Referring to figure 3.1 EA is confident with the mappings obtained and may need to use it retrieving information from knowledge base.

SWRL rules in section 2.4.2 have an implication between antecedent (body) and consequent (head). The intended meaning can be interpreted as : if the conditions in the antecedent hold, the given condition in the consequent must also hold. SWRL has inference capabilities through the SWRL rules. Rules are edited in SWRLTab *see section* 2.4.2 and are considered as instance data in Protégé. Protégé together with SWRLTab do not support SWRL rule execution. SWRL rules require availability of rule engines to be executed. A common rule engine is JESS and others are Drools, Algernon and Bossam. These rule engines perform reasoning using a set of rules and a set of facts being inputs. New facts that are inferred act as inputs for firing more rules. The land cover ontology base application requires the ability to extract information in this ontology. For knowledge extraction, a query language SQRWRL *see section* 2.4.2 supports querying of ontologies.

5.2.1 Limitations of OWL and SWRL

The limitations that both SWRL and OWL have are as follows considering the discussion in section 2.4.2

- OWL and SWRL provide useful standards for expressing concepts and instances of the land cover classification systems ontology.
- Open world assumption of OWL and SWRL preclude queries for negation.
- SWRL makes a very limited assumptions about how rules are executed and specifically does not define mechanism for recursive application of rules i.e., it involves incomplete inference.
- The number of built-ins and the facts that they do not have specified output variables complicate the implementation. Working with values that change over time is difficult because of RDF's monotonicity assumption– once a triple is asserted, it can neither be retracted nor changed.

5.2.2 Limitations of OWL-DL

The reasoning procedures in Description Logics are decidable i.e they terminate both from positive and negative responses. This makes them less expressive compared to full Horn like rules (SWRL) which are more expressive, though, it leads to undecidability [OCKT+05]. For example, it is possible to state that all instances of *Forest* concept must have at least one part that is *Road* in expression 5.1. It is hard to express that if Land use/cover (Forest),f, and a Road,r, have the $code, i$, thus , f has r as a part through relation *hasPart* since it consists of three variables.

$$Forest \sqsubseteq \exists hasPart.Road \tag{5.1}$$

$$Forest(?a) \ \wedge Road(?r) \ \wedge code(?f, i) \ \wedge id(?r, i) \rightarrow hasPart(?f, ?r) \tag{5.2}$$

Based on the relationship in expression 5.1, inferences cannot be made to infer new knowledge between instances from the existing data.

Table 5.1: Properties of the intended site model

Category properties	Site	Distance_from Residential	RoadType	Location	Soil Type	Slope
Asserted properties	hasSpecifications	isAtDistanceFrom	isConnectedTo	isAtLocation haslatitude hasLongitude	hasSoil	hasSlope
Inferred properties	Potential_site					

Table 5.2: Factors making a location good site for siting vacation cabin

Slope	Residential	RoadType	Location	Soil
Gentle	$\geq 1500m$ away	minor	EA region	Sandy

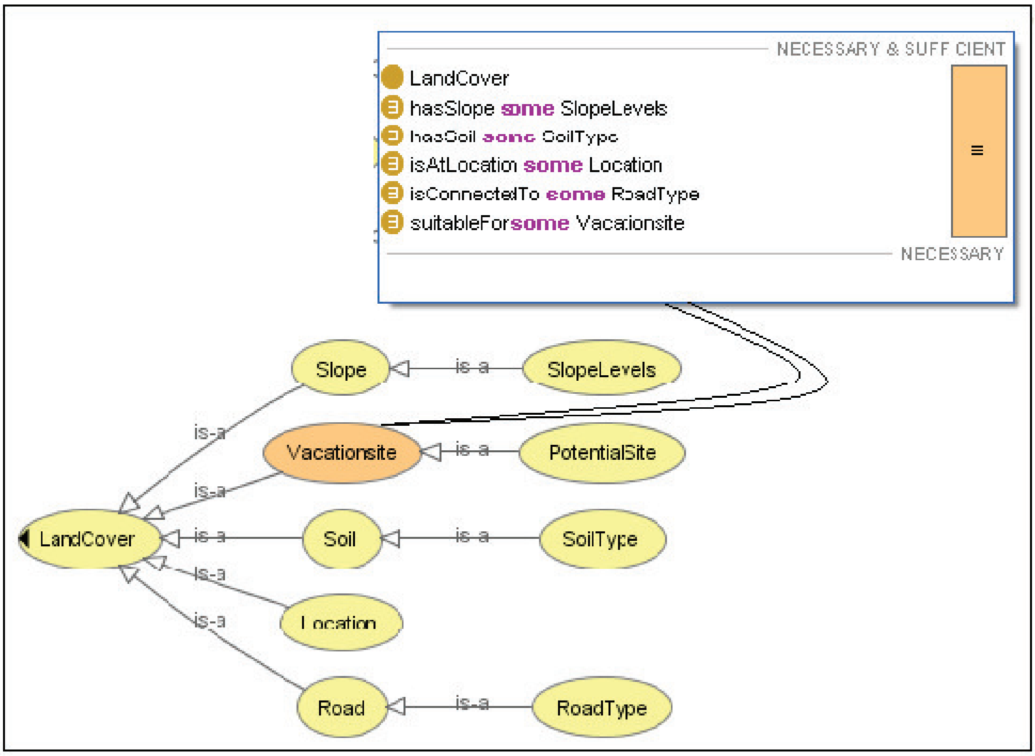

Figure 5.1: Knowledge representation of the land cover concepts of the vacation site using ontology. The properties of the vacation site are in the box.

Example

The EA may need to issue queries that involve geospatial operations. This requires processing or the combination of results of multiple data sources. These queries have the ability to infer new information from disparate data sources. For instance, a query such as :

Find topographic area that is potentially good for siting a vacation cabin?

Locations of topographic areas and information about good site factors are provided in tables 5.2 and 5.1. The relevant land cover concepts and properties for sitting the vacation cabin are collected as shown in table 5.1. Table 5.2 shows the specification that the EA use.

Assuming that the confidence measure (similarity value) between the compared concepts in section 4.1 satisfies the intended application for the EA. Semantic rules can then be applied to determine potential zones for sitting vacation cabin using specifications given in table 5.2. This task involves use of more than one data source and this necessitates geospatial processing to determine if a topographic area is in a potential zone. Assuming that the EA region is at location $(Latitude : 472573m, Longitude : 253017m)$.

Listing of properties beyond each land cover concept is used to describe the land cover concepts characteristics. Taking into account the values inside the properties, we can informally have two categories of properties *asserted* and *inferred* properties in table 5.1. Asserted properties allow to input known facts (explicit knowledge). Inferred properties are unapparent facts (implicit knowledge) leading to inference using known facts. In addition to ontology created in figure 5.1, dealing with semantic heterogeneity, rules are developed in section 5.3 for the inference of domain knowledge (land cover information), spatial and thematic retrieval.

5.3 Retrieval of geoinformation using rules and facts

SWRL rule is used to transcribe and refine the ontology which consists of facts and rules. Using it in land cover application facilitates knowledge interchange with other tools and applications. The hierarchy 'is-a' relation of the ontology is represented as facts as in figure 5.1. It represents hierarchical relationship between land cover. Based on knowledge about the environment information, location facts of the cabin site can be generated using rules.

To choose the preference site for the location of the cabin considering the specifications in table 5.2. The following rule is created , converted to inferred facts and transferred to OWL-model using a rule engine (JESS). Figure 5.2 shows the rule created in SWRLTab and conversion to inferred facts in JESS.

1. $CabinSitting(?site) \wedge$
2. $hasSlope(?site, ?gentle) \wedge$
3. $isAtDistanceFrom(?site, ?distance) \wedge$
4. $swrlb : greaterThanOrEqual(?distance, 1.5) \wedge$
5. $isConnectedTo(?site, ?minorRoad) \wedge hasSoil(?site, ?sandy) \wedge$
6. $isAtLocation(?site, ?location) \wedge$
7. $hasLatitude(?location, ?latitude) \wedge$
8. $hasLongitude(?location, ?longitude) \wedge$
9. $swrlb : add(?distLatitude, 472573) \wedge$
10. $swrlb : add(?distLongitude, 253017) \wedge$
11. $swrlb : greaterThanOrEqual(?distLatitude, 1500) \wedge$
12. $swrlb : greaterThanOrEqual(?distlongitude, 1500) \wedge$

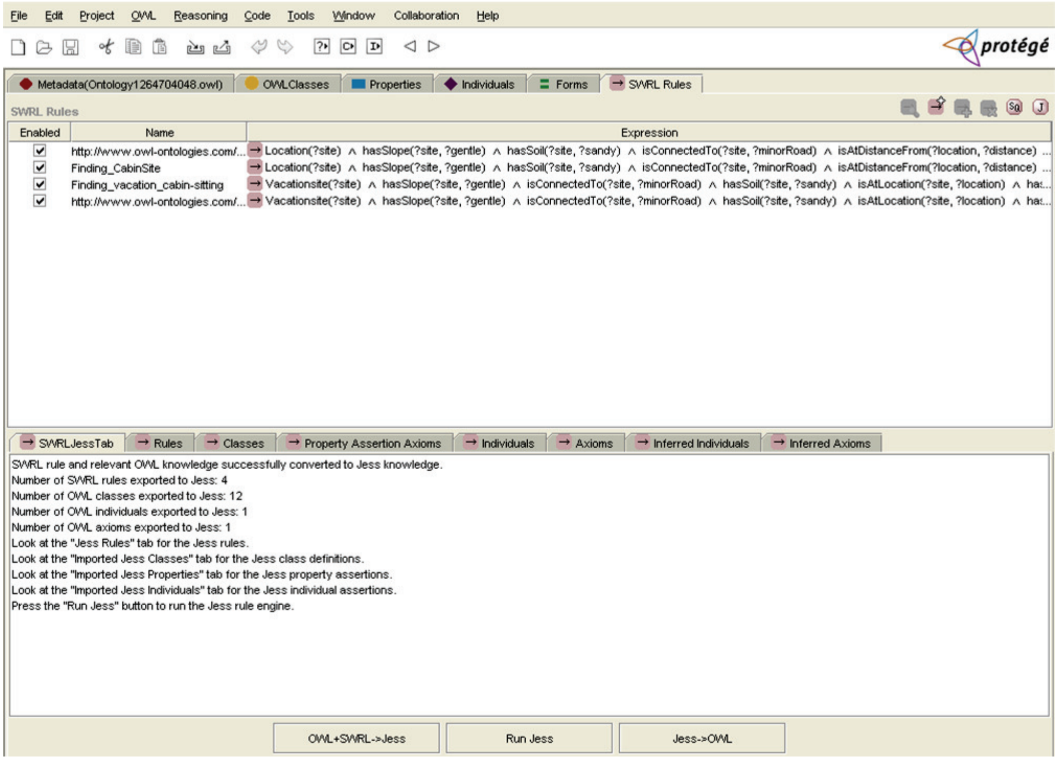

Figure 5.2: Editing of rules in SWRLTab using SWRL language and running them in rule engine(JESS). These lead to inferred facts which are transferred back to knowledge base.

13. $\rightarrow Potential_site(?site)$

From this rule, line (1) restricts the instances at which this rule is applied to one in the class *CabinSitting*. The variable name *?site* is assigned to them. Other lines constrains the matching instances to those having the specified qualities. This leads to meeting the EA preference shown in line (13). The execution of these rule is that if a potential site is at current location ($Latitude : 253017m, Longitude : 472573m$) but $\geq 1500m$. In this clause, the values allowing deviations in latitude and longitude are sufficient to infer about the potential site for cabin. This rule has been used to find the semantic relationships thus integrating land cover information based on the given facts.

To retrieve the information stored in the land cover OWL-DL ontology, SQWRL built-in library is used in the constructed rule. The following SQWRL query retrieves the vacation site instances and their values in the knowledge base.

1. $Vacationsite(?site) \wedge isAtPosition(?site, ?latitude) \wedge$

2. $isAtPosition(?site, ?longitude) \wedge$

3. $swrlb : equal(?longitude, 253017) \wedge swrlb : equal(?latitude, 472573) \wedge$

4. $isAtSlope(?site, ?gentle) \wedge$
5. $hasSoilType(?site, ?sandy) \wedge isLinkedTo(?site, ?minorRoad) \wedge$
6. $isFarFrom(?site, ?distance) \wedge$
7. $swrlb : greaterThanOrEqual(?distance, 1500) \rightarrow$
8. $sqwrl : select(?site, ?latitude, ?longitude, ?distance, ?gentle, ?sandy, ?minorRoad)$

Line (8) is the query that gets the information from the knowledge base. The results of the retrieved information from the knowledge model are shown below. Figure 5.3 shows the retrieved information and the respective values of the cabin site.

1. $?site, ?latitude, ?longitude, ?distance, ?gentle, ?sandy, ?minorRoad$
2. $CabinSite, 472573, 253017, 1500,$ "$gentle$", "$sandy$", "$minorRoad$"

From the retrieved information, it is evident that SWRL and SQWRL allow the knowledge-level encoding of geographical rules and queries. This is by use of concepts from the ontology.

5.4 Summary

Earlier on, it was difficult for the EA to establish the potential site from the ontology. This is because OWL-DL has limitations *see section* 5.2.2. These built-in operators in SWRL, make it much easier to describe the potential land cover concept intended by the user (EA). The instance (CabinSite) of the land cover concept (VacationSite) and its values have been retrieved from the ontology using rules. This process shows the expressivity power that rules have.

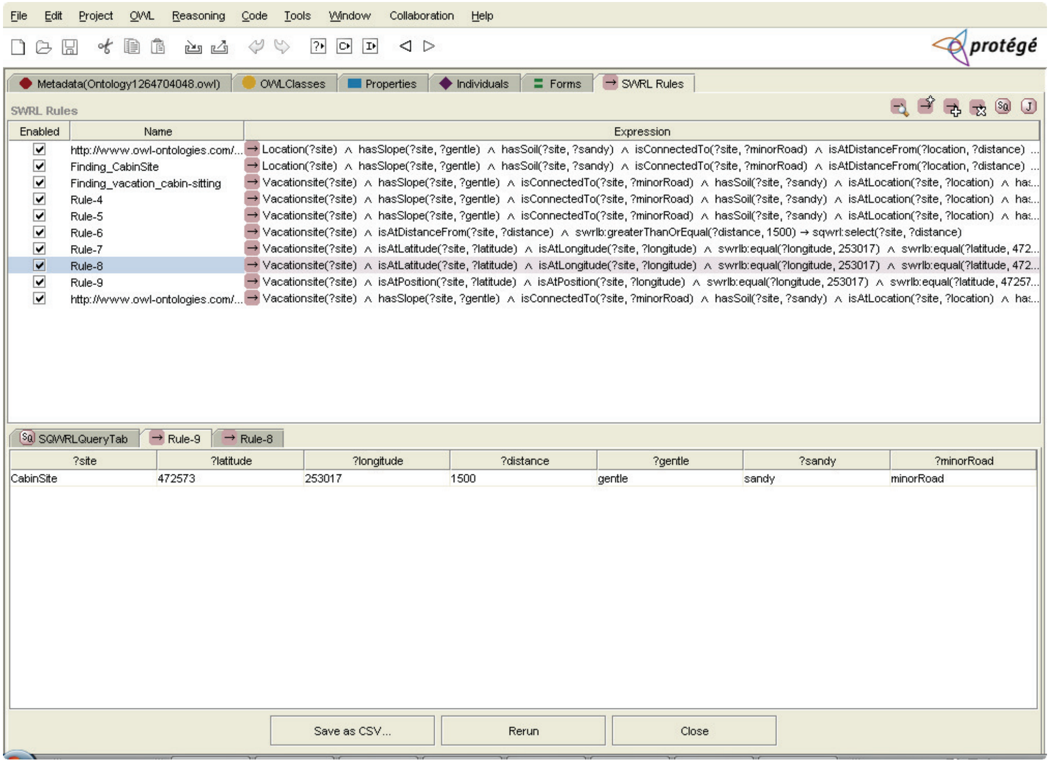

Figure 5.3: The output of querying the ontology to retrieve the instance CabinSite and its values. These information belong to the land cover concept, vacation site. This shows the power of rules in retrieving instances from the ontology.

Chapter 6

Results and Discussions

6.1 Introduction

This chapter discusses the results that was derived. This research study led into the creation of application specific ontologies and evaluation of semantic similarity measurement as an ontology mapping method. It also identified the impact of revising the concept definition and exploited ontology mapping with rules in defining the the relation that holds between two compared instances and in an application example. The following section is a discussion for each result derived from this research study. This sets the pace for discussion based on earlier contribution.

6.2 Discussion on the results

Result 1: Establishment of mappings between concepts from two different ontologies

The research question (1) is related to the development of ontologies and similarity measurements using SIM-DL. This has been discussed thoroughly and answered in sections 3.5 and 4.1. However, this is given in detail in section 4.3 after its validation using the framework detailed in section 4.2. The validation framework serves as a verification of the similarity measurements obtained using SIM-DL. This framework is an extension of Tversky's set-theoretic similarity i.e, Feature Contrast Model (FCM) [Amo77] introduced in section 4.2. It provides the similarities of the matching set. This is applied in datasets of both similar granularity and dissimilar granularity. With regard to the kind of dataset, in our case raster, the framework is applied. For the case of vector datasets of polygon dimension, an approach for finding a matching set is proposed which is open for future research.

In assessing the success level of SIM-DL, it is observed that the similarity measures compared to our framework do not match. This is evident in figures D.2 and D.3. The land cover categories may not have been correctly classified and the class descriptions not fully defined. The semantic deficiencies because of unsupervised analysis and extraction of land cover definitions could have also contributed to this mismatch. Limited area is represented by a single pixel, thus is different from similarity of a large area. In a larger area, pixel forms only a part. Therefore, single pixel is inadequate. This is because spatial unit changes from an individual pixel to a multi-pixel region. This can also contribute to the difference.

The language limitation at which the concepts are represented in SIM-DL restricts the level of describing the concepts. Since the natural language description is full of information, SIM-DL is limited. It does not support the datatype properties and functional properties. In addition, it does not support nested value restriction. This is because, the level of expressivity becomes complex, thus not supported by SIM-DL reasoner. The similarity measure can satisfy at the concept level which may or may not be useful for an application. This is because at the concept level, less semantic information about the land cover can be achieved unlike at the instance level. For example, most specific concept i.e a concept that has all individuals as instances and the most specific description that satisfy the property of individuals. This leads to satisfaction about the land cover data thus fit for use. The integration of measures for role-based constructors and role intersection, the similarity measure using SIM-DL meets the needs of finding how similar the concepts are and provides the solution to the challenge of establishing whether the compared concepts represent a similar notion.

This research study has revealed that additional object measures need to be taken into account. These measures consist of geometric levels which is the actual set of pixels. The pixel value (class value) represent semantic level. To have an optimal semantic information, available content and context information has to be applied.

Result 2: Definition of mappings between instances

At the concept level, mappings may or may not satisfy. This is because less semantic information or rather the information is abstract. This leads to achieving of generalized mappings from the object properties. For better improvement, instance information which are datatype properties are used. Defining mappings between two instances is not trivial. However, the solution can be the use of rule based approach. The rules used are created using SWRL language discussed in section 2.4.2 and its application is utilized in section 4.6. The expressivity of SWRL rules has made it possible in defining the mappings between instances. SWRL built-ins are used to delegate the comparison of two instances according to the similarity measure obtained using SIM-DL. The mapping relation based on the similarity value and a defined threshold, the compared instances are categorized as equivalent, disjoint and subsumption relation. The identified possible instances are compared based on their entire set of properties. The semantic similarity between two corresponding instances is based on the syntactic similarity of the datatype properties and the semantic similarity of the object properties. Applications of these mapping relations in an example case is discussed in section 4.6.

Result 3: How are the mappings represented?

Representations of mappings can be considered from the viewpoint of the expressiveness in terms of operators and functions. This mapping representation has been explored in section 4.4. The mapping element is described by 5-uples shown in section 4.4. From a set of two compared concepts from two ontologies concerning a particular application. The relations are defined in section 4.4 to be equivalent, disjoint and subsumption relation. These relation categories are interpretations of the overall similarity values for the compared concepts. If the similarity between two entities is higher than a specified threshold, the compared entities are considered to be equivalent. If the similarity is less than the given threshold, the mapping representation is disjoint.

However, the challenge that still prevails is how can these representations be made self-explanatory to the user? How is the threshold determined? These questions can be answered in the future. Having an adequate mapping representation can allow plausible use. But these mappings can be interpreted differently by different people. This is with respect to the users goal. We can observe that mapping relation only depends on the threshold and each similarity value is separately considered. A further research is needed to have a mapping representation that is essential for the management, sharing and reuse of these mappings. This would support a common understanding between users.

Result 4: How can the resulting mapping be put into an application and maintained

This research question is discussed in detail in section 4.5. Conditions in form of rules are formulated in an application example. These rules involve the addition or removal of some definitions to the existing concept definitions. The rules used are aimed at finding out how a change in concept meaning can influence the mappings, hence sensitivity. The formulated rules are based on the ontology and on applying these rules, a context is specified. This context is defined by the semantic description of the concept. The similarity is then measured between the compared concepts and compared to similarity before the modification. From the modification of concept meaning, it can be observed that SIM-DL is sensitive to a number of features in finding corresponding elements, hence context sensitive.

Result 5: How to exploit ontology mappings in an application?

Using OWL-DL has proofed to provide an abstract view in the land cover concept level. Hence semantic relationships among the instances are difficult to discover. SWRL rules therefore have proofed to provide procedural knowledge power in discovering semantic relationships among the land cover instances. The SWRL rules together with EA knowledge model are used to locate the site for building a vacation cabin. This is done through integration of the land cover concepts and the site specifications from the EA. This has depicted the inference of new and implicit knowledge. However, some implicit spatial relationships such as location point-in-polygon land cover or polygon-in -polygon was not possible to deduce from the spatial location of land cover information. In this situation, if the EA wants to determine the land cover region at which the vacation cabin is to be located in. This follows our observation about the limitations of SWRL indicated in section 5.2.1.

Following the ontology developed to encode the land cover (vacationsite) concept, instance with the respective values of the site has been retrieved using rules. SQWRL has been used in this case to query the knowledge base. Knowledge level encoding of complex location rules and queries using concepts from land cover ontology is thus possible with rules. These rules can freely mix with concepts also with ontology concepts representing the domain entities of the user(EA). Attribute information such as distance, slope etc. in this application are important for information retrieval in the knowledge base.

6.3 Conclusions and Recommendations

To this end, this research study has answered all the formulated research questions. Accordingly, all the objectives of the research study have been met. The conclusions of this research study are drawn with respect to each research question as defined in section 1.4 of chapter 1.

6.3.1 Conclusions

How can similarities be established between two ontologies?

We developed ontologies and applied the semi-automatic mapping technique by use of similarity measurement approach. In this respect, the concepts of the ontology modelling principles by [NM01] were adapted, similarity measurement technique by [JKS$^+$07] and validation framework using the extension of Tversky's Feature Contrast Model (FCM) [Amo77].

Semantic data extraction and definition of concepts formed the basis of similarity measurement. The similarity measurement using SIM-DL served as a platform where the measures between search concept and target concept are calculated. The output of this similarity measurement facilitates the mapping relations thus integration and retrieval of geoinformation. To have an effective data integration, the fundamental step of mapping ontologies is taken into account. Thus necessary mappings are made between the ontologies before the decision is made to carry out the integration task. The similarity measures are evaluated by finding a matching pair using the validation framework. The validation output compared to the SIM-DL results,though there is mismatch on comparing, we can still make an observation that semantic similarity measurement outcome using SIM-DL are plausible. This qualifies similarity measurement as an ontology mapping method to be essential in finding corresponding concepts prior to integration. It also helps in establishing whether any compared concepts based on the object properties, represent a common notion.

How can the mappings be defined between two instances and their representation?

Definition of mappings between two instances have been defined using SWRL rules. These rules because of their power of expressivity have made it possible for datatype properties to be compared according to similarity measure and the defined threshold. The relation between instances are represented either as equivalent or disjoint (differentFrom) or subsumption relation using SWRL rules. Rules can therefore be applied to define mappings and represent mappings between instances. They help in describing the relation between the compared instances, similarity measure being the basis of this definition.

How do the similarity measurement methods specify representation of mappings?

Based on this research study, the mappings between the compared spatial concepts are represented in 5-uples. The mapping relation being one of the elements in the 5-uples is categorized as equivalent, disjoint and subsumption relation. These relations are determined by the similarity measure and the defined threshold.

How can the resulting mapping be put into application and maintained?

From this research study, we observed that changes in the concept meaning with time or correction of errors cause the mappings between the concepts to changes. This shows similarity measurement techniques must be sensitive to a number of ontology characteristics to find corresponding elements, thus mappings.

How can the mappings be exploited in an application?

In this case, SWRL rules and ontologies have been used to retrieve the geoinformation. This has made it possible for the ontology mappings to be applied in geoinformation retrieval using facts and the rules. SWRL rules can therefore be used to provide procedural knowledge power in enhancing limitations of the ontology inference especially in the discovery of the semantic relationships among instances. The SWRL rules are utilized together with ontologies in the finding of a location to meet the user preference and in the retrieval of instances from the knowledge base. Based on the observations made on the use of the rules and ontologies in this application example, we can deduce the following. Known facts can be updated as ontological knowledge by the user and the rules can provide a base for describing how to meet the user preferences. However, SWRL as it is now does not offer everything that is necessary to support this use case because of limited built-ins.

This research study has demonstrated how (semi)automatic ontology mapping can be used to determine similarity between concepts and instances from different ontologies. Further, it has shown how to establish whether the compared concepts represent similar notions. This approach in this research study has taken into account the characteristics of the concepts (object properties). These properties can help in solving different issues in the area of semantics. Ontologies being one way of modelling application and service semantics, it allows mapping specifications. This can potentially overcome the problem of integration and retrieval of geoinformation. The need for ontology mapping is on. This is in line with the growth in the number and diversity of geographical applications and services. We have seen that, semantic similarity relies on representation. That is whether and to what degree one concept is similar to another depends on what is said about both concepts.

The use of similarity based approach in ontology mapping generates inference and categorizes objects into kinds either when we do not know exactly what properties are relevant or when we cannot separate an object into separate properties. It is evident that similarity can be used as a default measure to reason about the semantic geoinformation.

However, the results within this research study are based on a limited number of case studies and semantic similarity measurement methods. We cannot claim that semantic similarity measurement using SIM-DL has been fully achieved. This approach is a starting point for the geoinformation integration.

6.3.2 Recommendations and Future Work

As we worked further on developing ontology and using similarity measure tool (SIM-DL), we noticed that similarity reasoner takes a longer time computing over a larger ontology. However, for simple ontology, the similarity computation time is very short. In addition, SIM-DL does not support the data type properties in case of individuals and more expressive constructors, thus making logic go beyond $ALCHQ$. These issues need to be addressed in future to overcome these limitations. In addition, other similarity matchers such as OWL Lite Alignment(OLA) algorithm which is claimed to be a very sophisticated matcher can be used in the future. Unfortunately, it is in perpetual re-engineering at the moment. Another validation approach such as finding a matching pair in vector dataset of polygon dimension can be looked at in the future.

Other case studies such as location based services and mobile mapping can be used in the future to assess the semantic similarity. This will ensure semantic data quality for sharing between different geographical information communities. This requires the exploitation of Web Feature Services(WFS) in the future by using ontology and SWRL rules in imposing logical constraints. WFS can solve some problems such as different understanding of similar features apart from feature attributes, mistakes in the attribute values and in geometry of features.

The implementation in this research study shows that key constraints and inferences of land cover classification concepts cannot be expressed in OWL and SWRL due to fundamental limits of these rule languages. Semantic matching rules can as well be used to describe relations using properties. However, this semantic matching rules only deal with domain knowledge for understanding of the land cover data. Other reasoning rules that cover spatial relation operators and spatial operations such as RuleML can be explored in the future. For instance, spatial rules can be used to determine the directional, topological and metric relations between geospatial components. An example could be how to use the rules in evaluating the direction and distance from the location of a cabin site to a swimming pool.

In addition, how can these spatial operation rules be used to generate new concepts and instances from existing environmental information. This demands for an integration of this approach with other techniques such as OO jDREW reasoning engine and GeoSWRL [KD07]. Further, some spatial operations such as spatial union and spatial intersection from existing spatial objects can generate new spatial objects using rules can be looked at in future.

Appendix A

Introduction to Terminologies

A.1 Summmary of the terminologies

Table A.1: Summary of the terminologies

Summary of the terminologies			
Ontology Approach	**Input**	**Output**	**Description of the output**
Ontology Mapping	Two (heterogeneous) ontologies A and B covering similar domain	Mapping relation of similar concepts or instances and degree of confidence based on data instances and statistics. Some kind of link between the two ontologies.	Maintains semantic consistency and coherence of the original ontologies while comparing and mapping across different ontologies.
Ontology Integration	Two (heterogeneous) ontologies A and B covering similar domains	An altered target ontology,say B, if A is source ontology	An integrated ontology is developed
Ontology Alignment	Two or more (heterogeneous) ontologies (and B) over the same domain, where heterogeneity has been resolved	Set of corresponding concepts or instances,degree of similarity	Maintains semantic consistency and coherence of the original ontology while comparing and mapping across different ontologies.
Ontology Merging	Two or more (heterogeneous) ontologies (and B) covering highly overlapping or identical domain	New ontology $X$is developed	An integrated ontology X is developed
Ontology Matching	Two heterogeneous ontologies	Alignment(set of correspondences)	It's a set of correspondence (pairs) or similarity matrix.

Appendix B

Tools for Geo-semantic Modelling and Mapping

B.1 Survey of ontology editing and mapping tools

B.1.1 Ontology editing tools

Table B.1: Examples of ontology editing tools

Semantic Web tools					
	Protégé-2000	OilEd	OntoEdit	WebODE	Ontolingua
Developers	SMI	University of Manchester	Ontoprise	UPM	KSL
Pricing policy	Open source	Free ware	Free ware and licensed	Free web access license	Free web access
Extensibility	Plugins	No	Plugins	Plugins	None
Inference services	FaCT, Pellet, RacerPro	FaCT	OntoBroker	Prolog	none
Usability graphical taxonomy	Yes	No	No	Yes	Yes

B.1.2 Ontology mapping tools

B.2 Description Logic reasoners

Table B.2: Representatives of ontology mapping tools

Ontology Mapping Tool	Mapping approach	Input type	Output type	Automation level	Availability
Protégé prompt suite	Heuristic based on lexicals and structure	OWL and RDF based ontologies	Merged ontology with suggestions to guide the userin creation of merged/ aligned ontology	semi-automatic approach to perform ontology merging and alignment.	Available as plug-in in Protégé-2000
Chimaera	Heuristic based(only within the ontolingua environment)	Multiple ontologies. Use different concepts	Merged ontology with suggestions	Semi-automated and guides the user	Available as part of ontolingua environment
GLUE	employs machine learning techniques to find mappings	Taxonomic structure of ontologies.	Mapping between concepts	Automated except when there is need for user verification and correction of mappings	Not available
SIM-DL	Machine learning of description logics	Two or more concepts or ontologies	delivers similarity values and rankings (SR).	automated	Free and open source

Table B.3: Description logics reasoning engines

Reasoners					
Reasoner	**Test version**	**Logic**	**Implementation**	**License**	**Comments**
Racer	1.7.23 and 1.7.24	$SHRIQ(D)$	CommonLISP	free license	discontinued after 1.7.24(authors went commercial with RacerPro)
RacerPro	1.9	$SHRIQ(D)$	LISP	commercial	DIG-only interface is free but not as flexible as the original RACER
FaCT++	1.1.3	$SHOIQ(D)$	C++	GPL	difficulties with large scale hierarchies
Pellet	1.5	$SHROIQ(D)$	Java	MIT	Full support of OWL 1.1 specification, overall performance on par with RacerPro, often faster with the higher complexity description logics
SIM-DL	beta2.3.1e	$ALCHQ$	Java	GPL	Similarity reasoner for various description logics from ALC family upto $ALCHQ$. Used via DIG or Protégé plug-in called Sim-cat.

Appendix C

Use Case Land Cover/use Categories

C.1 Land use/cover agricultural area category types

Table C.1: Land use/cover agricultural area category types

Land use/cover categories		
Classification system	Code	Agricultural area Category
CORINE	2.1.1	Non-irrigated arable land
	2.1.2	Permanently irrigated land
	2.1.3	Rice fields
	2.2.1	Vineyards
	2.2.2	Fruit trees and berry plantations
	2.2.3	Olive groves
	2.3.1	Pastures
	2.4.1	Annual crops associated with permanent crops
	2.4.2	Complex cultivations
	2.4.3	Land principally occupied by agriculture, with significant areas of natural vegetation
	2.4.4	Agro-forestry
LGN 4	1	Pastures
	2	Maize
	3	Potatoe
	4	Beet
	5	Cereals
	6	Other agricultural crops
	8	Greenhouses
	9	Orchads
	10	Flower bulbs

Table C.2: Land use/cover Artificial surfaces(CORINE)/Urban area(LGN5)category types

Land use/cover categories		
Classification system	Code	Artificial surfaces/Urban area Category
CORINE	1.1.1	Continuous urban fabric
	1.1.2	Discontinuous urban fabric
	1.2.1	Industrial or commercial units
	1.2.2	Road and rail networks and associated land
	1.2.3	Port areas
	1.2.4	Airports
	1.3.1	Mineral extraction sites
	1.3.2	Dump sites
	1.3.3	Construction sites
	1.4.1	Green urban areas
	1.4.2	Sport and leisure facilities
LGN 4	18	Continuous Urban area
	19	Built-up rural area
	20	Deciduous forest in urban area
	21	Coniferous forest in urban area
	22	Built-up area with dense forest
	23	Grass in built-up area
	24	Bare soil in built-up area

Table C.3: Land cover UMD and IGBP category types

Land cover categories		
Classification system	Land cover code	Land Cover definition
IGBP	0	Water Bodies
	1	Evergreen Needleleaf Forests
	2	Evergreen Broadleaf Forests
	3	Deciduous Needleleaf Forests
	4	Deciduous Broadleaf Forests
	5	Croplands
	6	Closed Shrublands
	7	Open Shrublands
	8	Woody Savannas
	9	Savannas
	10	Grasslands
	11	Permanent Wetlands
	12	Croplands
	13	Urban and BuiltUp
UMD	0	Water bodies
	1	Evergreen Needleleaf Forests
	2	Evergreen Broadleaf Forests
	3	Deciduous Needleleaf Forests
	4	Deciduous Broadleaf Forests
	5	Mixed forests
	6	Woodlands
	7	Wooded Grasslands/Shrublands
	8	Closed Bushlands or Shrublands
	9	Open Shrublands
	10	Grasslands
	11	Croplands
	12	Barren
	13	Urban and Built-up

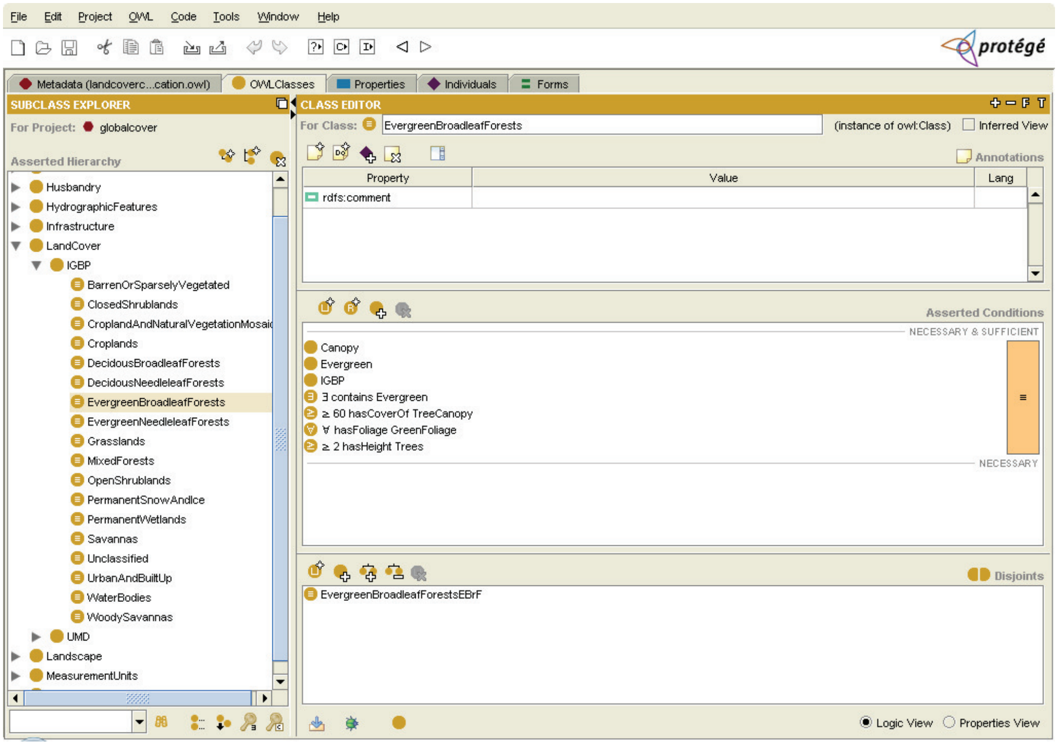

Figure C.1: Concept construction in protégé

Appendix D

The Output of Geo-semantic Similarity Measurements and Validation

D.1 Syntax and semantics of $ALCHQ$

Table D.1: Syntax and semantics of $ALCHQ$
(adapted from [JKS$^+$07])

Syntax	Semantics	Name
$\top$	Δ^I	Top
$\bot$	$\emptyset$	Bottom
A	$A^I \subseteq \Delta^I$	Atomic concept
R	$R^I \subseteq \Delta^I \times \Delta^I$	Atomic role
$\neg C$	$\Delta^I \; C^I$	(Full) negation
$C \equiv D$	$C^I = D^I$	Concept equality
$C \sqsubseteq D$	$C^I \subseteq D^I$	Concept inclusion
$R \equiv S$	$R^I = S^I$	Role equality
$R \sqsubseteq S$	$R^I \subseteq S^I$	Role inclusion
$C \sqcap D$	$C^I \cap D^I$	Concept intersection
$C \sqcup D$	$C^I \cup D^I$	Concept union
$\vee R.C$	$a \in \Delta^I \mid \forall b.(a,b) \in R^I \rightarrow y \in C^I$	Value restriction
$\exists R.C$	$a \in \Delta^I \mid \exists b.(a,b) \in R^I \wedge y \in C^I$	Existential quantification
$\leq nR.C$	$a \in \Delta^I \lvert \lvert b \in \Delta^I \rvert (a,b) \in R^I \wedge b \in C^I \rvert \leq n$	Qualified max. number restriction
$\geq nR.C$	$a \in \Delta^I \lvert \lvert b \in \Delta^I \rvert (a,b) \in R^I \wedge b \in C^I \rvert \geq n$	Qualified min. number restriction

D.2 Similarity measurement process

The following is an overview of similarity measurement process by SIM-DL as explained in section 4.1.1

- *Overall similarity*(Sim_u) The overall similarity between two concepts C_s and C_t. Similarity between disjunctions $C_{s_1} \cup .. \cup C_{s_n}$ and $C_{t_1} \cup ...C_{t_m}$ is measured according to equation D.1. $\omega_{i,j}$ is applied on the overall similarity to act as adjustable factor for relative importance. The sum of $\omega_{i,j}$ is always 1. SI is a set of tuples (C_s,C_t) chosen for comparison.

$$\mathrm{Sim}_u(C_s, C_t) = \textstyle\sum_{C_s,C_t \in SI} \omega_{i,j} \times Sim(C_{s_i}, C_{t_j}), \textit{(source [JKS}^{+}\textit{07, JRSK08])} \quad \text{(D.1)}$$

source [JKS+07, JRSK08]) Similarity is calculated for each element of the set of tuples C_{s_i}, C_{t_j}. each of this set is formed by intersection in $ALCHQ$ normal form and similarity measured by Sim_i. Sim_i is the function that determines similarity on this level. It is a normalized sum derived from similarity functions for the constructors involved [JKS+07, JRSK08].

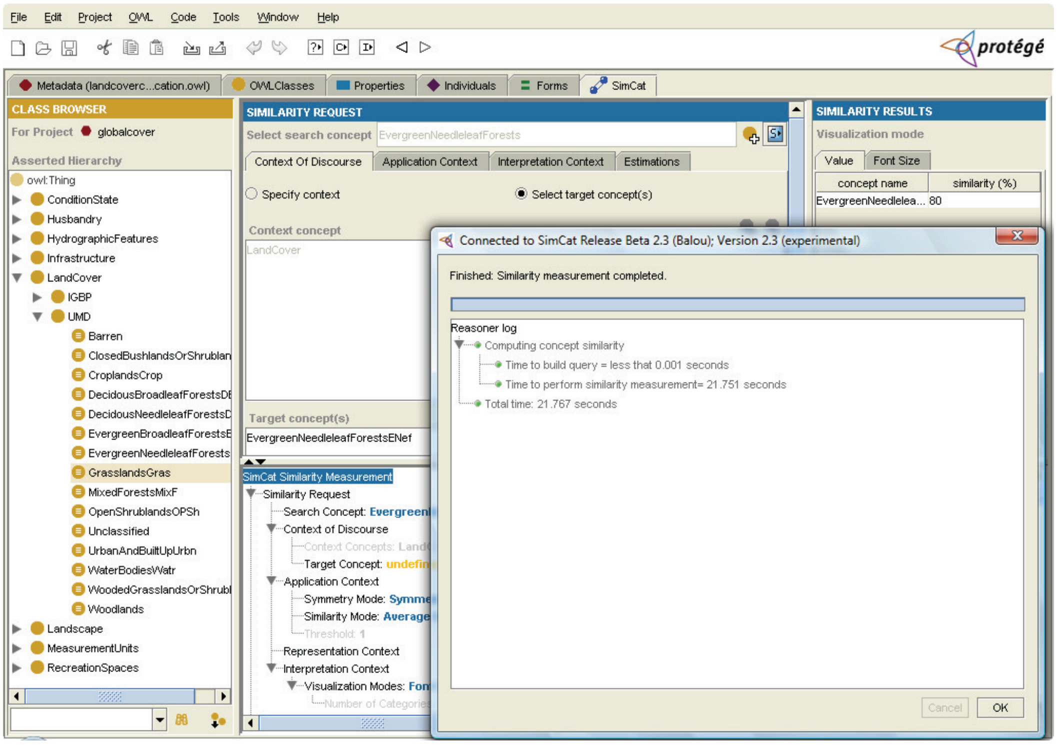

Figure D.1: SIM-DL plug-in in Protégé:Finished similarity measurement. This figure shows the reasoning of SIM-DL in similarity measurement process and the output of the measurement of the compared concepts.

D.3 SIM-DL similarity results

Table D.2: Similarity measure for ten land cover classes calculated with SIM-DL for UMD and IGBP dataset

Class code	Search concept, C_s	Target concept, C_t	Similarity, $S(\%)$
0	Water bodies	Water bodies	75
1	Evergreen Needleleaf forest	Evergreen Needleleaf forest	80
2	Evergreen Broadleaf forest	Evergreen Broadleaf forest	80
3	Decidous Needleleaf forest	Decidous Needleleaf forest	77
4	Decidous Broadleaf forest	Decidous Broadleaf forest	77
5	Mixed forests	Mixed forests	49
6	Closed shrublands	Woodlands	25
7	Open shrublands	Wooded grassland/shrublands	22
8	Woody Savannas	Closed Bushlands/shrublands	0
9	Savannas	Open Shrublands	0
10	Grasslands	Grasslands	80

D.4 Validation results

D.5 Comparison of the semantic similarity from SIM-DL and validation framework.

D.6 Graphical representations

Table D.3: Similarity measure for ten land cover classes calculated with SIM-DL for CORINE and LGN4 dataset

Class code CORINE	Class code LGN4	Search concept (C_s)	Target concept (C_t)	Similarity(%)
112	2	Discontinuous Urban Fabric	Maize	2
211	0	Non-Irrigated Arable Land	No class	-
242	5	Complex cultivation Patterns	Cereals	1
243	0	Land Principally Occupied By Agriculture	No class	-
231	8	Pastures	Greenhouses	3
121	9	Industrial/ Commercial Units	Orchards	0
142	0	Sports and Leisure Facilities Fabric	No class	-
132	0	Dumpsites	No class	-
124	21	Airports	Coniferous forest In Built-Up area	3

Table D.4: Similarity judgement by validation framework for ten land cover classes for IGBP and UMD

Class DN code(IGBP)	Class DN code (UMD)	Similarity judgement(%)
0	0	98
1	1	62
2	2	47
3	3	47
4	4	59
5	5	78
6	6	50
7	7	47
8	8	53
9	9	44
10	10	67

Table D.5: Similarity judgement by validation framework for nine land cover classes for CORINE and LGN 4

Class code CORINE	Class code LGN4	Search concept (C_s)	Target concept (C_t)	Similarity judgment (%)
112	2	Discontinuous Urban Fabric	Maize	2
211	0	Non-Irrigated Arable Land	No class	-
242	5	Complex cultivation Patterns	Cereals	40
243	0	Land Principally Occupied By Agriculture	No class	-
231	8	Pastures	Greenhouses	36
121	9	Industrial/ Commercial Units	Orchards	25
142	0	Sports and Leisure Facilities Fabric	No class	-
132	0	Dumpsites	No class	-
124	21	Airports	Coniferous forest In Built-Up area	6

Table D.6: Similarity judgement by validation framework compared to SIM-DL for UMD and IGBP

Class DN code	SIM-DL(%)	Validation framework(%)
0	75	98
1	80	62
2	80	47
3	77	47
4	77	59
5	49	78
6	25	50
7	22	47
8	0	53
9	0	44
10	80	67

Table D.7: Similarity judgement by validation framework compared to SIM-DL for CORINE and LGN 4. The numbers in the X-axis are the DN codes for the land cover concepts that were compared. In the Y-axis are the similarity values.

Class code CORINE	Class code LGN4	Search concept (C_s)	Target concept (C_t)	SIM-DL(%)	Validation (%)
112	2	Discontinuous Urban Fabric	Maize	2	2
211	0	Non-Irrigated Arable Land	No class	-	-
242	5	Complex cultivation Patterns	Cereals	1	40
243	0	Land Principally Occupied By Agriculture	No class	-	-
231	8	Pastures	Greenhouses	3	36
121	9	Industrial/ Commercial Units	Orchards	0	25
142	0	Sports and Leisure Facilities Fabric	No class	-	-
132	0	Dumpsites	No class	-	-
124	21	Airports	Coniferous forest In Built-Up area	3	6

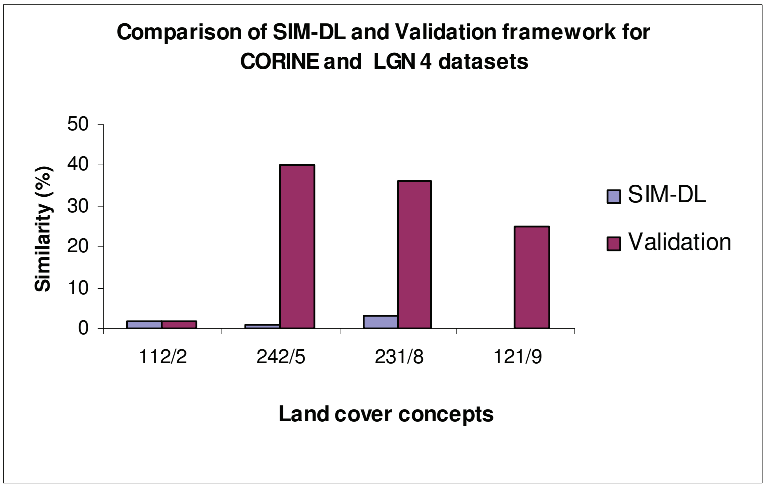

Figure D.2: Representation of the comparison of the similarity measures from SIM-DL and the validation framework. This is for the CORINE and LGN 4 dataset. The numbers in the X-axis are the DN codes for the land cover concepts that were compared. In the Y-axis are the similarity values.

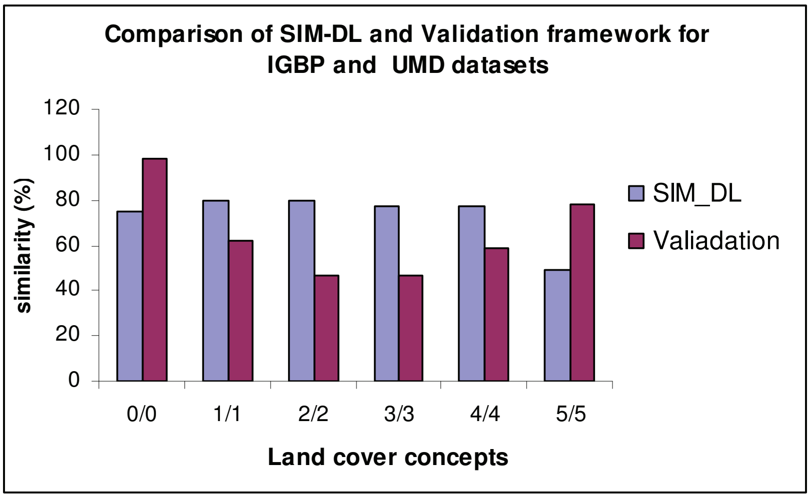

Figure D.3: Representation of the comparison of the similarity measures from SIM-DL and the validation framework. This is for the IGBP and UMD datasets. The numbers in the X-axis are the DN codes for the land cover concepts that were compared. In the Y-axis are the similarity values.

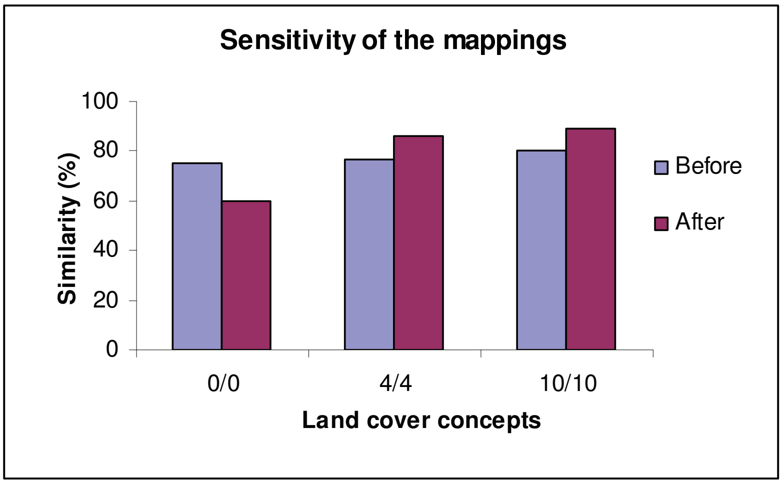

Figure D.4: Comparing the mappings before and after modification of the concept meaning. The numbers in the X-axis are the DN codes for the land cover concepts that were compared. In the Y-axis are the similarity values.

Bibliography

[AFCLL08] A. Annoni, A. Friis-Christensen, R. Lucchi, and M. Lutz. Requirements and Challenges for Building a European Spatial Information Infrastructure: INSPIRE. *Creating Spatial Information Infrastructures*, 2008.

[Amo77] T. Amos. Features of similarity. *Psychological Review*, 84,(4):327–352, 1977.

[BF96] P. Burrough and A. Frank. *Geographic Objects with Indeterminate Boundaries*. Gisdata. Taylor & Francis Ltd, pages 374, 1996.

[Cen09] Land Processes Distributed Active Archive Center. Land Cover Type Yearly L3 Global 500 m SIN Grid: https://lpdaac.usgs.gov/lpdaac/products/modis_products_table/land_cover/yearly_l3_global_500_m/mcd12q1, Access date 10th November, 2009.

[CFLGP03] O. Corcho, M. Fernández-López, and A. Gómez-Pérez. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering.*, 46(1):41–64, 2003.

[CSH06] N. Choi, I. Y. Song, and H. Han. A survey on ontology mapping. *Sigmod Rec.*, 35(3):34–41, 2006.

[DMQ05] D. Dou, D. McDermott, and P. Qi. Ontology Translation on the Semantic Web. *Journal on Data Semantics (JoDS)*, pages 35–57, 2005.

[Ehr07] M. Ehrig. *Ontology Alignment: Bridging Semantic Gap*, volume 4 of *Semantic Web and Beyond*. Springer, pp. 254, 2007.

[EM95] M. Egenhofer and D. Mark. Naive Geography. *Conference on Spatial Information Theory,(COSIT'95),Lecture Notes in Computer Sciences*, pages 1–15.Semmering, Austria, (988)1995.

[ES07] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag, Heidelberg (DE), pp. 333, 2007.

[fBO00] The National Center for Biomedical Ontology. Protégé.The Protégé project. http://protégé.stanford.edu/., Access date 10th July, 2009.2000.

[FCM06] F. Fonseca, G. Câmara, and A. Monteiro. A Framework for Measuring the Interoperability of Geo-Ontologies. *Spatial Cognition & Computation*, 6, (4):307–329, 2006.

[Gru93] T. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5,(2):199–220, 1993.

[GS04] R. Goldstone and J. Son. Similarity. *On thinking: Towards a theory of thinking. The Parmenides Foundation.*, 2: pp.42, 2004.

[GvH04] A. Grigoris and F. van Harmelen. *Web Ontology Language: OWL.Handbook on Ontologies in Information Systems*. Springer-Verlag, 2004.

[Hal06] M. Hall. A Semantic Similarity Measure for Formal Ontologies (With an application to ontologies of geographic kind). Master's thesis, Alpen–Adria Universitat Klagenfurt, 2006.

[HG01] F. Hakimpour and A. Geppert. Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach. *Proceedings of International conference on Formal Ontologies in Information Systems FOIS'01. ACM Press*, page pp.12, October 2001.

[Jan08] K. Janowicz. Kinds of Contexts and their Impact on Semantic Similarity Measurement. 2008.

[JKS+07] K. Janowicz, C. Keßler, M. Schwarz, M. Wilkes, I. Panov, E. Martin, and B. Baumer. Algorithm, Implementation and Application of the SIM-DL Similarity Server. *In proceedings of the Second International Conference on GeoSpatial Semantics (GeoS 2007)*, pages 18 pp.,Mexico City, Mexico, 2007.

[JRSK08] K. Janowicz, M. Raubal, A. Schwering, and W. Kuhn. Semantic Similarity Measurement and Geospatial Applications. *Transactions in GIS*, 12,(6):651–659, 2008.

[JW09] K. Janowicz and M. Wilkes. SIM-DLa: A Novel Semantic Similarity Measure for Description Logics Reducing Inter-concept to Inter-instance Similarity. In *ESWC 2009 Heraklion: Proceedings of the 6th European Semantic Web Conference on The Semantic Web*, pages 353–367, Berlin, Heidelberg, 2009. Springer-Verlag.

[KC07] S. Kaza and H. Chen. Evaluating ontology mapping techniques: An experiment in public safety information sharing. *Decision Support Systems , Information Technology and Systems in the Internet-Era.*, 45(4):714–728, 2007.

[KD07] W. Kammersell and M. Dean. Conceptual Search: Incorporating Geospatial Data into Semantic Queries. *Advanced Information and Knowledge Processing. In: Terra Cognita - Directions to the Geospatial Semantic Web, Athens, GA*, pages 47–54, 2007.

[Kie08] B. Kieler. Semantic Data Integration Across Differentt Scales: Automatic Learning of Generalization Rules. *In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Beijing,vol. XXXVII,(2):685–690, 2008.

[KKT05] M. Kavouras, M. Kokla, and E. Tomai. Comparing categories among geographic ontologies. *Computers & Geosciences,*, 31(2):145–154, 2005.

[Kok06] M. Kokla. Guidelines on Geographic Ontology Integration. *In: International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences (ISPRS Technical Commission II Symposium)*, Vienna,vol.XXXVI (II):67–72., 2006.

[KRJ+08] C. Keßler, M. Raubal, K. Janowicz, Meersman R., Tari Z., and Herrero P. et al. (Eds.). The Effect of Context on Semantic Similarity Measurement. *In Proceedings of the 3rd International IFIP Workshop On Semantic Web & Web Semantics (SWWS 07).November 29-30 2007; forthcoming,LNCS 4806*, Vilamoura, Portuga,vol.II:1274–1284, 2008.

[LSK+07] M. Lutz, J. Sprado, E. Klien, C. Schubert, and I. Christ. Overcoming semantic heterogeneity in spatial data infrastructures. *Computers & Geosciences*, 35(4):739–752, 2007.

[MDH05] H. Mizen, C. Dolbear, and G. Hart. Ontology Ontogeny: Understanding how an ontology is created and developed. *Proc. First International Conference on Geospatial Semantics (GeoS 2005),LNCS 3799*, pages 15–29, 2005.

[MEH04] D. Mark, M. Egenhofer, and S. Hirtle. Ontological Foundations for Geographic Information Science. *In Mc-Master, R. and Usery, E., editors, A Research Agenda for Geographic Information Science*, pages 1–8, 2004.

[NM01] N. Noy and D. McGuinness. Ontology development 101: A guide to creating your first ontology. Technical report, SMI-2001-0880, Stanford Medical Informatics, 2001.

[NM02] N. Noy and M. Musen. Evaluating Ontology-Mapping Tools: Requirements and Experience. *In Proc. 1st workshop on Evaluation of Ontology Tools (EON2002), EKAW02,*, 2002.

[Noy04] N. Noy. Semantic Integration: A Survey Of Ontology-Based Approaches. *SIGMOD Record*, 33(4):65–70, 2004.

[NSD+01] N. Noy, M. Sintek, S. Decker, M. Crubézy, R. Fergerson, and M. Musen. Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, 16,(2):60–71, 2001.

[OCKT+05] M. O' Connor, H. Knublauch, S. Tu, B. Grosof, Mike Dean, William Grosso, and Mark Musen. Supporting Rule System Interoperability on the Semantic Web with SWRL. *ISWC 2005.*, LNCS 3729:974–986, 2005.

[OCSPA09] M. O' Connor, R. Shankar, D. Parrish, and Das A. Knowledge-data integration for temporal reasoning in a clinical trial system. *International Journal of Medical Informatics*, vol. 78:S77–S85, 2009.

[OZ08] P. van Oosterom and S. Zlatanova. *Creating spatial information infrastructures:Towards the Spatial Semantic Web.* CRC Press: Boca Raton, Florida, 2008.

[P9́4] A. Pérez. From Knowledge Based Systems to Knowledge Sharing Technology: Evaluation and Assessment. Technical report, Stanford University, 1994.

[PSB99] D. Peuquet, B. Smith, and B. Brogaard. The Ontology of Fields. Technical report, of a specialist meeting held under the auspices of the Varenius Project .Santa Barbara: National Center for Geographic Information and Analysis, 1999.

[Sch08] A. Schwering. Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS*, vol:12,(1):5–29, 2008.

[SHD+03] S. Staab, I. Horrocks, S. Decker, M. Kifer, B. Grosof, J. Angele, and G. Wagner. Trends and Controversies:Where Are the Rules? *IEEE Intelligent Systems*, 18:76–83, 2003.

[SLKB06] J. Stoter, R. Lemmens, B. Köbben, and N. Bakker. Semantic Data Integration in a Multiple Representation Environment. *In: Proceedings of Joint ISPRS working groups II/3 and II/6 workshop on multiple representation and interoperability of spatial data : 22-24 February, 2006*, Hannover:8 pp., 2006.

[SM98] B. Smith and D. Mark. Ontology and Geographic Kinds. *Paper presented at the International Symposium on Spatial Data Handling*, Vancouver, BC, Canada:308–320, 1998.

[SM01] B. Smith and D. Mark. Geographic categories:an ontological investigation. 15(7):591–612, 2001.

[SP07] E. Sirin and B. Parsia. SPARQL-DL:SPARQL Query for OWL-DL ,3rd OWL Experiences and Directions Workshop (OWLED-2007). *In Proc. of 3rd OWL Experiences and Directions Workshop (OWLED-2007)*, 2007.

[TW00] H. Thunnissen and A. Wit. The National Land Cover Database of the Netherlands. *In: Beek, K.J., Molenaar, M. (Eds.), Proceedings* *of the XIX Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS) Geoinformation For All. Lemmer, GITC, 2000. Int. Arch. Photogramm. Remote Sensing. (Part B7/3).*, vol.33:223–230, 2000.

[UG96] M. Uschold and M. Gruninger. Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, 11(2):93–155, 1996.

[UG04] M. Uschold and M. Gruninger. Ontologies and Semantics for Seamless Connectivity. *SIGMOD Record*, vol.33,(4):58–64, 2004.

[ZS06] A. Zhdanova and P. Shvaiko. Community-driven Ontology Matching. *In Proc. 3rd European Semantic Web Conference (ESWC), Lecture notes in computer science*, 4011:34–49, 2006.