Mathematisch-Naturwissenschaftliche Fakultät
Wegelerstr. 10
53115 Bonn

Rheinische Friedrich-Wilhelms-Universität Bonn

THE KNOWLEDGE-BASED SEARCH FOR WATER-RELATED INFORMATION SYSTEM FOR THE MEKONG DELTA, VIETNAM

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinische Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Tran Thai Binh
aus
Hochiminh City, Vietnam
Bonn December 2013

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms Universität Bonn

1. Referent: Prof. Dr. Klaus Greve
2. Referent: Prof. Dr. Gunter Menz
Tag der mündlichen Prüfung: 04.12.2013
Diese Dissertation ist auf dem Hochschulserver der ULB Bonn
http://hss.ulb.uni-bonn.de/diss online elektronisch publiziert.

Erscheinungsjahr: 2014

For my parents

my wife

and my sons

Acknowledgements

There are countless people who have supported and encouraged me in completing this study. I would like to express my deep gratitude to all of the people who have supported me during my research.

First of all, I would like to thank the DAAD (Deutscher Akademischer Austausch Dienst – German Academic Exchange Service), the DLR (Deutsches Zentrum für Luft- und Raumfahrt – German Aerospace Center), and the HCMIRG (Hochiminh City of Institute of Resources Geography) for giving me the opportunity to participate in the doctoral research program.

I would like to express sincere appreciation to the DLR – DFD – LA for their extended long-term support and especially to Dr. Claudia Künzer, the leader of WISDOM team for her support and continuous encouragement.

I would like to express my gratitude to my Principal Supervisor, Professor Dr. Klaus Greve, Bonn University, for his valuable advice and guidance during the three and a half years of my study. I also appreciate the support of my Associate Supervisors: Professor Dr. Gunter Menz.

A special thanks to Verena Jaspersen who always came up with very good ideas and suggestions and who was very patient with me (particularly with my poor written English). I also thank to Malte Ahren who help me to print out this thesis.

I would like to thank Dr. Thilo Wehmann and Florian Moder, German colleagues and all colleagues in the WISDOM team at DLR. I would to thank Dr. Lam Dao Nguyen, Pham Bach Viet and all colleagues in GIRS (Geographic Information system and Remote sensing Research Center).

This thesis would never have been completed without the encouragement and devotion of my family – my wife, Nguyen Thi Phuong Chi, and my sons Tran Huu Duc and Tran Huu Phuc. Thank for their continuing support and patience during this period. Thank for their encouragement and believing in me, which helped me to pursue my research study towards the end.

Last but not least, I would like to thank my parents, Tran Van Hoat and Vo Thi Xuan, who support me spiritually throughout my life.

Tran Thai Binh

February 2013

Summary

In recent years, the World Wide Web has strongly changed way of sharing and accessing data. Moreover, with new methods of data collection are developed we have much more data today. However, it is not straightforward to integrate and to discover data or information from different systems, different fields of research as well, especially when users need to find and retrieve the relevant data for their demands. Normally, users get lost in a huge amount of irrelevant search results or may miss relevant data or information. The issue happens because the data are heterogeneity, which are various in formats and organized under different schemas and likely named in different terms to describe the meaning. Thus, it is necessary to have a proper solution to ensure interoperability between different systems. This study proposes an innovation way to describe the meaning of data on how they relate to each other based on the expert knowledge and common dictionaries in order to provide a search result more precise and sufficient for user queries.

The thesis focuses on applying the ontology to discovering and retrieving data for the WISDOM Information System (IS), a Web-based information system for water related information system in Mekong Delta, Vietnam. The proposed approach applies the hybrid ontology and the WISDOM IS is devided into three main domains: i) Data domain, ii) Observed Object domain and iii) Application domain.

Data Domain contains classes that present the properties of datasets, e.g. format type; geometric resolution – pixel size; spatial representation – line, point, polygon or pixel; and spatial relation - which area the datasets relate to; and thematic reference classes of datasets.

Observed Object Domain consists of classes that describe physical and non-physical objects related to the water subject, i.e. “man-made feature”, “natural” and “social”, called observed objects. Phenomena are also presented concerning observed objects. The relationships in this domain are described independently from user’s tasks.

Application Domain describes the user’s tasks, divided into types, e.g. response task, monitoring task, etc. The user tasks are described in relation to observed objects, which are the main concerns of these tasks.

The relations between domains are based on the expert knowledge and common dictionaries. These relations describe how the data concern to each other, to phenomena or to observed

objects. The real world object observing by users task are describe in relating with the phenomenon in order to provide all relevant data set just for one search.

This study also builds a prototype. The result returns from the prototype are evaluated to prove the sufficiency of the proposed approach. The evaluation uses the common criteria, i.e. precision, recall and average precision. The evaluation proves that the proposed approach is good and has high ability to apply in practice.

This study concluded that ontology can resolve the semantic heterogeneity of data. It can describe the properties of dataset and the relations of dataset's topic on the real world object, phenomena and users' tasks as well. The proposed approach can be applied not only for water related domain, but also for another domain.

Curriculum Vitae

Name: TRAN THAI BINH

Scientific Publications

1	***Ontology based approach for water related information system for Mekong delta, Vietnam***	GISIDEAS 2012, H CMC, Vietnam	10/2012
2	***Ontology based description of satellite imageries for application based data query***	EnviroInfo 2011, Milan, Italy	10/2011
3	***Ontology based approach for Geospatial Semantic Web***	ACRS 2010, Hanoi, Vietnam	11/2010
4	***Use of remotely sensed data and GIS to detect changes of riverbank in Mekong River***	Seminar on "Remote sensing applications in riverine and coastal engineering", HCMC, Vietnam	12/2001
5	***Using GIS to management Transportation Infrastructure of HCMC***	The 8th Conference on Science and Technology - HCMC University of Technology	04/2002
6	***Using GIS for natural resources management***	Seminar at HCMC University of Social Science and Humanities	12/2007

CONTENT

Acknowledgements iii
Summary iv
Curriculum Vitae vi
CONTENT ix
Figures xii
Tables xv
Glossary xvi
1. INTRODUCTION AND OBJECTIVES 1
1.1. Introduction 1
1.1.1. Motivation of this study 1
1.1.2. Definitions of fundamental scientific and technical terms 5
1.1.2.1. Observed objects 5
1.1.2.2. Phenomena 5
1.1.2.3. Tasks 5
1.2. Objectives of the thesis 6
1.3. Structure of the thesis 8
2. LITERATURE REVIEW 9
2.1. Introduction 9
2.2. State of technology 12
2.2.1. Existing standards 12
2.2.1.1. The Open Geospatial Consortium (OGC) Standards 12
2.2.1.2. The International Standardization Organization (ISO) standards 14
2.2.1.3. Summary 18
2.2.2. Ontology 19
2.2.3. Database Connection 23
2.3. Research Review 25
2.3.1. Data Integration 25
2.3.2. Task Ontology 30
2.3.3. Existing Ontologies 31
2.3.4. Ontology Mapping 34
2.4. Conclusion 38

3. METHOD 39
3.1. Overview of approach 39
3.2. Data Domain 45
3.3. Observed Object Domain 48
3.4. Application Domain 50
3.5. Spatial and Temporal Domain 51
3.6. Relational database (RDB) to Resource description framework (RDF) 53
3.7. Ranking 54
3.8. Conclusion 58
4. WISDOM INFORMATION SYSTEM CONTEXT 60
4.1. Introduction 60
4.2. Collected Data in WISDOM / Data model in WISDOM 62
4.2.1. Fields of research 62
4.2.2. Data management model 64
4.3. Conclusion 71
5. IMPLEMENTATION OF PROTOTYPE 75
5.1. Proposed approach applied in the WISDOM Information System 75
5.2. Data domain 78
5.3. Observed object domain 83
5.4. Application domain 86
5.5. Spatial and Temporal domain 88
5.6. Implementation of a prototypical Graphical User Interface 88
5.6.1. The used tools and software. 89
5.6.2. The Graphical User Interface 89
5.7. Conclusion 91
6. EVALUATION 93
6.1. Precision and recall 93
6.2. Average precision 100
6.2.1. Average precision at seen relevant documents 100
6.2.2. Average precision in combination with recall 105
6.3. Conclusion 106
7. CONCLUSION 108
7.1. Summary of findings 108
7.2. Conclusion 110

7.3. Recommendation 111
Appendices 113
A. List of ISO/TC 211 Standards 113
B. ISO 19115:2003 114
C. List of ProductGroup 115
D. The Relationships and Properties in Data Domain 116
E. The Relationships and Properties in Observed Object Domain 118
F. The Relationships and Properties in Application Domain 120
G. JAVA Code 122
H. SPARQL 125
References 127

Figures

Figure 1.1: The three dimensions of heterogeneity at the conceptual aspect 4

Figure 2.1: Ontology approaches 21

Figure 2.2: Example RDF 22

Figure 2.3: Example RDF and RDFs 22

Figure 2.4: Marius Podwyszynski's approach 27

Figure 2.5: Semantic Translation Specification Service in M. Lutz proposed approach 28

Figure 2.6: Abstract of Athanasis Nikolaos approach 29

Figure 2.7: SWEET ontologies and their interrelationships 32

Figure 2.8: AGROVOC web page 34

Figure 2.9: An example of ontology mapping 36

Figure 3.1: The Overview of thesis's approach 41

Figure 3.2: Main classes and main relationships of domains 42

Figure 3.3: User search for observed object 44

Figure 3.4: User search for phenomenon 44

Figure 3.5: User search for phenomenon with a particular task 45

Figure.3.6: Outline of Data domain 47

Figure 3.7: Example for relationships of classes and individuals 48

Figure.3.8: Outline of Observed object domain 48

Figure 3.9: Abstract model for inference of datasets 49

Figure.3.10: Abstract of Task Domain 50

Figure.3.11: Outline of Temporal Domain 51

Figure.3.12: Abstract of Spatial Domain 52

Figure 3.13: An example for the ranking of cover area property 55

Figure.3.14: An example for the "bestFit" property 55

Figure.3.15: User query related to temporal property of dataset 56

Figure 4.1: Location of Mekong Delta, Vietnam 61

Figure 4.2: WISDOM research fields 63

Figure 4.3: Aspects of data within WISDOM IS 65

Figure 4.4: An example for spatial reference schema 67

Figure 4.5: Thematic reference used to search data sets in WISDOM 70

Figure 4.6: The relationship between dataset and thematic reference via product group 71

Figure 4.7: Thematic reference variable 72

Figure 4.8: Spatial reference variable 72

Figure 4.9: Temporal reference variable 72

Figure 4.10: The GUI of WISDOM Information System 73

Figure 5.1: Integration of approach into existing system. 76

Figure 5.2: Flowchart of approach. 77

Figure 5.3: The abstract hierarchy of Data domain 79

Figure 5.4: Properties are used as the definition for HighResolution class 82

Figure 5.5: Outline of the classes hierarchy of the observed object domain 84

Figure 5.6: Classes hierarchy of application domain 87

Figure 5.7: An example of a direct relationship of FirstAid task 87

Figure 5.8: An example of indirect relationship of Monitoring task 88

Figure 5.9: The GUI of the prototype for Observed Object search 90

Figure 5.10: The GUI of the prototype for phenomena search. 91

Figure 6.1: An example of precision and recall 94

Figure 6.2: The precision (1) 97

Figure 6.3: The precision (2) 98

Figure 6.4: The recall values 99

Figure 6.5: The AP at seen relevant documents compares with the value of 0.50, 0.80 and 1.00 for all test cases 104

Tables

Table 2.1: Core metadata for geographic datasets (ISO 2003) 18

Table 4.1: Administrative units in Vietnam are stored in the spatial reference table 68

Table 4.2: Example of the spatial reference model in the WISDOM IS 68

Table 4.3: Examples of “product-theme” entity relation model in WISDOM IS 71

Table 5.1: Object properties and data properties in data domain 81

Table 5.2: An example for flood’s effect from Schramm (Schramm et al. 1986) 85

Table 6.1: The list of test cases 96

Table 6.2: The precision of the test cases have been done by testers 97

Table 6.3: The recall of the four test cases 99

Table 6.4: Example of average precision from two different systems 101

Table 6.5: The average precision of the test cases 103

Table 6.6: The AP at seen relevant documents compare with 1.00, 0.80 and 0.50 104

Table 6.7: The AP in combination with recall for test cases 106

Glossary

Term	Description
AGROVOC	The thesaurus are created by Food and Agriculture Organization of the United Nations, contains more than 40000 concepts in up to 22 languages covering topics related to food, nutrition, agriculture, fisheries, forestry, environment and other related domains. (http://aims.fao.org/website/AGROVOC-Thesaurus/sub)
D2RQ	The D2RQ Platform is a system for accessing relational databases as virtual, read-only RDF graphs. (http://www.w3.org/2001/sw/wiki/D2RQ)
Eclipse	An open development platform comprises of extensible frameworks, tools and runtimes for building, deploying and managing software across the lifecycle. (www.eclipse.org/)
FAO	Food and Agriculture Organization of the United Nations. The main effort of FAO is achieving food security for all - to make sure people have regular access to enough high-quality food to lead active, healthy lives. FAO's mandate is to raise levels of nutrition, improve agricultural productivity, better the lives of rural populations and contribute to the growth of the world economy. (http://www.fao.org/)
Jena	Jena (Apache Jena™) is a Java framework for building Semantic Web applications. Jena provides a collection of tools and Java libraries to help you to develop semantic web and linked-data apps, tools and servers. (http://jena.apache.org/)
Metadata	Metadata is structured information that describes, explains, locates, and otherwise makes it easier to retrieve and use an information resource.

	(Section 2.2.1.2)
Observed object	The observed object is the object in the real world, which can be described by datasets. (Section 1.1.2.1)
OGC	The Open Geospatial Consortium (OGC) is an international industry consortium of 477 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards. (http://www.opengeospatial.org/ogc)
OWL	The Web Ontology Language (OWL) is a knowledge representation language for authoring ontologies. It is designed for use by applications that need to process the content of information instead of just presenting information to humans. (Section 2.2.2)
Pellet reasoner	The Pellet reasoner is a program able to infer logical consequences from a set of asserted facts or axioms. (http://clarkparsia.com/pellet/)
Phenomenon	Phenomenon is any observable occurrence, normally, it is refers to an extraordinary event. (Section 1.1.2.2)
RDB	A relational database is a collection of data items organized as a set of formally described tables from which data can be accessed easily. A relational database is created using the relational model. (http://www.linkedin.com/skills/skill/Relational_Databases)
RDF	Resource Description Framework (RDF) uses a simple structure statement “Subject – predicate – object” to describe resources or to present the relation between resources in structure resource – property – resource/literal. (Section 2.2.2)

RDFs	Resource Description Framework Schema (RDFs) was developed based on RDF characteristics but it is extended to describing about classes of resource and their properties. (Section 2.2.2)
SPARQL	SPARQL defines a standard query language and data access protocol to be used with RDF data model. (Section 2.2.3)
User's task	Task is defined as an action to a response to a phenomenon.\ (Section 1.1.2.3)
Water mask	Using satellite imagery, all the water bodies are mask into a single layer for the normal water level. In addition, a wet seasonal or a flooded seasonal satellite image is used to extract the flood boundary. This process will delineate the normal flood levels. (http://www.systemecology.com/services4.html)
WISDOM	**W**ater-related **I**nformation **S**ystem for the sustainable **D**evelopment **o**f the **M**ekong Delta in Vietnam (WISDOM) is a bilateral research project between Germany and Vietnam. (http://www.wisdom.caf.dlr.de/)

1. INTRODUCTION AND OBJECTIVES

1.1. Introduction

The first chapter is organized as follows. After an introduction to the specific problem of interest motivating this study, definitions of fundamental scientific and technical terms are given. Objectives are outlined in the next section, followed by the section for the structure of the thesis.

1.1.1. Motivation of this study

Finding and accessing sufficient data or information to answer scientific questions are a crucial task in many different fields of research (Klien et al. 2004; Zhan et al. 2008). Nowadays, the amount of available data and information are increasing dramatically with the development of the World Wide Web (WWW). Invented since 1990 by Tim Berners-Lee, the WWW has significantly grown (W3C 2000) and nowadays it can be considered to be the most effective tool to share information and data. At the present, the WWW consists of more than eight billion pages (WorldWideWebSize 2012). Data providers normally generate data for their personal use or applications, thus the published data are based on the own perspective of providers (Navarrete 2006). As a consequence, despite containing a huge amount of information, the way of information provided in the WWW is very heterogeneous. This makes searching for particular information difficult as common users might experience, that a search result is unsuitable or irrelevant to the given keywords in the query given by the user in a searching machine (Lutz et al. 2009).

Hence, one of the most current challenges in this context is to design and improve the way on how to extract information which is valuable and tailored to certain user group interest, out of a large amount of obtainable resources (Han et al. 2006).

Not only in the field of the WWW, the amount of available data is increasing day by day because of the development of collecting data methods (Mena et al. 1998; Han et al. 2006). This counts also for other information technology (IT) related disciplines such as geographic information systems (GIS) for example. GIS captures, manages, analyses and displays all forms of geographically referenced information (GIS.com 2012). The situation in this particular IT domain is very similar. Gaining the right information out of a GIS becomes increasingly challenging. Since the late 1970s, most of the geographic information systems were based on proprietary commercial products running mostly on desktop computers (Coppock et al. 1991; Navarrete 2006). Those systems were built for different thematic purposes and aspects. That was very difficult to exchange and share data between organizations, because they might use different data standards, developed by software providers for various thematic and commercial purposes (Navarrete 2006). There was a significant move from isolated desktop programs to programs which can run as an internet service and interact with heterogeneous systems and platforms (Sriphaisal et al. 2006). WWW enables data providers to share information and to avoid the inefficient and redundant data handling by centralizing information through applying state of the art internet based technologies (Athanasis et al. 2009). That leads to the need of the still on-going research trend, the so-called *interoperable GIS* (Goodchild et al. 1999), in order to integrate and share information between different systems (Yuan 1997). This term can be compare with Spatial Data Infrastructure (SDI) which was devised by US National Research Council in 1993. SDI indicates a framework which facilitates the creation, exchange, and use of geospatial data and related information resources across an information-sharing community (ESRI 2010).

According to Thorsten Reitz, interoperability is defined as *"The ability of systems to exchange information automatically"* (Reitz 2008). From a software engineering perspective, interoperability implies an open system that can integrate software components (Navarrete 2006). In this aspect, Open Geospatial Consortium (OGC) develops and promotes standards for open interfaces, protocols, schemas etc. to exchange geospatial data and instructions between different systems, by defining voluntary specifications to enable syntactic interoperability (OGC 2012d). OGC plays an important role in solving the heterogeneity of geospatial software by developing specifications at

multi-levels, which enables developers to build software by integrating different modules in accordance with OGC specifications. Examples here are web based interfaces namely a few as interfaces of visualization of the Web Mapping Service (WMS), download services realization of Web Feature Service (WFS) or Web Coverage Service (WCS), searching data of the Catalogue Service for the Web (CSW), and processing services of Web Processing Service (WPS) (OGC 2012d).

From an information perspective, the term interoperability indicates the need to share information. Information was created independently dealing with different aspects about facts in the real world with minimum or no communication between systems. Different requirements and technics for generating geodata brings various data models attempting to describe the world, and consequently, generating heterogeneous information (Bishr 1998).

Friis-Christensen (Friis-Christensen et al. 2005) identified three different types of heterogeneity as the follows:

- Syntactic heterogeneity: geodata from resources can have different data formats. Spatial data may be represented through various models (vector or raster) or they may refer to different spatial coordinate systems.
- Structural heterogeneity: geographical features can be represented by using several geometrical and data schemas. A geographic feature can be represented by distinct geometric features. For instance, roads can be represented by either polygons or lines or multi-temporal techniques (Bishr 1998).
- Semantic heterogeneity: the real world may be categorized in many ways by agents (persons or organizations) use various mental models. These categories correspond to thematic concepts, therefore, we can observe that semantic is mainly related to the thematic component of the geographic information (Navarrete 2006).

OGC provides specifications for standardizing service interfaces and exchange formats. These specifications define a common format for representing geographic information avoiding syntactic heterogeneity (Lutz et al. 2006). Specifications of ISO 19115 (Ostensen et al. 2002), which is a standard for metadata, defines how geographical information and associated services should be described, including the identification, the extent, the quality, the spatial and temporal schema, spatial reference and the distribution of digital geographic data. (ISO 19115:2003). The metadata describes the structure of the representation schema

in dataset, and it is an important tool to deal with structural heterogeneity. However, there is no standard that deals with semantic heterogeneity (Vaccari et al. 2009; Yan et al. 2011).

In terms of semantic heterogeneity, there are several concepts of how to model the real world. Three types to describe the world can be identified basing on the classification provided by KnowledgeWeb (KnowledgeWeb 2005) (Figure 1.1).

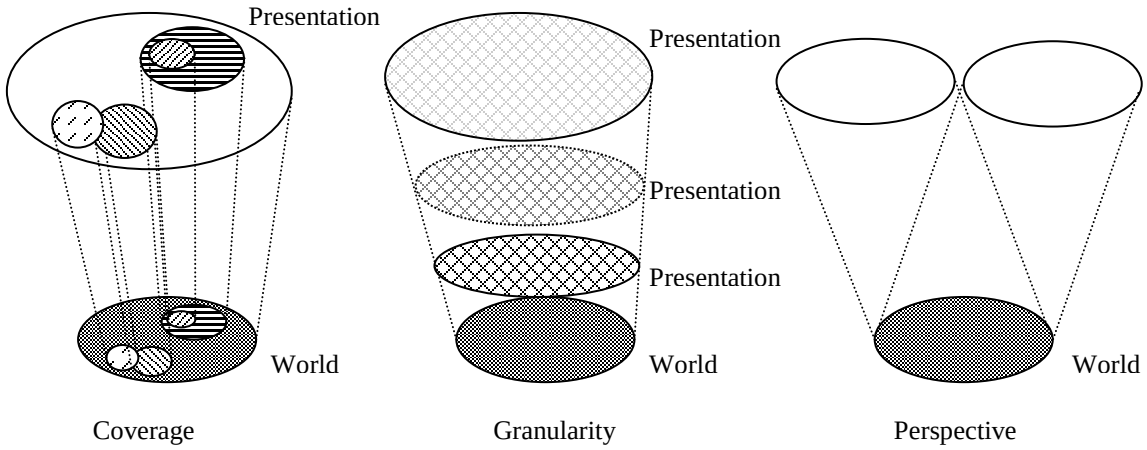


Figure 1.1: The three dimensions of heterogeneity at the conceptual aspect (source: (KnowledgeWeb 2005))

- Coverage: several models cover different parts of the world. The models may overlap in some parts.
- Granularity: One model provides a more detailed description than the others.
- Perspective: two models are the results of observing the real world from different points of view. This is the typical case of different disciplines.

The semantic heterogeneity is obviously the most complex one (KnowledgeWeb 2005). This study addresses the semantic heterogeneity in a geo-database consisting of geodata collected from multidisciplinary scientific, and proposes an approach providing a sufficient way to discover and retrieve all relevant data for user requests.

1.1.2. Definitions of fundamental scientific and technical terms

1.1.2.1. Observed objects

Data are created or collected in order to describe the status of objects in the real world. When searching for data, users actually want to get the information of an object at a certain time for a particular location. In this study, the object(s) in the real world, which can be described by datasets, are hereafter defined "observed object" divided into three groups.

- Infrastructure Feature: define classes which describe features made by human being such as road network, industrial area etc.
- Natural Feature: define classes which describe natural features such as soil, water resource etc.
- Social Feature: define classes which describe features related to human activities such as economy, education etc.

1.1.2.2. Phenomena

In point of fact, when users search for data of an observed object, they are interested not only in the object itself, but also in the related phenomenon. For instance, water level describes the status of a river, and it is useful to monitor flood – flood is a phenomenon. Phenomena are described in this study in relation to observed objects (the influence between observed objects and phenomena). This does not include the cause of the phenomena.

1.1.2.3. Tasks

In this thesis, task is defined as an action to a response to a phenomenon. For example, flood rescue and flood monitoring. For a certain phenomenon with different tasks, users

need disparate data or information (for example, flood with task of rescue needs information about transportation, health care system and population, while task of monitoring task needs information on water level). By this approach, for a certain task, users can retrieve all data sufficient for their planned actions.

1.2. Objectives of the thesis

Water-related Information System for Sustainable Development of the Mekong Delta project (WISDOM) is a bi-lateral project between Vietnam and German government has been established focusing on development and implementation of an innovative water-relate information system containing all the outcomes and results of the different research disciplines involved in the project (WISDOM 2011). The main objective of this study is to define a method to design and implement ontology into the WISDOM Information System – the web based information system for WISDOM - in order to gain more precise querying results. This includes the intention to provide users an efficient tool to discover and retrieve relevant data for specific tasks in the field of water-related information. The proposed approach is also evaluated by widely agreed criteria.

Since ontologies can describe data in a meaningful, machine readable way, based on the defined ontologies in this thesis, the end-user can discover and retrieve data via Web services more accurately and efficiently at three levels.

(1) To provide more accurate and more reasonable results for users, especially for non-GIS users who have less experience on how to search for geodata,

(2) To provide an innovative way to retrieve all the relevant data of a phenomenon for users, and

(3) To provide an innovative way to retrieve all the relevant data for a certain task.

All three aspects will be developed in the context of this thesis. The ontology provided by this study can be extended to other Environmental Information Systems, which covers one or more domain(s) other than water-related domain.

In general, the study focuses on an ontology-based description of data in order to facilitate users to efficiently search for data. To reach that goal, it is vital to answer the following research questions.

- **How to apply ontology to describe semantic of data sources?** The existing works on ontology are reviewed in order to figure out the reasonable way to present the semantic of data source.
- **How to describe data in relationships of observed objects and phenomena?** The influence between phenomena and observed objects are adopted from current definitions and common knowledge. This is crucial part of the study; it determines the relevant datasets for a particular user search. This is an innovation way to facilitate user search for data.
- **How to improve user search for data in the context to their tasks?** Ontology is applied to describe the semantics of a dataset. The dataset attributes, the observed objects and tasks are described in separate domains such as the data domain, observed object domain and application domain. These three domains are connected via constraints that are defined by properties and rules. Using this system, users only need to provide their tasks and the observed object of interest to the system; as a result the system will return data based on predefined constraints stored in the ontologies. Users don't have to search for thematic groups or to search in a trial and error approach several times to retrieve all relevant datasets from the system.

The assessment of the returned result is also considered to prove the feasibility of the approach. The approach provides an effective way of searching data or information. Users can get all relevant data for their tasks in an optimal system just by one search. Analyzing returned results from system and comparing it to user's expectations will be done. Based on that an evaluation will be done in order to improve the ontologies and specify the missing attributes in the database. The feedbacks from the evaluation will be used to improve database schema.

1.3. Structure of the thesis

To fulfill the research objectives, this thesis is structured into seven chapters. The chapters are briefly described as follows:

Chapter 1 Introduction and objectives: depict the issues regarding to the semantic heterogeneity of data that motivate the thesis "the Knowledge-based search for water-related information system for the Mekong Delta, Vietnam", then define some special terms used in this thesis. The objectives are presented as new way to search for data.

Chapter 2 Literature review: Reviews the current state of the art literature related to this study. The ideas and reused ontologies are presented in conclusion section. This chapter presents the current ideas on how to applied ontology to solve the data semantic heterogeneity.

Chapter 3 Methodology: Introduces the study method to achieve an ontology-based discovery of water-related information. This section presents the approach for an ontology-based retrieval, including a description of the workflow of the user interface.

Chapter 4 Collected data in WISDOM: This chapter analysis the variety of collected data in the WISDOM IS and how they are organized. This chapter also describes the data structure and its heterogeneity and analysis how difficult it is to manage the data in terms of semantic.

Chapter 5 Implementation: Prototypical implementation of the study case, which includes the technical description of the applied query language, software and programming language.

Chapter 6 Evaluation: Assessment of research result regarding advantages and disadvantages.

Chapter 7 The summary of findings, the conclusions and recommendations are presented in this chapter.

2. LITERATURE REVIEW

2.1. Introduction

This chapter reviews existing approaches in the field of applying ontology for web based information systems. The review does not only focus on applying ontology for data discovery, which were integrated into the system from different research fields, but it also provides a review of relevant semantic problems for web based information systems, such as ontology mapping and connection between an ontology and a SQL database.

For the field of Geo-Information, integrating data from different sources to provide the value added information by combining and analyzing different data is the most important objective (Zhao et al. 2005), because data is distributed and heterogeneous, which makes it difficult to achieve precise query results. Limitations due to heterogeneity have been mentioned by Stuckenschmidt, Friis-Christensen (Stuckenschmidt 2003; Friis-Christensen et al. 2005). Reasons for heterogeneity are due to different syntax, different structure or different semantic (see section 1.1.1).

The data heterogeneity have been addressed by the Open Geospatial Consortium - OGC (OGC 2012d), who develops and promotes standards for open interfaces, protocols, schemas etc. to exchange geospatial data and instructions between different systems, by defining voluntary specifications to enable syntactic interoperability. OGC provides specifications for standardizing service interfaces and exchange formats, these specifications provide a common format for representing geographic information avoiding syntactic heterogeneity (Lutz et al. 2006). It also enables the cataloguing of geographic information (Klien et al. 2004). Though the OGC-Compliant catalogues support discovering, organization, and access to geographic information, they do not yet provide methods to solve problems of semantic heterogeneity (Bernard et al. 2004; Klien et al. 2004), thus the returned results are too narrow or too large (Hochmair 2005) .

To deal with structural heterogeneity, specifications of ISO 19115, the metadata, defines how to describe geographical information and associated services, including the identification, the extent, the quality, the spatial and temporal schema, spatial reference and the distribution of digital geographic data. It may be used for other forms of geographic data such as map, charts, textual documents (ISO 19115:2003). Currently, there is no standard that deals with semantic heterogeneity.

In general, current available geographic information systems are organized by spatial, thematic and temporal aspects (Bernard et al. 2005; Athanasis et al. 2009; Podwyszynski 2009; Gebhardt et al. 2010b). Users can explore data by querying thematic, regional and temporal attributes. Results can be browsed or downloaded for further analyses or processing (Athanasis et al. 2009). The systems can manage various spatial and non-spatial datasets and their distinct aspects. However, it can be difficult, especially for novice users to give the correct search terms. They cannot estimate how many filter criteria should be utilized in order to find the data which are the most relevant to their task. If the user is looking for something that has not been categorized in the way she or he thinks, they will get imprecise results (Hochmair 2005; Athanasis et al. 2009).

The information retrieval techniques are commonly based on a specific encoding of available information, e.g. fixed classification codes, or simple full-text analysis (Visser et al. 2002). One real-world object can be described by different terms (synonyms). For example, water level can be water height or water depth. On the other hand, one term can describe different objects (homonym) (Zhao et al. 2006). Thus, keyword search may return results, which do not really relate to what a user wants to search for (Bernstein et al. 2002; Bernard et al. 2004). The underlying problem is that keywords are a poor way to capture the semantics of data or information (Lutz et al. 2009). As a result, it is hard for users to search and retrieve appropriate data for a certain task.

In short, within the existing systems, the returned result from the system is sometimes mismatch or inappropriate to the query because of the missing implementation of semantic capabilities. As a result, users have to change keywords or search criteria several times. In the worst case, they are not able to find the data or information they need, even if it exists in the system (Hochmair 2005). Furthermore, it is time consuming and lowers the acceptance of such a system tremendously when they need relevant data for a certain task

(Washington et al. 2008). In that situation, they must search several times for each particular dataset and related documents by modifying their search parameter. In other words it can be also described as a trial and error or searching by iteration. For example in the case of WISDOM Information System (IS), users want to analyze the land cover affected by flood within the WISDOM IS. Therefore, they have to search for water mask datasets (the datasets present the distribution of surface water), land cover datasets from satellite images, province or region area, legal documents and planning programs of the current region, etc. Every dataset belongs to different categories, so the users have to go through each category by hand and search for more than one time to get all data (Tran et al. 2010) (see chapter 4 for more detailed in how user can search for data and how data are organized in WISDOM IS).

To answer the user queries, it is not always the case, that databases have exact data to meet the user request, but the system can provide similar or relevant data. Even in such a case, it is also difficult to retrieve all relevant data for a certain task (Nigro et al. 2008). The problem here is not only a lack of data, but there is also the issue that a lot of data are returned from the system. Sifting through all to find relevant information can be a complicated, lengthy and frustrating process (Washington et al. 2008). These constrains can be resolved, by implementing a semantic description of data as well as the description of thematic reference groups (Zhan et al. 2008).

Geodata are organized in several models by different aspects (Becker et al. 2012). The solution to semantic heterogeneity relies on ontology since it provides a formal specification of the mental model underneath geodatasets (Navarrete 2006). Ontology emerges as best solution to solve the semantic problem of data for particular domain (Xiao 2006) . It is a “formal, explicit specification of a shared conceptualization” (Gruber 1995) playing a vital role in describing the meaning of the data in which the computer can understand data to apply meaningful data discovery automatically (Zhan et al. 2008), providing semantic descriptions to offer more precise results for user requests. It is useful not only for sharing understanding, but also for evolving as a basis for improved data usage, achieving semantic interoperability, developing advanced methods for representing and using complex metadata, correlating information, knowledge sharing and discovery (Fensel 2001). That means, ontologies do not only describe the meaning of datasets, but

they can also describe the relations between datasets in order to provide all relevant data or information for a certain user request.

2.2. State of technology

This section presents the state of technologies which can deal with the heterogeneity of data.

2.2.1. Existing standards

Geodata are published on web based information system using geospatial web services (Khaled et al. 2010). The geospatial web services change the way of designing, developing and deploying spatial information systems and applications (Zhao et al. 2006), and help users to access geodata. However, data is heterogeneous. It comes in various formats. It is organized under different schemas and may use different terms to describe its meaning. Thus, it is necessary to have a solution to ensure interoperability between different systems (Zhang et al. 2005). OGC and ISO/TC211 play main roles in standardizing geospatial web services, especially in designing interoperable software components for the access and processing of spatial data (Orchestra 2007). These standards are described in detail in following sections.

2.2.1.1. The Open Geospatial Consortium (OGC) Standards

The geodata are scattered via WWW since they are published using different formats and schemas. The users want to have access to data and information from several systems without copying and converting whole datasets (Riedemann et al. 2003; Bacharach 2008). The OGC has developed open standards in order to meet these needs. “The OGC is an international industry consortium of 483 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards”

(OGC 2012d). It provides specifications for standardizing service interfaces and exchange formats. These specifications define a common format to represent geographic information avoiding syntactic heterogeneity. According to (Bacharach 2008), some of the main OGC standards can be listed as below

- The OpenGIS® Sensor Observation Service (SOS) provides a web based interface to make sensors and sensor data archives accessible (Bacharach 2008; 52North 2012; OGC 2012c).
- The OpenGIS® Open Location Service Interface Standard (OpenLS) specifies interfaces which enable integration between different wireless networks and devices. That provides access to multi content repositories and service frameworks (Bacharach 2008; OGC 2008).
- The OpenGIS® Simple Feature Interface Standard (SF) defines the common way to store and access features in vector data, i.e. points, lines and polygons, etc.
- The OpenGIS® Geography Markup Language Encoding Standard (GML) defines an XML grammar for expressing geographical features.
- The OpenGIS® Catalogue Service Interface Standard (CS) provides interface for publishing and accessing collections of descriptive information (metadata) about geospatial data, services and related resources (Bacharach 2008).
- The OpenGIS® Web Map Service Interface Standard (WMS) produces geo-registered map images from distributed geospatial database for a certain query over the internet (OGC 2004).
- The OpenGIS® Web Feature Service Interface Standard (WFS) provides a way to retrieve or modify individual features of geodata on the internet.
- The OpenGIS® Web Coverage Service Interface Standard (WCS) enables interoperable access to geospatial "coverages" such as satellite images, digital aerial photos, and digital elevation data (Bacharach 2008).

This section focuses only on the standards which deal with the data syntactic heterogeneity, i.e. WMS, WFS WCS. They provide the interfaces for the access to geospatial data from one or more sources (Yanfeng et al. 2006; Stopper et al. 2011).

The WMS provides a simple Hypertext Transfer Protocol (HTTP) interface, the application-level protocol that is used to transfer data on the web (SiliconPress 2002), to retrieve and display georeferenced map images from multiple remote and heterogeneous sources (OGC 2004; Stopper et al. 2011). In the request, the users define the area of the earth surface where they want to focus on, and the layer of data. The results returned from the server are the graphical visualization of geospatial data which come simultaneously from multi heterogeneous source in a standard image format (Zhang et al. 2005), i.e. georeference map images such as JPEG, PNG, etc. (Amirian et al. 2008). And then, the images can be displayed in a browser (Gwenzi 2010). By the end of 2005, the WMS became ISO standard, the ISO 19128:2005 Geographic information – Web map server interface (OGC 2005).

The WFS provides a way to create, modify and exchange geographic information on the Internet (OGC 2010b). Using WFS, The geometric descriptions of features in geodata returned from the WFS server are encoded in GML from multiple sources (Zhang et al. 2005; Stopper et al. 2011). The WFS server receives, reads and executes the request from the users. And then it returns the result in a feature set encoded in GML. The WFS becomes the ISO 19142:2010, the geographic information – web feature service (OGC 2010b).

The WCS provides a standard interface and operations that enables interoperable access to geospatial "coverage", i.e. satellite imageries, digital aerial photos and digital elevation data (OGC 2010a; OGC 2012b). It can be considered as an extension of WMS and WFS, since they cannot access coverages. The result returned from a WCS server is information about coverage and an output coverage which is encoded in a specified binary image format, such as GML, GeoTIFF.

The standards mentioned above provide an access to heterogeneous database (Amirian et al. 2008) by resolving the syntactic heterogeneity, however, there are still some constraints to be resolved such as semantic interoperability issues (Zhang et al. 2005).

2.2.1.2. The International Standardization Organization (ISO) standards

The International Standardization Organization (ISO) is the world's largest developer of voluntary International Standards (ISO 2012). ISO defines several standards such as documented agreements containing technical specifications or other precise criteria to be used consistently, i.e. rules, guidelines, or definitions of characteristics, to ensure that products, materials, processes and services are fit for their purposes.

In order to resolve the structural heterogeneity of the database (see section 1.1.1), the ISO/TC211 (ISO Technical Committee) defines several standards for geographic information (Ostensen et al. 2002). They provide many standard groups as shown below (ISO 2009) (see more details in Appendix A).

- Standards for specifying the infrastructure for geospatial standardization: infrastructure for the further standardization of geographic information (ISO 19101, ISO/TS 19103, ISO/TS 19104, ISO 19105 and ISO 19106)
- Standards for describing data model for geographic information: abstract conceptual schemas for describing the fundamental components of features as elements of geographic information (ISO 19109, ISO 19107, ISO 19137, ISO 19123, ISO 19108, ISO 19141, ISO 19111 and ISO 19112)
- Standards for geographic information management: focused on individual features and their characteristics, these standards are focused on the description of data sets containing information about one or, typically, many feature instances (ISO 19110, ISO 19115, ISO 19113, ISO 19114, ISO 19131, ISO 19135, ISO/TS 19127 and ISO/TS 19138)
- Standards for geographic information services: support the specification of geographic information services (ISO 19119, ISO 19116, ISO 19117, ISO 19125-1, ISO 19125-2, ISO 19128, ISO 19132, ISO 19133 and ISO 19134)
- Standards for encoding of geographic information: encoding standards are needed to support the interchange of geographic information between systems (ISO 19118, ISO 6709, ISO 19136 and ISO/TS 19139)
- Standards for specific thematic areas: standards is the area of geographic imagery (ISO/TS 19101-2 and ISO 19115-2)

Among standards defined by ISO, this review chapter just focuses on the ISO 19115:2003, the metadata, which deal with the structural heterogeneity of data. Metadata is often called data about data or information about information. Metadata is structured information. It describes and explains information resources and makes it easier to retrieve, use, or manage data (NISO 2004).

The ISO 199115:2003 is applicable to (ISO 2003):

- The cataloguing of datasets, the clearinghouse activities ("the Clearinghouse is a mechanism to exchange information and coordinate activities to enhance peace operation capacity building efforts"[(1)]), and the full description of datasets;
- Geographic datasets, dataset series, and individual geographic features and feature properties.

The ISO 19115 consists of more than 300 metadata elements (86 classes, 282 attributes, 56 relations) (Mavratza et al. 2007) (see the Apendix B for the full list of ISO 19115). However, most of the elements is optional, typically only a subset of elements, which is called the core, is used. The core elements mainly focus on describing the characteristics of a datasets to identify it, typically for catalogue purposes (ISO 2003; Mavratza et al. 2007). The core metadata mostly focus on answering the following question: (i) what does the topic the dataset relate to? (ii) which region does the dataset describe? (iii) what is the period of time when the dataset is valid? and (iv) Who is the contact person if the user wants to know more about or order the dataset? (ISO 2003). The core set consists of three kinds of elements (Table 2.1):

- Mandatory (M): mandatory elements
- Conditional (C): conditional elements. These elements are mandatory if a certain condition has been met.
- Optional (O): optional elements.

(1) http://www.state.gov/t/pm/ppa/gpoi/c20213.htm

Dataset title (M) (MD_Metadata > MD_DataIdentification.citation > CI_Citation.title)	**Spatial representation type (O)** (MD_Metadata > MD_DataIdentification.spatialRepresentationType)
Dataset reference date (M) (MD_Metadata > MD_DataIdentification.citation > CI_Citation.date)	**Reference system (O)** (MD_Metadata > MD_ReferenceSystem)
Dataset responsible party (O) (MD_Metadata > MD_DataIdentification.pointOfContact > CI_ResponsibleParty)	**Lineage (O)** (MD_Metadata > DQ_DataQuality.lineage > LI_Lineage)
Geographic location of the dataset (by four coordinates or by geographic identifier) (C) (MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_GeographicExtent > EX_GeographicBoundingBox or EX_GeographicDescription)	**On-line resource (O)** (MD_Metadata > MD_Distribution > MD_DigitalTransferOption.onLine > CI_OnlineResource)
Dataset language (M) (MD_Metadata > MD_DataIdentification.language)	**Metadata file identifier (O)** (MD_Metadata.fileIdentifier)
Dataset character set (C) (MD_Metadata > MD_DataIdentification.characterSet)	**Metadata standard name (O)** (MD_Metadata.metadataStandardName)
Dataset topic category (M) (MD_Metadata > MD_DataIdentification.topicCategory)	**Metadata standard version (O)** (MD_Metadata.metadataStandardVersion)
Spatial resolution of the dataset (O) (MD_Metadata > MD_DataIdentification.spatialResolution > MD_Resolution.equivalentScale or MD_Resolution.distance)	**Metadata language (C)** (MD_Metadata.language)
Abstract describing the dataset (M) (MD_Metadata > MD_DataIdentification.abstract)	**Metadata character set (C)** (MD_Metadata.characterSet)
Distribution format (O)	**Metadata point of contact (M)**

(MD_Metadata > MD_Distribution > MD_Format.name and MD_Format.version)	(MD_Metadata.contact > CI_ResponsibleParty)
Additional extent information for the dataset (vertical and temporal) (O) (MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_TemporalExtent or EX_VerticalExtent)	**Metadata date stamp (M)** (MD_Metadata.dateStamp)

Table 2.1: Core metadata for geographic datasets (ISO 2003)

As shown on Table 2.1, ISO 19115:2003 provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, distribution of digital geographic data and a method for extending metadata to fit specialized needs as well (ISO 2003). The datasets are described at least by title, topic (thematic group), the valid period of time of the dataset, language use in dataset and a short introduction. The metadata provides the user an over view about the dataset, so that the user can make a better choice on which dataset they need.

However, these descriptions are machine readable, not machine understandable. It causes some issues for the user when they use another language which is not used in the metadata. There is also constraint for integration efforts when the systems use different measure systems, for example kilometre and mile. Moreover, the different research fields may use disparate terms to describe the same real world object. That is the homonym or synonym case.

In short, by applying the specifications of ISO 19115:2003, the structural heterogeneity of data from different sources can be resolved (ISO 2003). However, semantic interoperability remains unaddressed in these standards (Gwenzi 2010), e.g. the relationship between datasets in terms of the observed object (see section 1.1.2.1).

2.2.1.3. Summary

As mentioned above, the current standards of OGC and ISO can resolve the syntactic and structural heterogeneity of data (see section 1.1.1), however, they cannot describe geodata in a context, what they mean and how they relate to each other in different research fields, for example how the data of water level relates to flood information, and how flood information relates to another information such as residential area, rice fields, etc. The descriptions using existing standards are human readable, but structured information extraction via a machine is hardly possible. Ontology is one of the candidates which can solve the constraints of these standards. It is described in more details in the next sections.

2.2.2. Ontology

In order to solve several of aforementioned issues, a "Concept of Ontologies" has been initially introduced by (Gruber 1995), described as "a formal explicit specification of a shared conceptualization". Conceptualization refers to an abstract model of how people commonly think about a real thing in the world. The concepts and relations have explicit names and definitions, the so-called explicit specification. Knowledge described in the ontology is accepted by a community via a shared conceptualization that enables reuse of domain knowledge.

Ontology plays a main role in developing a way to share common understanding of information among humans and software agents (Musen 1992). In this way, it is a fundamental prerequisite to improve data usage, achieving semantic interoperability, developing advanced methods for representing and using complex metadata, correlating information, knowledge sharing and discovery (Noy et al. 2001). This is achieved by a set of predefined vocabulary in certain areas of expertise, and relationships between them (Gruber 2008), which can be understood by both humans and computers. Ontology includes the following components (W3C 2009):

- Classes are a key component of ontology, also known as the concept. Most ontologies are focused on building classes, which are organized in a hierarchical structure to describe the types of objects in a domain of interest. For example,

"organisms" is a class in the context of biology. A class may have subclasses such as "animal" and "plant".

- Aspects (slots) are properties of each concept describing various features and attributes of the concept. For example, the concept of organisms can be described by aspects of the situation with the properties of motion; it is “moving” or “standing”. Formally, aspects mean the relationship between individual types and attributes, between individual and classes or between classes. However, in some cases the term property or role is used rather than aspect.
- Constraints (role restriction or Facet) are description of restrictions on the meaning of the concepts and relations between concepts. The motion condition in the above example has two values, but only one value at a certain time can be applied. Organisms cannot “move” and “stand” at the same time.

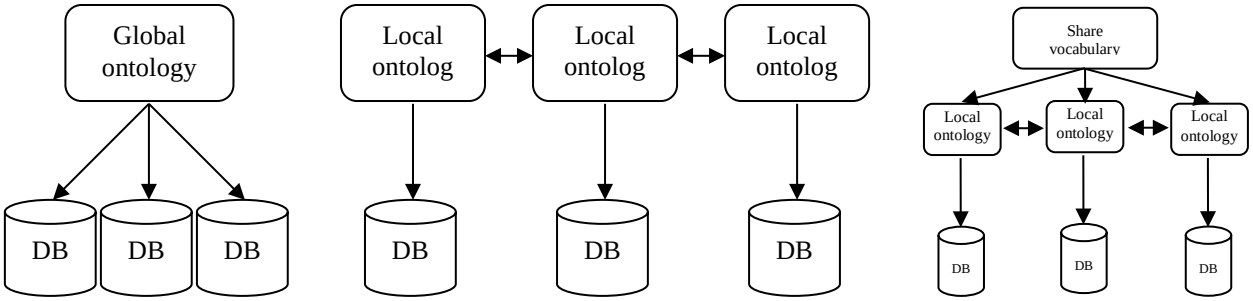
An ontology, together with a set of individual instances of classes, constitutes a knowledge base (Noy et al. 2001). The individuals are defined as objects perceived from the real world such as peoples, animals or automobiles etc. Ontology, with these components, can describe the semantics of the information sources and makes the contents explicit.

Although, ontologies are used for the explicit description of the information source semantics, there is no single correct methodology for designing an ontology (Noy et al. 2001). There are three different ways to apply ontologies (Wache et al. 2001) (i.e. single approach, multi approach and hybrid approach).

Single approach (Figure 2.1a) has one global ontology; all the information sources are related to only one ontology. It can be considered as a hierarchical, terminological database. It may consist of several specialized ontologies. It is useful when all information sources to be integrated provide nearly the same view on a domain. Single ontology approach is susceptible for changes in the information source, because it needs changes in the global ontology and in the mapping to the other information sources. SIMS (Services and Information Management for decision Systems) (Arens et al. 1996) is a typical example for this approach.

Multi approach (Figure 2.1b) describes each information source by separate ontologies, so it simplifies integration and supports changes in source. There is no shared vocabulary

between ontologies, so inter-ontology mapping is needed to communicate between information sources. With this approach it is difficult to compare different source ontologies, because it does not have a common vocabulary. An example of this approach is the OBSERVER system (Mena et al. 1996).



a) Single approach b) Multi approach c) Hybrid approach

Figure 2.1: Ontology approaches
(source: (Wache et al. 2001))

Hybrid approach (Figure 2.1c) is a combination of the aforementioned approaches. Each source is described by its own ontology and shared vocabulary is built to share the basic terms of a domain. The advantages are that new sources can easily be added, it supports acquisition and evolution of ontologies and source ontologies are comparable because of shared vocabulary. However, existing ontologies cannot easily be reused; designers have to redevelop from scratch. The framework in (Cruz et al. 2003) is example for this approach. The proposed approach on this thesis applies the hybrid approach. The system is described in independent domains, they are linked together via relationships and properties (see chapter 5 for more details).

To apply ontology, RDF (Resource Description Framework), was published in 1999 by W3C, can be used to describe objects (called resources) and their relationships on the web in a machine-understandable way (W3C 2010a). In other words, metadata of data is available on the web. RDF uses a simple structure statement "Subject – predicate – object" to describe resources or to present the relation between resources in structure resource – property – resource/literal (Figure 2.2). RDF uses an extensible URI-based vocabulary

with the XML syntax; hence exchange between different operating systems is easily possible. Any resource can be described with RDF statement (W3C 2010b).

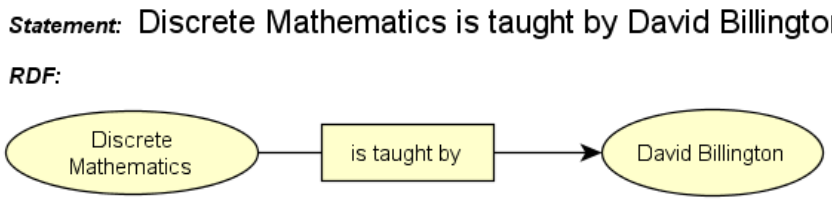


Figure 2.2: Example RDF

Source (Antoniou et al. 2008)

Since RDF is limited to the description of resources with classes, properties and values, further schema based on RDF was developed to extent the functionalities of RDF and broaden the potential application. **RDFs** (Resource Description Framework Schema) was developed based on RDF characteristics but it is extended to describing about classes of resource and their properties (Figure 2.3) such as class and subclass relations, "domain" and "range" restriction of properties. RDFs does not provide actual application-specific classes and properties, but the framework to describe it. Classes in RDFs are much like classes in object-oriented programming languages. This allows resources to be defined as instances of classes, and subclasses of classes. (W3C 2010b)

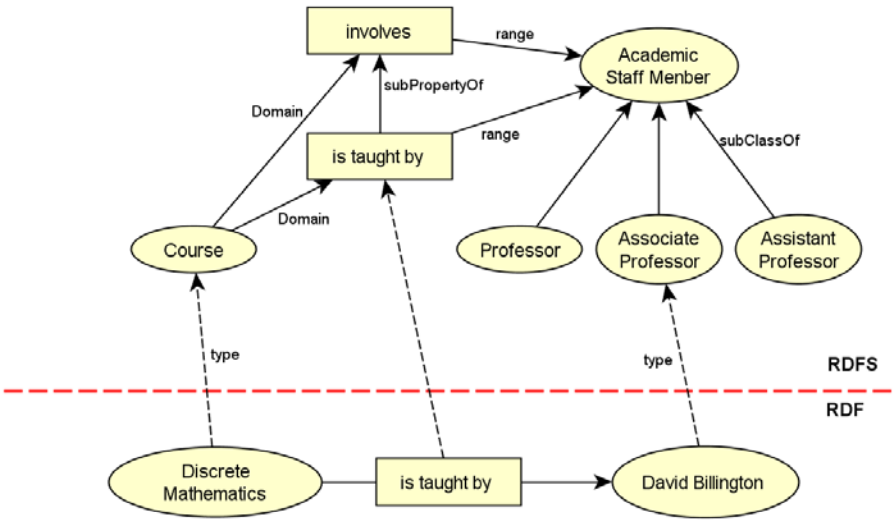


Figure 2.3: Example RDF and RDFs

Source (Antoniou et al. 2008)

RDF and RDFs were developed to provide basic capabilities for describing resources, but they specify fairly loose constraints on vocabularies. **OWL** (Web Ontology Language) was built on top RDF and RDFs adding supplemental constraints to increase the accuracy of implementation of a given vocabulary. RDFs specifies fairly loose constraints on vocabularies. OWL adds supplemental constraints that increase the accuracy of implementations of a given vocabulary. These constraints allow additional information to be inferred from the data, though it may not be explicitly represented in an ontology (for example if an individual Martin is in class Student, and the class Student is a subclass of the class Person, a reasoner will infer that Martin is a Person) (Powers 2003). OWL uses XML syntax and is a recommendation of W3C for semantic web (W3C 2012).

The **Reasoner** is a program able to infer logical consequences from a set of asserted facts or axioms. There are many available reasoner, such as Racer, Pellet, Fact++, Hermit etc. (Pan 2005; Tsarkov et al. 2006; Sirin et al. 2007; Fahad et al. 2008). A comparison of them are presented by Dentler (Dentler et al. 2011). The reasoner comparison is beyond the scope of this thesis.

2.2.3. Database Connection

OWL describes data and relationships between data items strictly. But, actually, data itself are mostly stored in a relational database (RDBs). Data have to be transfer to RDF at lower level, and then constraints of OWL can be applied at higher level. To access existing database content without replicating the entire database into RDF, the mapping of vast quantities of data from RDB to RDF has been the focus of many researches in diverse domains and has led to the implementation of generic mapping tools as well as domain specific applications (Wu et al. 2006; Roset et al. 2008; Ramanujam et al. 2009; Sahoo et al. 2009; Freitas et al. 2011). This study does not focus on the method on what is the best mapping methodology from RDB to RDF. The comparison of tools and languages to map RDB to RDF is presented in (Hert et al. 2011). Hert et al. compare nine different mapping

languages. Based on their conclusion, this thesis applies D2RQ as mapping tool. A detailed analysis would go beyond the scope of this thesis. This section describes only applied tools and languages within this study. These tools and languages are applied in order to avoid duplicate content of the RDB into RDF and to achieve an efficient way of querying data.

D2RQ, an open source software, is one of the most widely used mapping languages due to its flexibility and compatibility. It can be used to specify which concepts on the ontology correspond to which concepts in the database (Roset et al. 2008). Then users can access a database with an ontology based query language.

One of the most popular programming languages nowadays is Java, it is a high-level programming language developed by Sun Microsystems (Java 2012). Java can develop and run on any device equipped with Java Virtual Machine (JVM). Jena has been selected as Java development environment because it is an efficient open - source framework for Java based on W3C recommendation for RDF and OWL. Among a lot of features to efficiently implement ontology features, Jena provides a programmatic environment for RDF, RDFS and OWL, SPAQRL (Simple Protocol and RDF Query Language) including a rule - based inference engine (Curé 2005).

SPARQL defines a standard query language and data access protocol to be used with RDF data model. SPARQL works for any data source that can be mapped to RDF. SPARQL allows users to write globally unambiguous queries; it can explore data by querying unknown relationships, perform complex joins of disparate databases in a single, simple query and transform RDF data from one vocabulary to another to extend ontology (MSDN 2012) (see the Appendix H for more detail). The Pellet reasoner (Sirin et al. 2007) will be used to infer additional information based on OWL rules and constraints. It is an open-source Java based OWL reasoner. It can be used in conjunction with Jena libraries.

To create RDF file, Protégé is used. It is an open source Java tool, is extensible, and provides a plug-and-play environment that makes it a flexible base for rapid prototyping and application development (Protégé 2012). It implements a rich set of knowledge-modeling structures and actions that support the creation, visualization, and manipulation of ontologies in various representation formats. Protégé offers a graphical user interface

which allow ontology developers to focus on conceptual modeling without requiring to know syntax of an output language, such as RDFS or OWL (Noy et al. 2001).

2.3. Research Review

Research in the field of information discovery and retrieval is manifold. The review in this section focuses on approaches, which apply ontology in the field of ontology mapping, data integration, and task ontology. Although, only data integration and task ontology are related to this study, ontology mapping is briefly reviewed to complete the picture of applied ontology approaches.

2.3.1. Data Integration

Data integration involves combining data originating from different sources and providing users with a unified view (Lenzerini 2002). Data integration emerges from the increase of the need to share existing data. It is subject of extensive theoretical work with numerous unsolved problems.

Although, the emergence of Extensible Markup Language (XML), which was designed to transport and store data (W3C), has created a syntactic platform for web data standardization and exchange, it has several limitations. Schematic data heterogeneity may still persist, depending on the XML schemas used, e.g. nesting hierarchies. In addition, semantic heterogeneity may persist even if both syntactic and schematic heterogeneities do not occur, e.g., naming concepts differently (Cruz et al. 2005).

Since, ontologies provide an explicit and formal specification of a shared conceptualization, and are able to facilitate knowledge sharing and reuse. Ontology emerges as a solution to solve the heterogeneity of data integration. According to (Fonseca et al. 2002), an ontology represents a view of what exists in the world; a database schema represents what is stored in the database. The information that exists in the databases has to be adapted to be compatible with ontology classes. Fonseca et al. (Fonseca et al. 1999)

proposed an Ontology-Driven Geographic Information System (ODGIS) which can solve the problem of different conceptualization of the same real-world object. ODGIS acts as a system integrator. Ontology in such a system is a component, such as a database, cooperating to fulfill the system's objectives. The proposed system in ODGIS includes an ontology editor and its embedded translator plus a user interface to browse ontologies. Similarly, (Durbha et al. 2009) applied ontology approach to describe the theme in order to match disparate thematic definition schemas. Bernard (Bernard et al. 2003) in meanInGS project (Semantic Interoperability by means of Geoservices project) use ontology to describe the meaning of data and event service in order to overcome the language mismatch (synonyms and homonyms).

A further work (Podwyszynski 2009) presents an approach, that describes satellite imageries based on properties of imageries and related applications (Figure 2.4). The system is divided into two domains, the application domain and data domain; the application domain describes the application or phenomenon which users are working with. The data domain describes the properties of satellite imageries. The two domains are related with each other through a measurement component, e.g. sensor characteristics. Users are able to search data based on applications, e.g. sea surface temperature, which they are interested in without any knowledge of low level data characteristics. The approach prosed by Podwyszynski focus on only one research field that is to provide satellite imageries sufficient for user demands. He is not interested in providing relevant data for user search.

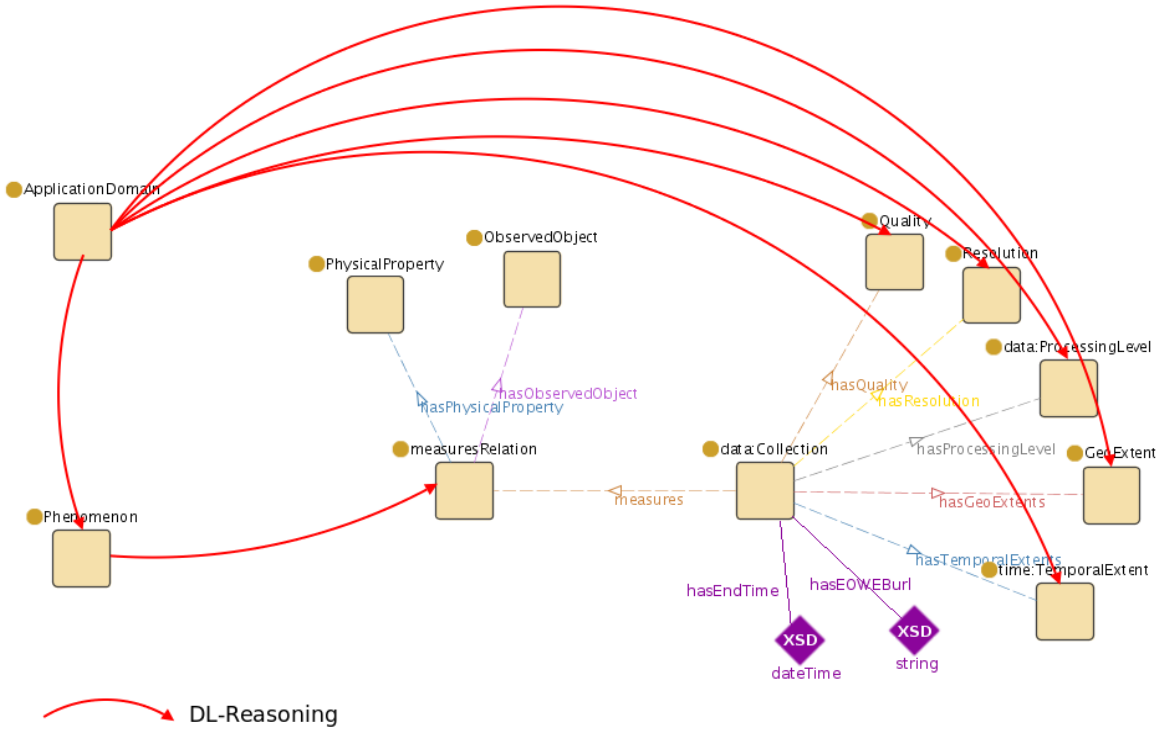


Figure 2.4: Marius Podwyszynski's approach
Source (Podwyszynski 2009)

The above mentioned approaches applied ontology in order to provide sufficient data for user query belonging to applications, but they do not describe how the data relate to each other. These descriptions are useful in case of no data exactly match to user requirement, in such a case, relevant data should be inferred providing to user.

There is an issue of scale conflicts which occur when attributes have different units or are presented in varying scale of measures (Vaccari et al. 2009), e.g. Different authors of geological maps have used different stratigraphic classifications at different times in history, leading to several synonymous and homonymous stratigraphic terms within the geological database. To solve the problem for integrating collected data from different classification system, M. Lutz (Lutz et al. 2009) proposed a hybrid ontology approach that has a Semantic Translation Specification Service (STSS) (Figure 2.5). The user request will be sent to STSS to find the appropriate data by using an ontology-based reasoner and then send data back to browse or download. This approach also provides a *Query Template* allowing users to construct queries based on the concepts of the ontology. Nevertheless, these templates limit the range of expressions of possible user queries. The concept

description can also be conceived as a query, so that the query concept can be either the concept description itself (existing concept of or from the domain) or a concept defined by users based on the concept and relation in the shared vocabulary. However, the STSS in this approach just focuses on translating the information from different classification or definition systems, for example different geological classification systems, different terminology. This approach focuses only on one thematic field.

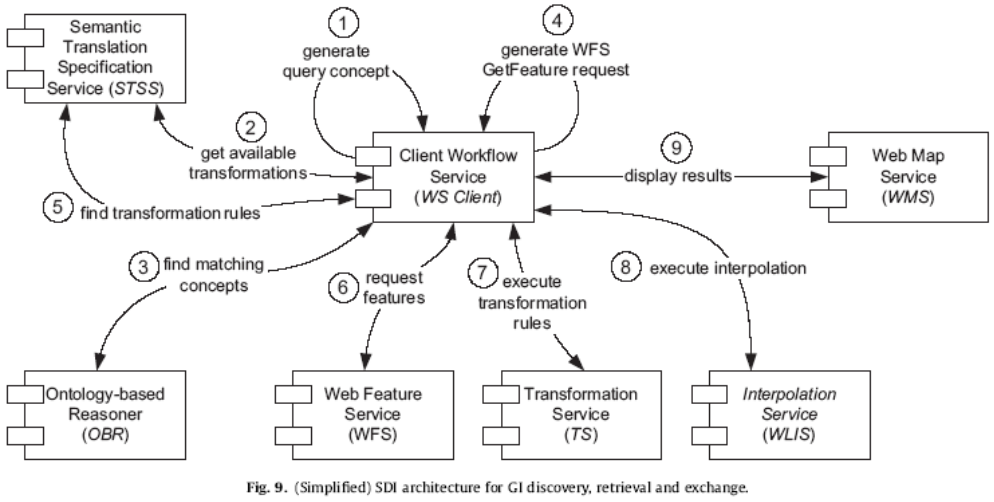


Figure 2.5: Semantic Translation Specification Service in M. Lutz proposed approach (Source: (Lutz et al. 2009))

BUSTER (Bremen University Semantic Translator for Enhanced Retrieval) project (Vöegle et al. 2003) also uses ontology to describe data content and classification systems, but they use simple ontologies and queries, which have only limited expressivity. Similar one is HarmonISA project (Hall et al. 2006), they use a complex similarity measurement between land use type definitions rather than compute the classification hierarchy (subsumption reasoning) with ontology.

Another approach to discover information on geodata services, (Athanasis et al. 2009) use multiple ontologies to describe the domain. It describes datasets under three schemas: the first one for ISO 19115, the second one for "phenomenon or theme" and the third one for data type hierarchy. Individual values (i.e. datasets) have been managed into schemas. One dataset can be classified under classes in phenomenon schema and data type schema; it also has "properties" in ISO 19115. Besides, it also has relationships with other datasets

via the "related-docs" property (Figure 2.6). With this approach, they missed the property which describes the relationship between datasets, e.g. "has-Effect-On" or "has-Event" properties. Because users normally need these properties for their research applications, e.g. when they want to search for "flood dataset" actually they want to search for datasets that can describe the effect of flood and of course some related documents regarding to the term flood. One of the advantages of approach presented by Nikolaos is addition relationship in the ISO metadata. It facilitates user search for data. But, there are no relation between schemas, thus, data providers use the graphic user interface to assign the relation of data with different schemas.

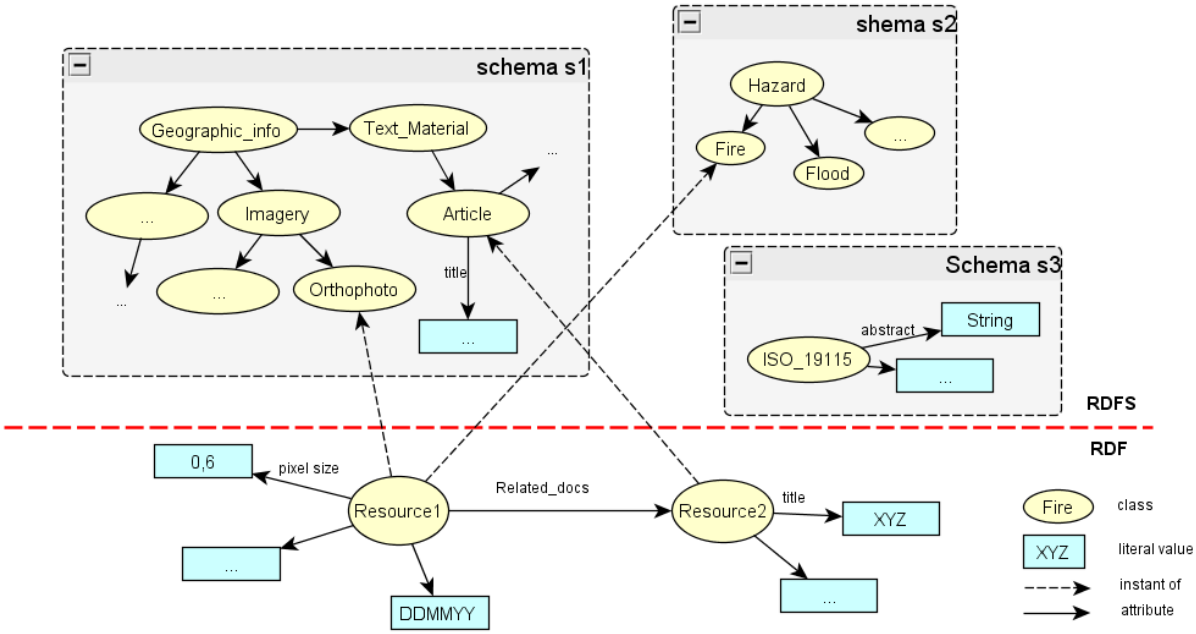


Figure 2.6: Abstract of Athanasis Nikolaos approach
(Metadata in experimental geoportal about natural disasters)
(Source: (Athanasis et al. 2009))

The approaches mentioned previously only focus on applying ontology-based search due to the meaning of information. Normally, the information will be further used in another processing program on desktop for particular application. Most of these tasks need more than one dataset, so that aforementioned approaches are still limited, because users have to search several times for information they need.

2.3.2. Task Ontology

Users need data to work for a certain task in a specific domain, for example they need water level data to monitor flood situations in the water-related domain. A user task can be defined as an action to a response to an event. A task guides cognition and perhaps even perception of objects in a given situation. Depending on the different tasks, we have different ways to observe and comprehend the reality as well as the components of the reality (Timpf 2002). According to (Guarino 1998), domain ontologies describe the vocabulary related to a generic domain, like water-related domain. Task ontologies describe the vocabulary related to a generic task, like monitoring or planning. Task ontologies are needed for domain ontologies for knowledge sharing, interoperability, and re-use of services because task ontologies can describe the reasoning concepts and their relationships occurring within a certain domain for a specific task (Timpf 2002).

There are approaches (Tran et al. 2007; Ikeda et al. 2009; Ren et al. 2010) which focus on the description of processing steps of a user task in order to discover the available services and then execute a business process. The task can be described through verb and noun extraction from a written description of the reasoning process (Ikeda et al. 1997). Verbs characterize actions and nouns characterize objects. The descriptions of relationships between actions and objects, and between actions themselves produce the task ontology (Timpf 2002). There are studies aiming to build graphic user interfaces to define a user task ontology (Welie 2001). However, those aspects are not subject of this thesis. The descriptions of user task in the aspect of data mining are issues of concern. That means, user tasks are described with sufficient data, instead of describing how the task is processed.

In current approaches, ontology is not applied to fully characterize data or formalize the relationships between concepts, including specification of which datasets are needed for a certain task (Wiegand et al. 2007). According to (Timpf 2002; Wiegand et al. 2007) , after being defined, a task ontology can be connected to certain domain ontology (application domain). Then a reasoner will infer and retrieve datasets sufficient for a specified task. In fact, users have to re-think what data sources are needed every time when they have a

particular task. There are some available sources but users are maybe not aware of them. To avoid that, (Wiegand et al. 2007) proposed a task-ontology, that describes user tasks in relationship to thematic groups of data. The proposed approach aims to build task ontology to facilitate data discovery for user tasks such as emergency response or planning activities. The main parts within the system are task ontology and data source ontology. The ontological restriction on this approach is "need". It can provide the list of datasets, which are needed for a particular task, but it does not show how it relates to the task. The ranking which dataset is the most appropriate for the user task is also mentioned as a future work of her approach.

Other approaches focus on spatial ontology and service discovery. In order to provide data belonging to a region, (Cai 2002) discusses the special characteristics of geographic information for information retrieval, such as spatial footprints in addition to thematic content. (Jones et al. 2004) suggested a spatial search engine which incorporates ontologies, geographic footprints, and spatial indexing to target spatially related data. A spatial query expansion using ontology of place was used in their search engine. To facilitate service discovery, (Klien et al. 2006; Paul et al. 2006) proposed an approach applying ontology to define the query concept with shared vocabularies. The shared vocabularies are registered for different domains to avoid the problems with simple key word based search due to naming heterogeneity. However, these perspectives are beyond the scope of this study, so we do not go into details.

2.3.3. Existing Ontologies

Scalability and reusability are the attributes of ontologies to be easily extendable. Concepts for a particular domain can adopt similar concepts and ideas from existing ontologies. There are many existing ontologies describing different things on the world. Thus, the existing ontologies and thesaurus are reviewed for reuse.

The most popular ontology model in earth science is Semantic Web for Earth and Environment Terminology (SWEET) (Li et al. 2008; SWEET 2012). It is a NASA funded

initiative consisted of more than 6000 concepts in 200 separate ontologies including thousands of terms relevant to Earth System Science and related concepts such as numerical units and other datasets. Earth system science knowledge based presented in SWEET ontologies are extendable and reusable (Raskin et al. 2005). SWEET ontologies are designed using the OWL language. It represents the world into separate ontologies and interrelation between them as in Figure 2.7.

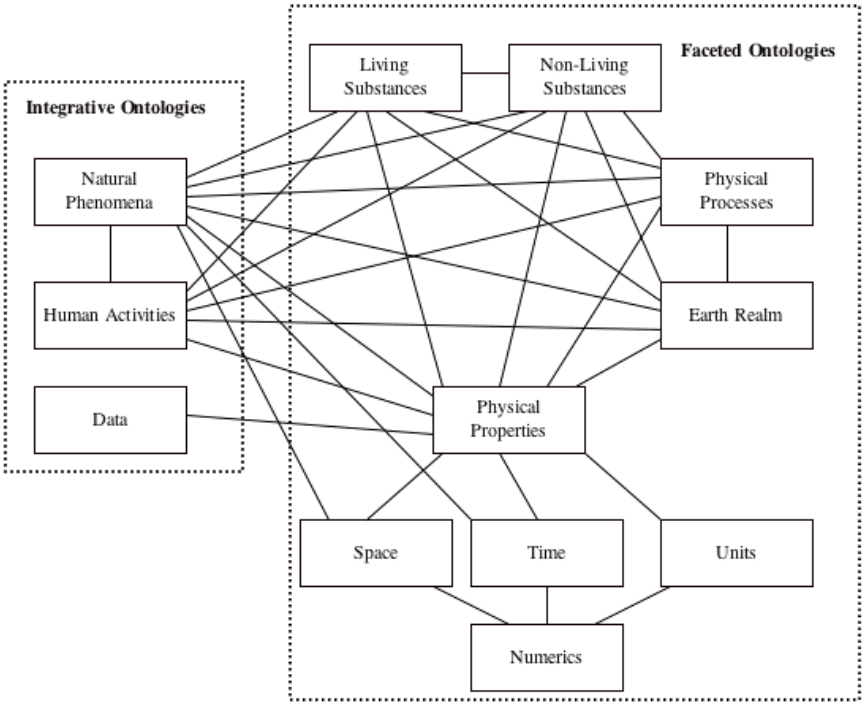


Figure 2.7: SWEET ontologies and their interrelationships
(Source: (Raskin 2005))

SWEET is divided into two groups – facets and integrative ontologies. Facets are hierarchy of homogeneous terms describing an aspect of the knowledge being codified. Terms in the hierarchy present atomic concepts (i.e. fluid and pressure) which may exist as concepts in the same or different ontologies. The structure of a faceted ontology is designed in a hierarchy structure – the higher level concepts are more general than the lowers, whereas, the integrative ontologies contain compound concepts, which are combinations of orthogonal concepts, i.e. fluid pressure. That is easy to combine different concepts for

different research fields and also extendable, since terms in facets can be added at any time.

Spatial ontology in this thesis is adopted from SWEET. It can be combined with Geonames - The GeoNames is a geographical database that covers all countries in the world and contains over eight million place names (GeoNames 2012) - in order to reuse the existing definitions about relation about administrative area such as near, neighbor etc. Concepts about phenomenon also adopt some concepts from SWEET which relate to water domain such as flood, drought (see chapter 5 for more details).

Food and Agriculture Organization of the United Nations (FAO) has defined a corporate thesaurus – the so called AGROVOC (FAO 2012). It contains more than 40000 concepts covering topics related to food, agriculture, environment and other related domains. It defines the hierarchies and relations between the terms. Some observed objects are adopted from AGROVOC to design a hierarchy classes related to land use. The AGROVOC web page is shown below (Figure 2.8). With a defined term, a list of related terms is shown including their explanations. The level of relations is defined as broader, narrower or related terms. The broader terms present a more general concept than the narrower ones. The broader and narrower terms are in the same branch of the hierarchy. The related terms are in different branch but they related to each other. For example, the term “air pollution” has broader term that is “pollution” and related terms are “greenhouse effect” and “atmosphere”.

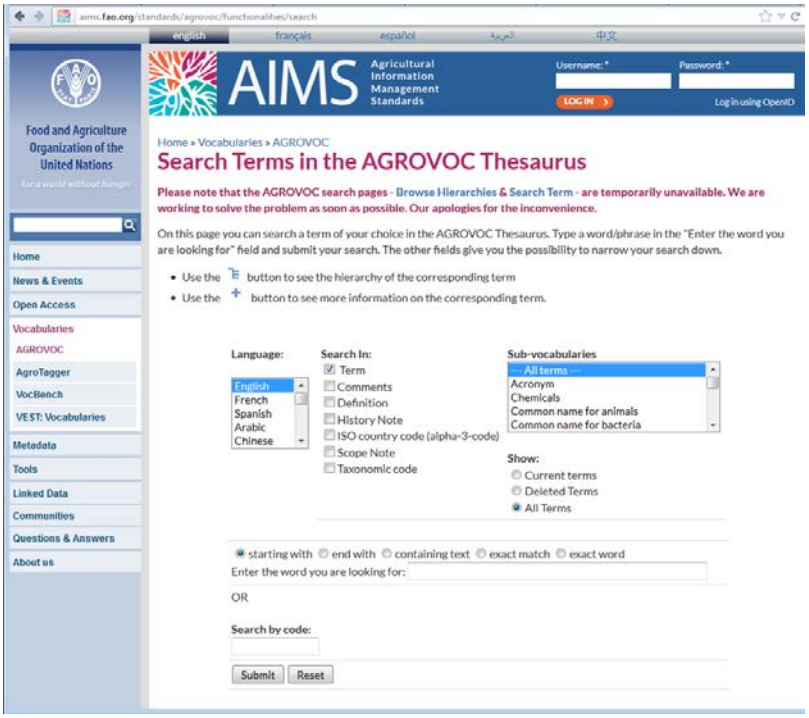


Figure 2.8: AGROVOC web page

For temporal domain, Feng Pan (Pan et al. 2005) proposed an ontology, that describes the basic temporal concepts and relations which is necessary for most simple applications, and also a recommendation of W3C. This thesis uses temporal ontology proposed by Feng Pan to store temporal value for dataset, i.e. valid date of data - start data and end date.

2.3.4. Ontology Mapping

An ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world, for example, what is human, and how does human relate to other things (Guarino 1998). However, human have different notions about real things in the world depending on their knowledge, their interest and even their culture (KnowledgeWeb 2005; Zhao et al. 2006). That situation leads to many different ways to describe the world, using ontology in a certain domain. Even in the case where a standard ontology has been established for a

particular domain, its customization to particular regions will result in heterogeneous ontologies. The problem arises because of the difference of specification of a conceptualization, and of the term in the domain and relations (Buccella et al. 2009). Many ontologies, which are in use today overlap in content (Noy 2009). Numerous attempts have been made to generate semantic "mappings" between different ontologies, or create aligned or integrated ones (Kavouras et al. 2005).

To solve the problems mentioned above, it is necessary to use ontology mappings geared for interoperability (Choi et al. 2006) in order to access data from different systems with a unifying view. Ontology mapping is the process whereby semantic relations are defined between two ontologies at conceptual level, which in turn are applied at data level transforming source ontology instances into target ontology instances.

Mapping could provide a common layer from which several ontologies could be accessed and hence could exchange information in a semantically sound manner. However, it is very common that mismatches between ontologies occur. Mismatches can are divided into two main categories: language mismatches (Chalupsky 2000; Madhavan et al. 2002), and ontology mismatches (Visser et al. 1997). Language mismatch happens when ontologies use different languages or different abbreviations or acronyms. Ontology mismatch happens when a real-world domain is represented in distinct ways, which means the concepts differ on how the world is modeled, and on how the concepts are related (for example, whale are classified as a mammal in one classification system, but it are a fish in another system). The heterogeneity also occurs during explication of the conceptualization, because the same concepts could be defined in different ways (Chalupsky 2000; Madhavan et al. 2002).

Ontology matching can be considered as an operation that takes two graph-like structures and produces a set of correspondences between the nodes of the graphs that correspond semantically to each other (Giunchiglia et al. 2007) (Figure 2.9). Then, these correspondences can be used for various tasks, including service discovery, composition and coordination, information retrieval operations, data schema mediation and translation. Thus, matching ontologies enables the knowledge and data expressed in the matched ontologies to interoperate (Vaccari et al. 2009).

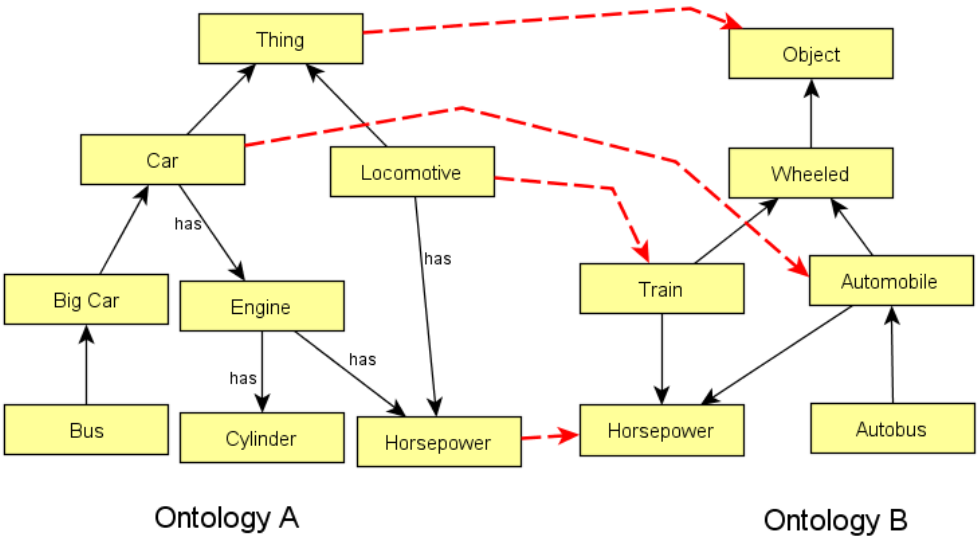


Figure 2.9: An example of ontology mapping
Source: (Abolhassani et al. 2006)

Choi categorized ontology mapping into three types (Choi et al. 2006). All three methods have in common, that they calculate the similarity between classes (concepts) within ontologies.

- Mapping between an integrated global ontology and local ontologies: This mapping specifies how concepts in global and local ontologies map to each other, how they can be expressed based on queries, and how they are typically modelled as views or queries. It is easy to define a mapping and to find mapping rules because an integrated global ontology provides a shared vocabulary, and all local ontologies are related to a global ontology. But the change of local ontologies or the addition and removal of local ontologies could easily affect other mappings to a global ontology. This mapping requires an integrated global ontology. This mapping supports the integration of ontologies for the Semantic Web, enterprise knowledge management, and data or information integration.
- Mapping between local ontologies: This category provides interoperability for highly dynamic, open, and distributed environments and can be used for mediation between distributed data in such environments. It avoids the complexity and overheads of integrating multiple sources. This mapping enables ontologies to be

contextualized because they keep their content locally. This category mapping has more maintainability and scalability because the changes (adding, updating, or removing) of local ontology could be done separately without regard to other mappings. This type of mapping does not have commonly shared vocabularies so it is difficult to find mapping rules between local ontologies.

- Mapping on ontology merging and alignment: This category allows a single coherent merged ontology to be created through an ontology merging process. The growing usage of ontologies or the distributed nature of ontology development has led to a large number of ontologies, which have the same or overlapping domains. These should be merged or aligned to be reused.

Actually, "A mapping between two models rarely maps all the concepts in one model to all concepts in the other. Instead, mappings typically lose some information and can be partial or incomplete" (Madhavan et al. 2002; Choi et al. 2006). In order to find an accurate ontology mapping, accurate similarity measurements between source ontology entities and target ontology entities should be considered.

In summary, ontology is an emerging discipline to solve the semantic heterogeneity to data. It has potential to improve information organization, management and understanding.

- Ontology can describe the semantic of data from sources created from different perspective and provide a unify view to users. It can be considered as ontology based data integration.
- Task ontology might be known as an approach for the case that users need data to carry out a certain task. Task ontologies describe the user task in relation to data. Applying task ontology. Users can discover and retrieve all data needed for their task.
- Since, ontology is a tool to represent knowledge; it also has differentiations because of perspective of producers. Ontology mapping is also a trend of research nowadays.

- The reusability is one of key features of ontology. It helps ontology designer can use existing knowledge from current systems instead of create a new one from scratch.

2.4. Conclusion

Ontology is applied for this thesis because it can represent a certain consensus about the knowledge with a richer internal structure as it includes relationships and constraints between the concepts (Palmer 2001). Ontology had been applied for information systems in many distinct aspects in order to improve the search results. Ontology-based approaches are the best solution to solve the semantic heterogeneity of data. The proposed approach in this thesis applies ontology to describe the relationship between dataset, observed objects, phenomenon and user task to facilitate user search for data in a water-related information system.

The idea presenting the observed objects in relating with phenomenon and dataset was adopted from Marius Podwyszynski (Podwyszynski 2009). This thesis adds the concepts describing the influence of phenomenon to observed objects to provide relevant data for a certain request. The idea of Athanasis Nikolaos (Athanasis et al. 2009) on describing data into different schemas is also considered. Instead of assigning data to each schema, this thesis will present an approach that describes the relationship between schemas. Data are just assigned to a class that present a real world object which can be observed by that data. This thesis also incorporates the concept presented in Nancy Wiegand about user tasks (Wiegand et al. 2007). But instead of describing task and phenomenon as one compound concepts in which data are assigned to, in this thesis, user tasks are described independently from phenomenon, that enables the combination between tasks and phenomenon. Concepts and term applied in this thesis are extracted from existing ontologies and thesaurus such as SWEET, GeoNames and AGROVOC.

3. METHOD

This chapter presents the research approach to accomplish the thesis objectives that can be summarized as to provide a list of relevant datasets for user requests. To reach the goal, this study proposes a semantic layer describing the semantics of data, observed objects, phenomena and user tasks (see section 1.1.2 for the explanations of these terms) built on top of a current geospatial information system. Domains and relations are described in more details as following.

3.1. Overview of approach

The geographic features are the core of geographic information that depends on the perception of the data provider and the needs of a specific application which determine the contents of the geographic information (OGC 2013a; OGC 2013b). A geodata is a set of geographic features which are objects of the real world associated with a certain location on the earth surface. Depending on the particular application, geographic features of point(s), line(s) and area(s) are modeled and represented in type of point, polyline and polygon. The spatial characteristics of geographic feature are quantitatively described by the geometric objects that are a combination of a coordinate geometry and a coordinate reference system. for accelerating computational geometry, topology is constructed in order to model how geographic features share coincident geometry, such as adjacent features – two provinces – share one edge. However, this study does not focus on these two aspects of geodata and assumes that the geodata is stored in an existing relational database.

In most of the current systems, geospatial data are stored in a Relational Database (RDB) (Laclavík 2006) and arranged into three aspects of thematic, spatial and temporal aspect. Datasets have a certain topic regarding to a particular region and have a period of time in which they are valid time. For example, census data has been assigned a thematic reference value that indicates the theme of data, which is statistic value about population. Census

data also has a temporal value which indicates the period of collected data; the spatial reference value presents a location, for which data was collected, hereafter called cover area. According to OGC, temporal and spatial aspects are integral parts of geographic information system. Traditionally, temporal characteristics of features have been treated as thematic feature attributes (OGC 2013b), e.g. a feature "Building" may have an attribute "date of construction". Nowadays, there are standards define the temporal and spatial schema, such as ISO 19107:2003 geographic information — spatial schema and ISO 19108:2002 geographic information — temporal schema. These are defined independent of thematic aspect. In this approach, the collected data is described by different domains coincide with three aspects of data, i.e. thematic, spatial and temporal aspects.

As shown in Figure 3.1, a semantic layer is proposed to build on top of the data layer of the existing system. The semantic layer has two sub layers – the mapping layer and the ontology layer. The attributes of datasets, stored in the RDB, are assigned to RDF applying D2RQ mapping language. In this way, these attributes comply with the constraints, rules and definitions predefined in the semantic layer to express the meaning of datasets. The semantic layer consists of separate domains describing the concepts and relationships between them (Figure 3.2). As mentioned before in section 2.2.2, concepts are represented by classes in ontology files. Instant values of classes are individuals of concepts, i.e. datacollection class is a concept, and individual datasets are instant values of that class.

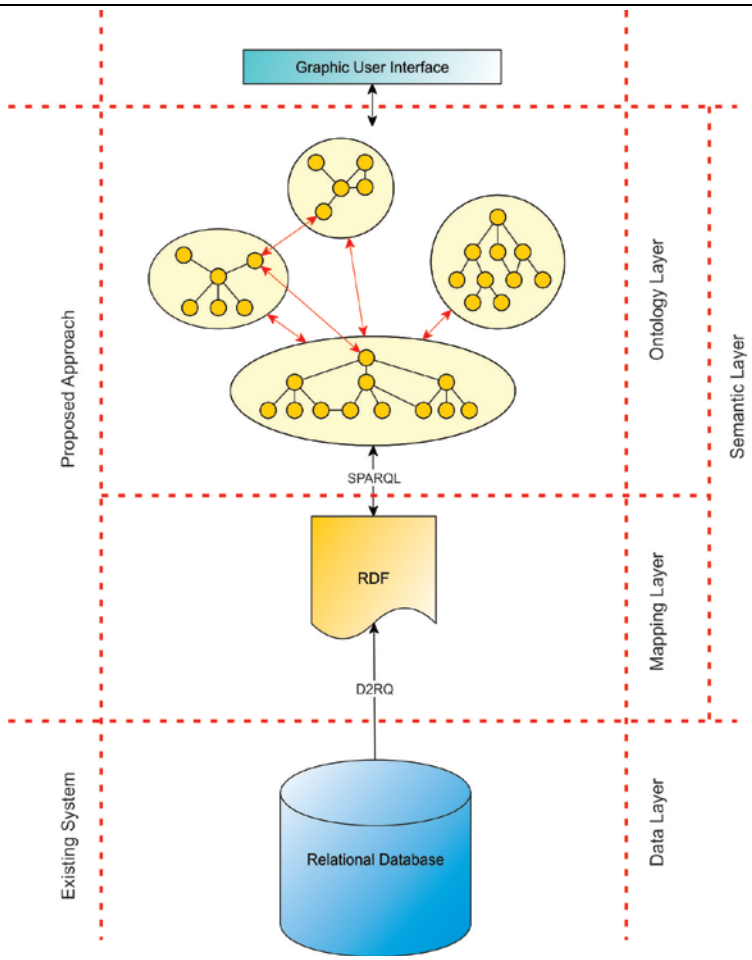


Figure 3.1: The Overview of thesis's approach

The overview of the relationships between domains is shown in Figure 3.2 demonstrating main classes of domains and main relationships between domains, i.e. Data domain, Observed object domain, Application domain, Temporal domain and Spatial domain.

Data Domain contains classes presenting the properties of datasets, e.g. format type; geometric resolution – pixel size; spatial representation – line, point, polygon or pixel; and spatial relation - which area the datasets relate to; and thematic reference classes of datasets. Datasets in the RDB are assigned as individuals to datacollection class and have relation to corresponding thematic classes and another class (see more details in section 3.2).

Observed Object Domain consists of classes that describe physical and non-physical objects related to the water subject, i.e. "man-made feature", "natural" and "social" which

are called observed objects (refer to section 1.1.2.1 - Observed objects). Phenomena are also presented concerning observed objects. The relationships in this domain are described independently from tasks. Therefore, the defined concepts in this domain are easy to combine with any tasks defined in application domain (see more details in section 3.3).

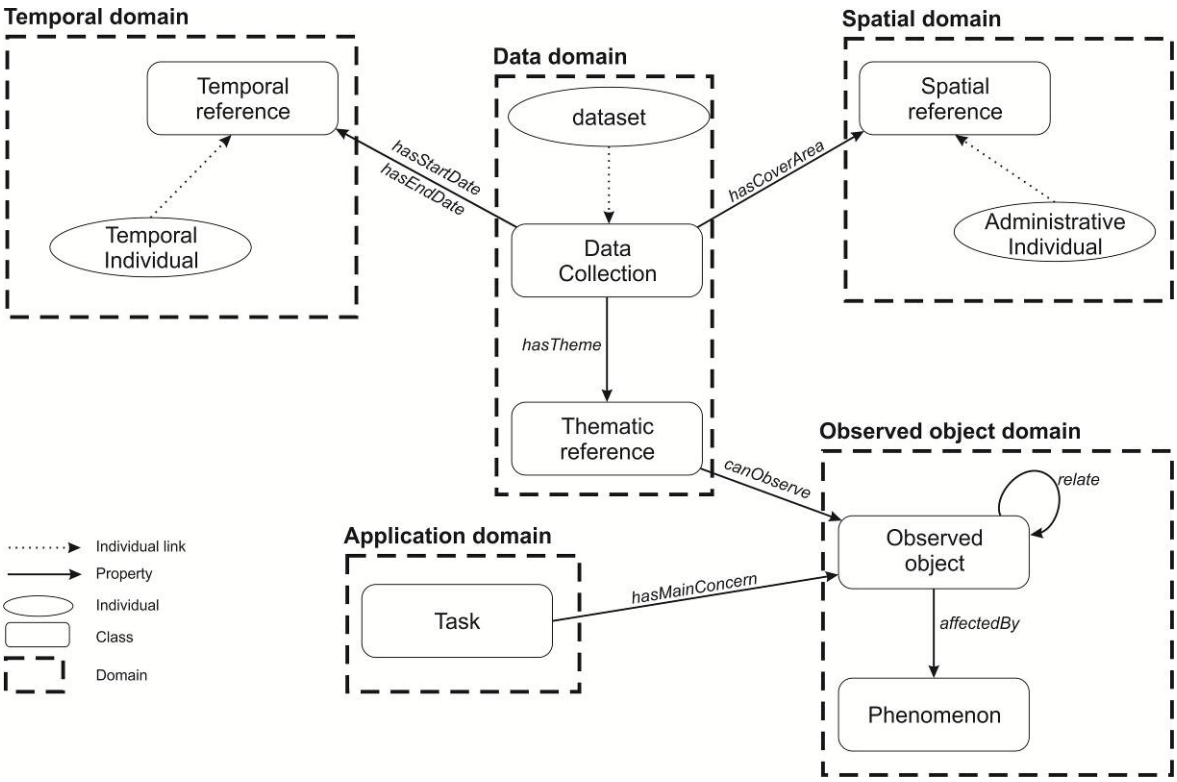


Figure 3.2: Main classes and main relationships of domains

Application Domain describes the user's tasks divided into types, e.g. response task, monitoring task, etc. The user tasks are described in relation to observed objects which are the main concerns of these tasks. The task acts like constraints to limit the returned result regarding to a certain phenomenon (see more details in section 3.4).

Temporal Domain has the main class Temporal Entity which stores the valid date values of datasets. The individuals in this domain are the start date and end date showing the period of datasets in which they are valid. In the case of datasets valid in just a point of time, e.g. water level dataset that contains information about water level at collected time; the start date and end date are the same value (see more details in section 3.5).

Spatial Domain stores values of the administrative areas as references for cover area of a dataset. The locations, where datasets are collected, are assigned to administrative areas containing those locations. For instance, no matter if these locations are surveying points (the stations collecting water level) or land use maps for a whole district, both are assigned to district level areas, which are called cover area (see more details in section 3.5).

In the semantic layer, the domains relate to each other by properties, Figure 3.2 depicts the main classes of domains and their relationships. The data domain connects to observed object domain by the properties "canObserve" which links thematic classes to observed object classes. User tasks in the application domain associate observed objects by "isMainConcern" property. Temporal and spatial domains contain values presenting attributes of datasets that linked the two domains to the data domain with "hasStartDate", "hasEndDate" and "hasCoverArea". These relationships can respond to typical searches described as follows.

- With a certain user query for an observed object (step (1) in Figure 3.3), a list of thematic classes related to observed objects is collected through semantic descriptions in the observed object domain (step (2)). And then, datasets that are assigned to thematic reference schema are retrieved (step (3)). Normally, users define when and in which area they are interested in (step (4), (5)). The list of returned datasets can be limited by temporal and spatial parameters of the user query (step (6), (7)). Finally, the list of datasets matching the user query is shown. The result returned to the user consists of all relevant datasets that relate to an observed object the user is interested in. That has been done based on the properties that describe the relationships between observed objects, i.e. "canObserve", "relate" in the observed object domain.

- With a certain user query for a phenomenon (Flood for example), the first step is to retrieve a list of observed objects, which are "affectedBy" or "canObserve" regarding to defined phenomenon (step (1) in Figure 3.4), and then the system works in a similar way to the previous case presented above.

- This case can be combined with a particular task (for example, phenomena: flood + task: rescue). With a particular task for a certain phenomenon, a list of data will be

changed by using the task as a constraint (step (11) in Figure 3.5). For example, with flood phenomenon, the result should be a list of data describing the objects which can observe water level (as they "can observe" flood) and paddy fields or residential areas (as those are "affected by" flood) and so on. But in combination with rescue task, the list should omit data for paddy fields but add the data about health care system and population density (as these are the "main concern" of rescue task)

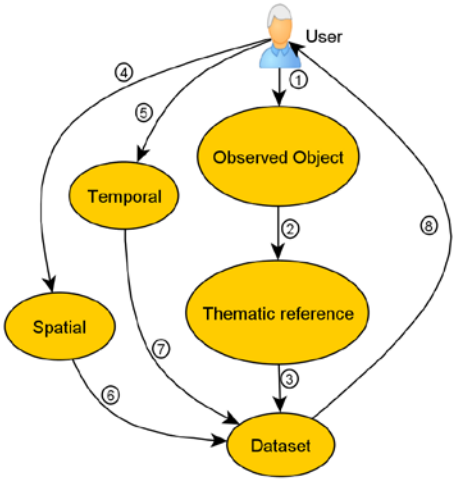


Figure 3.3: User search for observed object

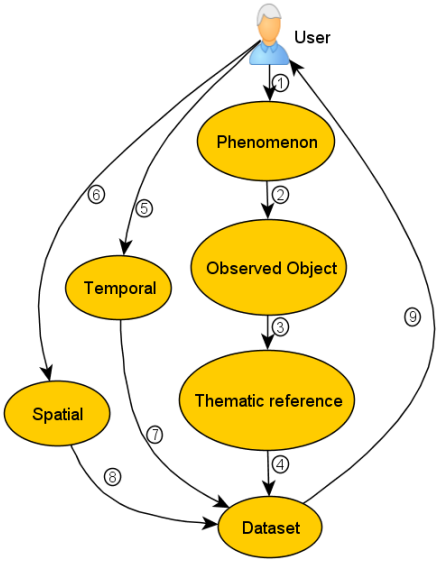


Figure 3.4: User search for phenomenon

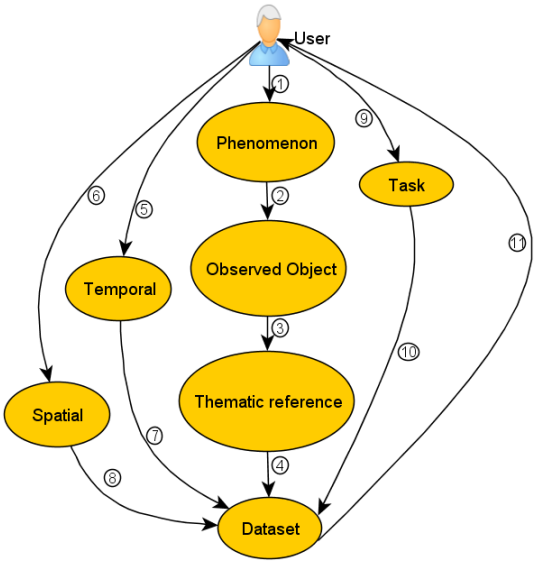


Figure 3.5: User search for phenomenon with a particular task

3.2. Data Domain

This domain describes attributes of datasets in a hierarchical structure with relationships between classes as shown in Figure.3.6. The hierarchy consists of classes presenting attributes and characteristics of datasets. All datasets are grouped by format type and assigned as individual values to subclasses under datacollection class. Besides classes containing the physical attributes, e.g. format type, spatial representation for vector data, geometric resolution for raster data etc. the data domain consists of data provider information in order to determine the origin of data. The datasets link to thematic reference classes which, in turn, are linked to observed objects in the observed object domain. Thematic reference concepts are adopted from the current system.

Temporal values presenting the valid dates of datasets are assigned as “start_date” and “end_date” stored in the temporal domain by properties “hasStartDate”, “hasEndDate”. Spatial reference values which present the related region assigned as cover area by datasets are presented in the spatial domain. The temporal and spatial domains are described in section 3.5 - Spatial and Temporal Domain. As the domains are designed independently, they apply the pre-existing definitions of ontologies and can be extended for future use.

The data domain includes definitions which allow the automatic classification of datasets into classes. For example, the high resolution imageries class is defined as a class holding raster based datasets that have pixel size smaller than 10 meters. Datasets which satisfy these variables belong to "HighResolution" class automatically. It also has constraints to ensure the consistency of the model, such as each dataset have only one format type or vector data have to have scale information.

There are properties presenting attributes of datasets such as "is_a" relationship – a relationship between a class and its sub classes; "hasProvider" – information about data provider; "hasResolution" – resolution of raster imageries; "hasCoverArea" – cover area of the datasets; "hasStartDate" and "hasEndDate" – start and end date of valid period of time of dataset. It has also properties presenting semantics of datasets, e.g. "bestFit" which holds knowledge related to the level of geometric resolution in relation to the area of interest the user states in a query. For example, high resolution imageries are suitable for small areas such as districts because they can detect the real-world objects in more details than the medium or the low resolution, while the large areas such as region covering many provinces should use low resolution or medium resolution imageries for overview purposes.

Furthermore, constraints are defined to ensure the consistency of the model or to check the integrity of database by running a reasoner, e.g. Datasets have only one format type such as raster, vector, tabular or text based. Raster based data have only one pixel size etc. and ensure the correct mapping between RDB and RDF. Reasoners detect missing values during the mapping process easily based on defined constraints in the ontology file.

Since, the observed objects in observed object domain have interrelations themselves (see section 3.3), the links between thematic reference and observed object classes enable to infer of the relations between datasets, even if these relations are not recorded in the database. For example, the datasets about the water level information can observe the status of the river under observation. A water mask dataset is a satellite imagery processing product in which all the water bodies are masked into a single layer presenting the distribution of surface water. So that it relates to water level. Therefore, it is inferred that water masks also relate to the river status.

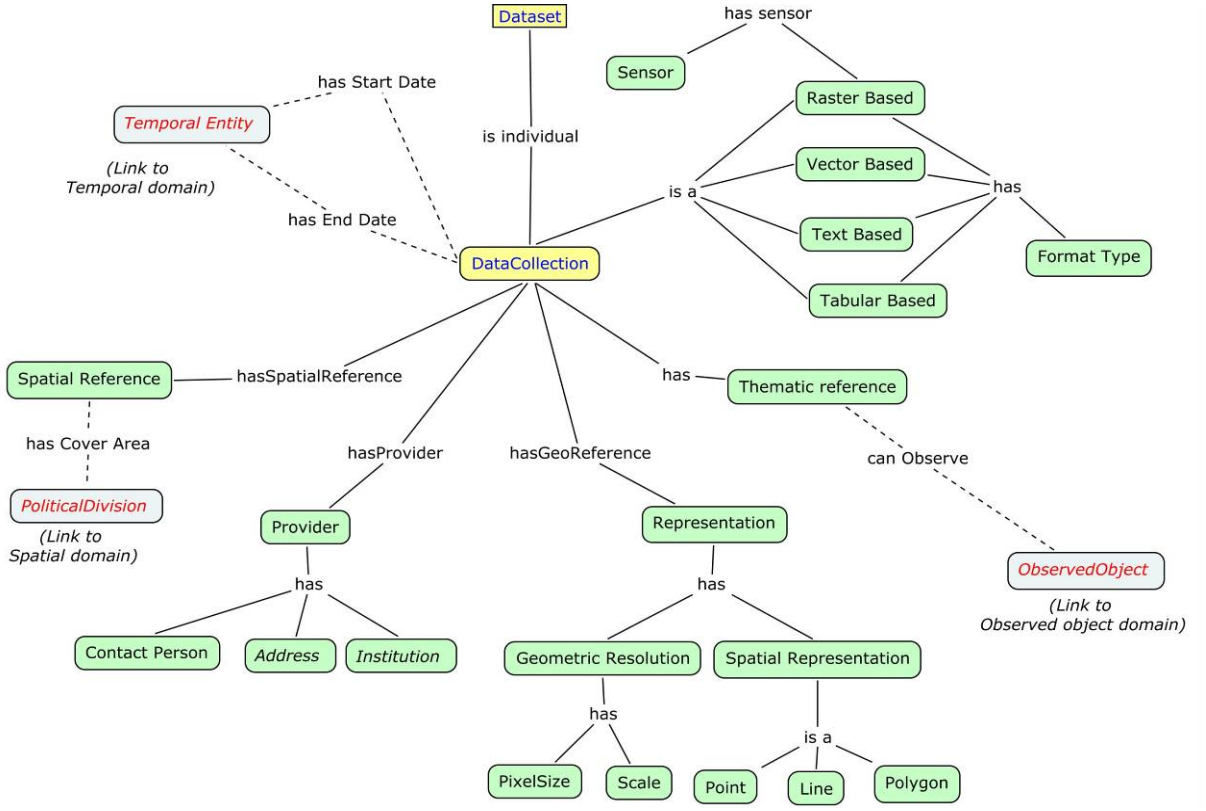


Figure.3.6: Outline of Data domain

In summary, this domain contains concepts presenting information about datasets. These concepts are arranged into classes in a hierarchical structure. The relationships of classes are represented by properties which are also used to construct definitions and constraints presenting the semantic of datasets. Datasets are assigned to this domain as individuals of the datacollection class; dataset attributes are individuals of corresponding classes. For example, a dataset with the name “land use map 2010 of province A” is provided by provider B that is depicted as below (Figure 3.7).

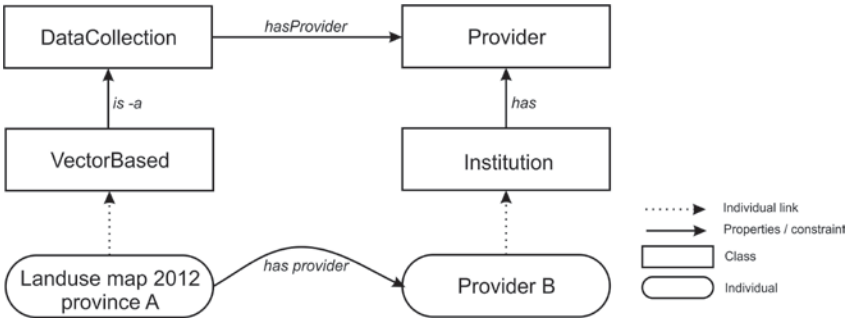


Figure 3.7: Example for relationships of classes and individuals

Despite containing attributes which are used to describe the semantics of a dataset, this domain does not describe everything about data. It is only an additional part of metadata.

3.3. Observed Object Domain

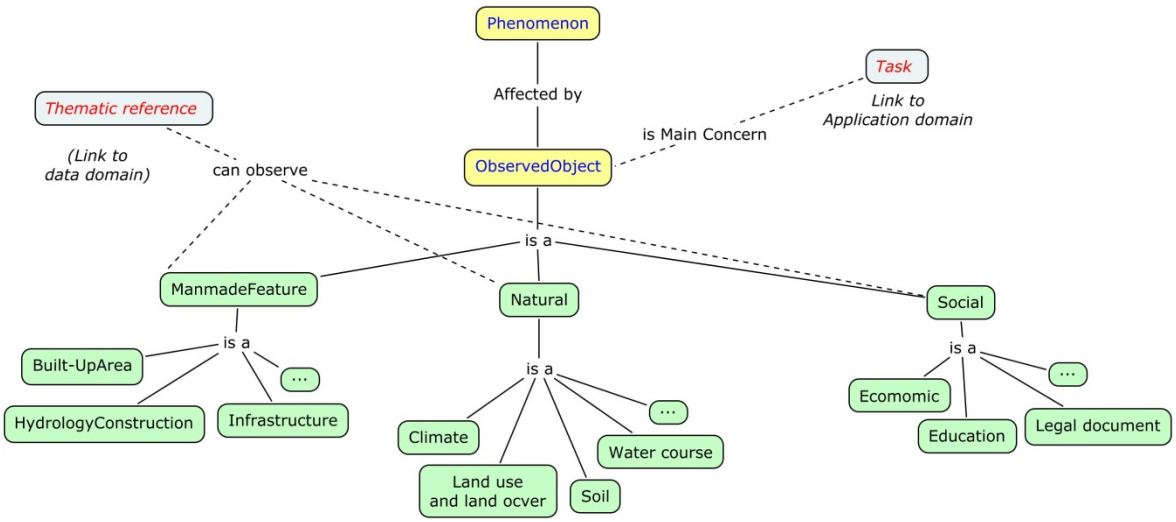


Figure.3.8: Outline of Observed object domain

This domain presents two main super classes that are ObservedObject and Phenomenon (Figure.3.8). A list of observed objects is grouped into three main classes – “man-made feature”, “natural” and “social” (refer to section 1.1.2.1 for definitions of these classes). Phenomena are described in terms of how they affect observed objects. This does not include the cause of phenomena.

Observed objects are presented in a hierarchical structure including properties that define how these objects relate to each other, e.g. the water mask has a relation to water level and the agriculture area relates to agriculture productions. This hierarchy adopts concepts coming from the section agriculture from AGROVOC, dealing with the level of relations which are defined as broader-, narrower-, related term in AGROVOC (see section 2.3.3 - Existing Ontologies). The higher level presents more general concepts than the lower. For example, “natural” class has the sub class “land use and land cover” which contains

"agriculture area", "industrial area" etc. Thus, "natural" is more general concept than "land use and land cover". The more detailed concepts are "agriculture area", "industrial area".

It is important to define the relationships between classes. That determines the results returned to user queries. In general, we can say that everything in the world relates to each other (Tobler 1970) – which counts for user queries as well, the system retrieves every dataset because they are related to each other. On the other hand, some relevant data will not be retrieved if relationships are not or are inappropriate defined, which are described as the precision and the recall issues (see chapter 6).

The main properties are used as follows.

- CanObserve: defines which thematic reference groups can observe which observed objects.
- Relate: presents the relationship between observed objects themselves.
- AffectedBy: describes the relations on how phenomena affect observed objects. It consists of sub properties such as "increase", "decrease" or "destroy".

These properties are used to infer which datasets relate to which phenomena and how datasets relate to each other in terms of their relations with observed objects. Figure 3.9 shows how the reasoners can infer datasets. With a certain phenomenon (case 1 in Figure 3.9), observed objects are retrieved via "affectedBy" property, then thematic reference via "canObserve" property. And finally, datasets which belong to the same thematic reference classes are retrieved. The datasets which relate to each other are found out due to the "relate" property in a similar way by a reasoner (case 2 in Figure 3.9).

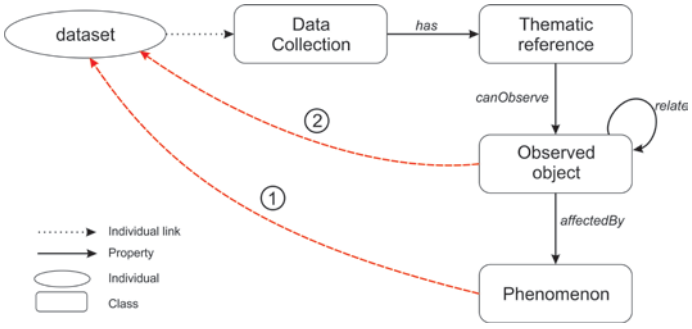


Figure 3.9: Abstract model for inference of datasets

3.4. Application Domain

This domain describes user tasks, which relate to the definitions of how much data are needed to carry out the tasks. However, this does not include the way how the tasks are performed. The user tasks are divided into sub tasks, which are defined independently from phenomena.

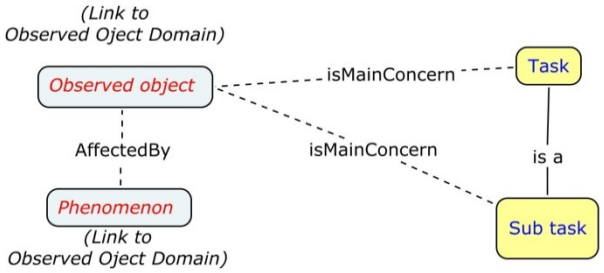


Figure.3.10: Abstract of Task Domain

The most important step in the application domain is to define sub tasks, and then assign the property “hasMainConcern” to observed object classes residing in the observed object domain. This property indicates particular observed objects which will be retrieved for a certain task.

The way of defining tasks is crucial for this domain. For this purpose, existing definitions are applied to define which observed objects are main concerns of tasks or sub tasks. For example, rescue task has sub tasks such as search and find, first aid and transportation network. Search and find sub tasks needs population data and the first aid sub task needs healthcare system information etc. Another example is monitoring task defined as the “measurement of environmental characteristics over an extended period of time to determine status or trends in some aspect of environmental quality” (Suter 1993). So it can be interpreted as follow: the main concerned observed objects of monitoring task are classes having “canObserve” property. In this case, the definition for “monitoring” class is constructed in Protégé as “hasMainConcern some (ObservedObject and (canObserve some Phenomenon))”. Since, the user tasks are defined independently from phenomena, they can

be combined for several purposes and that enables the scalability and transferability of the approach described here to another field of interest.

3.5. Spatial and Temporal Domain

The temporal domain applies the ontology defined by Feng Pan 2005(Pan et al. 2005) which provides the basic temporal concepts and relations that most applications would need, i.e. vocabularies for expressing facts about topological relations among instants, intervals, and events, together with information about durations, and about dates and times. Although, the approach of this thesis uses only "instant" class to describe the valid date of data, but the temporal ontology can be used later for automatic data processing.

The main class of this domain is the "instant" class (Figure.3.11). It has individual values describing the values of valid date of datasets. The properties "hasStartDate" and "hasEndDate" are used to define the links between datasets and individual values of the temporal domain (start date and end date).

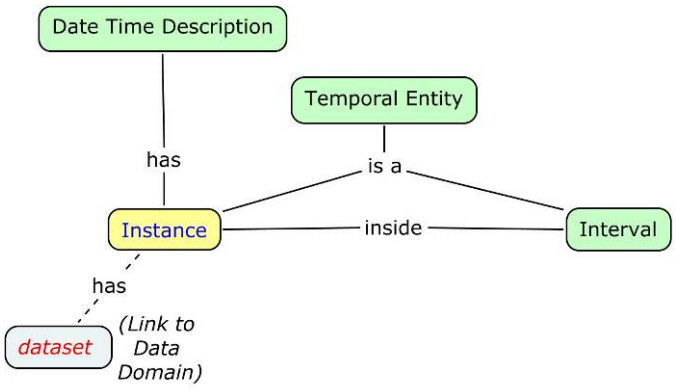


Figure.3.11: Outline of Temporal Domain

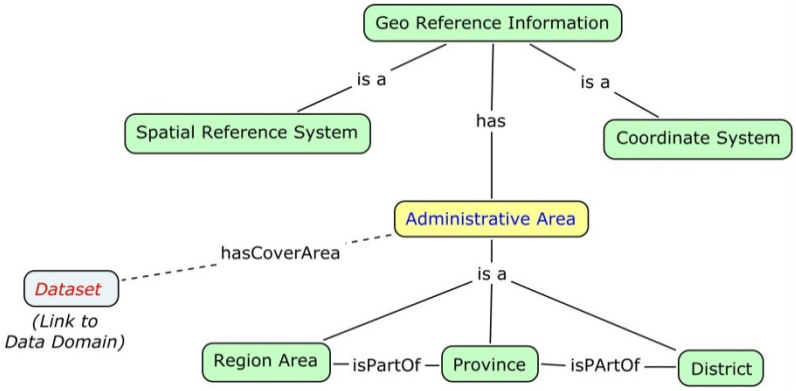


Figure.3.12: Abstract of Spatial Domain

The spatial domain is adopted from SWEET ontologies (see section 2.3.3), which suggests seven spatial ontologies including spatial coordinate, directions, distribution, extent, objects, scale and general spatial concepts. However, only Administrative Area class and its' sub classes are applied for spatial domain in this research (Figure.3.12). The geo reference information class is kept for future used in case of geodata from different geo reference system is integrated. As mentioned above in the section 3.1, the spatial characteristics of geographic feature are described by the geometric object which is a combination of a coordinate geometry and a coordinate reference system. In other word, the spatial domain should have information about geo reference system in order to define way to overlay various data from different geo reference system. However, in this study, geodata is assumed to be preprocessed into unique projection.

Administrative areas are classified into levels such as country, regions (including many provinces), provinces (including many districts) and districts. The lower level areas are linked to the higher level areas by property "isPartOf". These areas linked to datasets in the data domain by the "hasCoverArea" property. Similar to observed objects hierarchical structure, the datasets that relate to the higher administrative levels contain more general information than the datasets relate to lower levels. With queries for a certain area (e.g. province level), datasets which cover higher (region level) and lower levels (district level) of administrative areas are also retrieved, because they are related. This domain focuses only on the level of administrative areas. The spatial relationships such as nearby, far, left or right are not included.

3.6. Relational database (RDB) to Resource description framework (RDF)

Since data are stored in RDB, the values from RDB are mapped to instants of RDF by applied D2RQ tool. Values from RDB are assigned to RDF class through a mapping file. Essential values, which match to the defined concepts in ontology files, are mapped to RDF. That reduces the complexity and redundancy of the model.

The vital attributes which have to be mapped are

- Id: this value is used to distinguish datasets.
- Thematic reference values: these values determine how the datasets are connected to observed object and phenomena in observed object domain.
- Spatial reference values: determine the cover area of datasets.
- Temporal values: determine the valid period of datasets.

The other attributes can be information about format type, scale, resolution or provider to trim down the results returned.

The mapping file uses N3 based syntax – N3 is an assertion and logic language which is a superset of RDF (W3C 2011). It can be created with any text editor program. The mapping file defines how to connect to RDB, and the rules to map values from RDB to classes of RDF (D2RQ 2012). D2RQ acts as a mediate tool to connect RDB and RDF.

The structure of mapping file consists of three main parts (D2RQ 2012):

- Prefix: to shorten the mapping file
- Database connection variable, and
- Mapping rules: define which values from RDB are mapped to which classes in RDF.

D2RQ provides sufficient ways to map values from RDB to RDF such as (D2RQ 2012)

- Property bridge using information from another table: this property joins tables together by defined relationship between tables in order to get values mapped to RDF.

- Foreign key with multiple columns: The links between tables are defined by multi values from columns.
- Joining a table to itself: A table can be joined to itself in order to get two sets of information from one table. For example, in the case of administrative table, all areas are stored in the same table, they relate to each other by column parentID. It is necessary to link the table to itself to retrieve these values.

3.7. Ranking

Since, the proposed approach provides a list of relevant data for a user demand, there should be many datasets returned for a certain user query. Users will face the situation that they cannot distinguish which datasets provided by system are the most appropriate ones for their demands. That makes it necessary to rank dataset depending on their semantics, e.g. dataset cover areas, temporal entities, the level of relations, i.e. “canObserve” or “relate” (section 3.3), and “bestFit” property (section 3.2).

Regarding to the spatial match of datasets, three levels can be determined as follows (Figure 3.13).

- Exact match: datasets have cover areas which exactly match to the user request
- Lower level: datasets have cover areas which have a lower level in the administrative system comparing to users demands.
- Higher level: datasets have cover areas which have a higher level in the administrative system comparing to users demands.

The exact match is assigned as highest rank, then lower level and the higher level has lowest rank, because higher levels contain more general information than the lower levels.

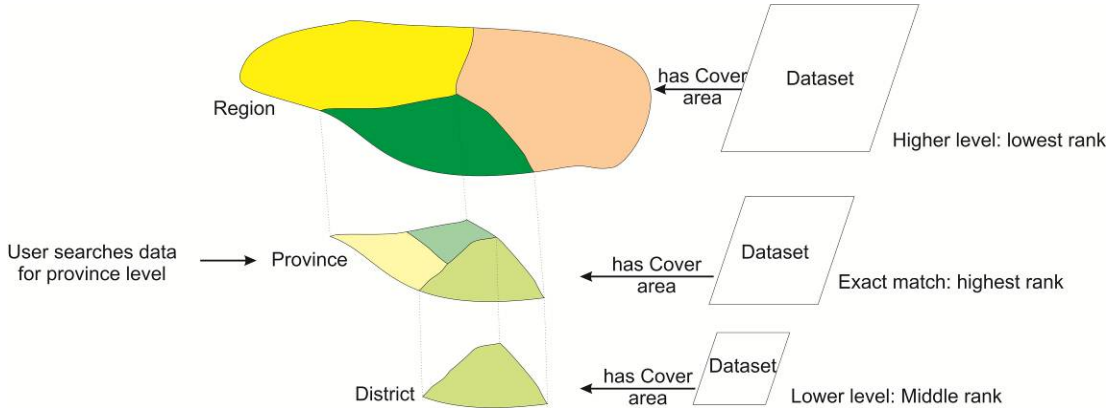


Figure 3.13: An example for the ranking of cover area property

Another possibility is to assess the suitability of datasets for query areas. As mentioned before, this thesis proposes a property so-called “bestFit”. An example for “bestFit” property is shown in Figure.3.14. The high resolution imageries can detect objects in more details than the medium or low resolution. These imageries are sufficient most appropriate for small areas such as districts. When users want to observe a large area on the other hand, they just need an overview. Thus, the medium or low resolutions are most appropriate (“best fit”) for overview purposes, i.e. province or region levels. This property applies to vector based data in the same way, since they have several scales with different levels of details. Dataset, which have “bestFit” property are assigned a high rank. This rank applies just for spatial datasets, i.e. raster and vector based data.

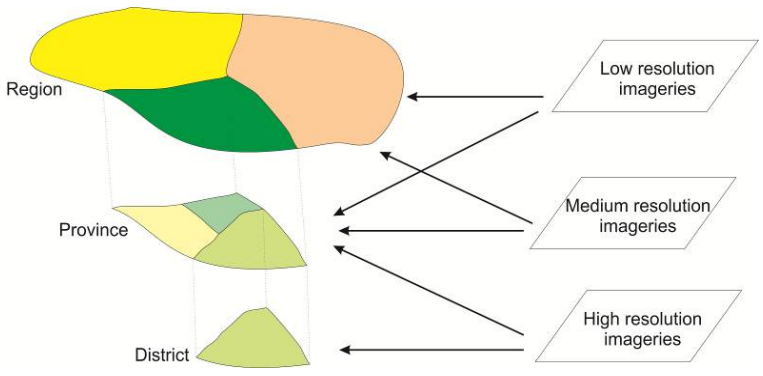


Figure.3.14: An example for the “bestFit” property

The temporal aspect is also a criterion to rank the results. There are three cases which are thinkable to indicate the relations between the requested period of time and the valid date of data (start date and end date) as shown in Figure.3.15. The Q_{min} and Q_{max} represent the lower and upper bounds of the query time range. The T_{min} and T_{max} represents the start and end date of the valid period of time of datasets.

- Inside: This case applies for datasets which have the valid dates fall into the bound of the query time range. The datasets which are valid in a point of time are included in this case. This case can be presented by the condition as shown below.
    - $Q_{min} < T_{min} \leq T_{max} < Q_{max}$
- Overlap: This case is applied when the bound of the query time equal (exact match) or fall into the valid dates of datasets. This case is shown by the condition below.
    - $T_{min} \leq Q_{min} \leq Q_{max} \leq T_{max}$
- Intersect: In this case, the bound of the valid date of data intersect with the bound of the query time range. It can be shown as follow.
    - $T_{min} < Q_{min} \leq T_{max} < Q_{max}$ or
    - $Q_{min} < T_{min} \leq Q_{max} < T_{max}$

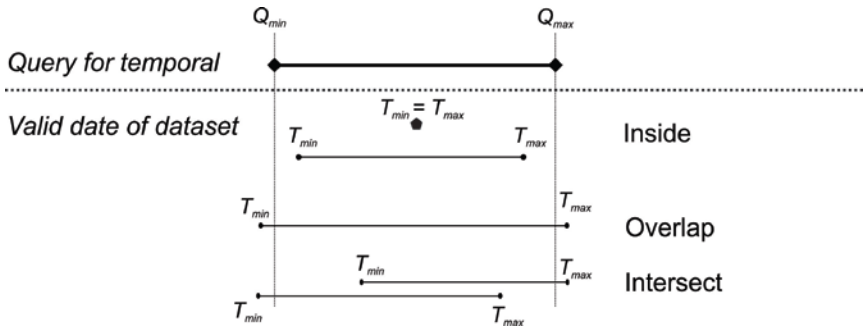


Figure.3.15: User query related to temporal property of dataset

The datasets are ranked based on the comparison of the bound of the query time range and the valid period of datasets. For calculation purposes, all the values of time are converted to monotonically increasing real number such as “Unix Time”, where time are track as a

running total of seconds [(2)]. The datasets are ranked based on how large the bound of valid date (the bound of dataset) cover the bound of query time range (the bound of query). The more the bound of dataset cover the bound of query, the higher these datasets are ranked. In the overlap case, the bound of dataset covers whole the bound of query, this is the highest rank. For every datasets which are valid in a point of time inside the bound of query are also ranked with highest point. For the remaining cases, the datasets are ranked by the function as shown below.

$$R = \begin{cases} \dfrac{(T_{max} - T_{min})}{(Q_{max} - Q_{min})} & Q_{min} < T_{min} < T_{max} < Q_{max} \quad (1) \\ \dfrac{(T_{max} - Q_{min})}{(Q_{max} - Q_{min})} & T_{min} < Q_{min} \leq T_{max} < Q_{max} \quad (2) \\ \dfrac{(Q_{max} - T_{min})}{(Q_{max} - Q_{min})} & Q_{min} < T_{min} \leq Q_{max} < T_{max} \quad (3) \\ 1 & Q_{min} < T_{min} = T_{max} < Q_{max} \quad (4) \\ 1 & T_{min} = Q_{min} \leq Q_{max} = T_{max} \quad (5) \end{cases}$$

Where: R is score which is used to rank the dataset.

By applying these functions, the higher score datasets have higher rank than the others.

Another method used for ranking applied in this thesis is the relationships between the observed object classes. These relationships are determined by properties which define the relative levels of classes (concepts) based on the hierarchy of classes extracted from existing dictionary or thesaurus. There are two types defined in this thesis as follows.

- Direct property: “canObserve” property. This property defines that a thematic reference class has close relation to an observed object. It is also applied for the relationships of observed objects and phenomena.
- Indirect property: “relate” property or “is-a” relationship. Indirect properties present loose relationships such as land cover classes relate to land use class.

[(2)] http://unixtime.info/facts.html

The indirect properties are defined to have lower ranking than direct ones. The property rank applies for queries about phenomena.

3.8. Conclusion

The approach described above is applied to describe the semantics of datasets in relation to thematic classes, observed objects and phenomena. Within this proposed method, the dataset are described based on their characteristics in data domain. They are linked to observed objects, phenomena in the observed object domain via the properties. The properties and relationship between domains enable users to discover and retrieve all relevant data for their searches. Furthermore, the task ontology in combination with the descriptions of phenomena facilitates users search. So, they can discover all available data for their tasks just in one search.

Actually, the returned list consists of many datasets. The ranking method provides a way to order datasets based on the appropriateness of data concerning the user demands. The total rank is a sum of three ranking aspects mentioned above, i.e. cover area, temporal and properties aspects. The total rank is calculated as shown below:

$$\text{Total Rank} = \text{Point}_{\text{MatchingLevel}} + \text{Point}_{\text{Temporal}} + \text{Point}_{\text{Property}}$$

Where:

- Total rank: the total point for dataset according to the appropriateness for user demand.
- $\text{Point}_{\text{MatchingLevel}}$, $\text{Point}_{\text{Temporal}}$, $\text{Point}_{\text{Property}}$: the point of data regarding to cover area, the valid period of times and the relationships of data comparing with user demand.

Therefore, there are three criterion used to rank data. Thus, the implementation allows users to choose which criteria is the most important for their queries. The equation changes to the ones shown below.

$$\text{Total Rank} = (W_{CoverArea}.Point_{MatchingLevel}) + (W_{Temporal}.Point_{Temporal}) + (W_{Property}.\ Point_{Property})$$

Where:

- $W_{CoverArea}$, $W_{Temporal}$, $W_{Property}$: the weight of which criteria is considered as the most important criteria.

This proposed approach commits to providing an innovative way to describe the semantic of data. It reduces the search time and provides more appropriate and accurate results for users. The bestFit property only assign to spatial datasets, so it is not used to rank for all datasets. It is shown in the result to provide a better way for the users to choose which datasets are appropriate to their need.

4. WISDOM INFORMATION SYSTEM CONTEXT

This chapter introduces briefly about the WISDOM project. The heterogeneity of collected data is addressed in the next section. The current data manage model is presented and the limitations of it are also pointed out. The limitations are resolved by applying the proposed approach.

4.1. Introduction

The Mekong River is one of the largest rivers in the world both in terms of its total length and mean of annual flow. Six countries – China, Myanmar, Thailand, Laos, Cambodia and Vietnam - have parts falling into its basin (MRC 2012). The Mekong Delta (MK), the last part of Mekong River, is located in the southern region of Vietnam covering an area of approximately 39,000 square kilometers. It is the largest agriculture and aquaculture production area of the nation and offers natural resources for several million inhabitants (Le 2010). The location of Mekong Delta is shown in Figure 4.1**.**

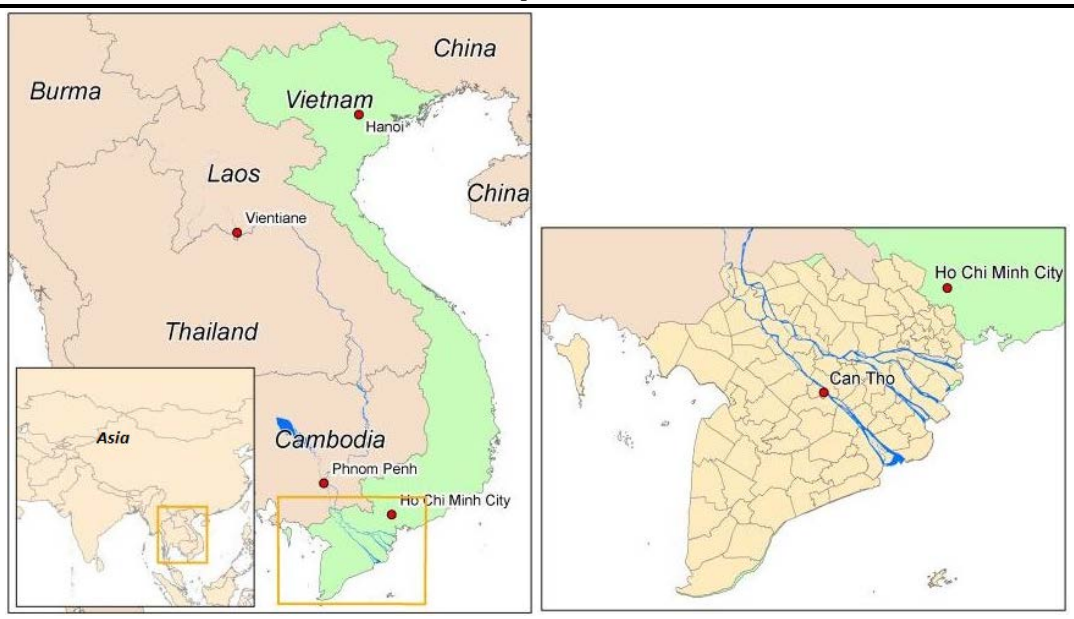


Figure 4.1: Location of Mekong Delta, Vietnam
(Source: (WISDOM 2011))

Pressure of population growth, changing climatic conditions and regulatory measures at the upper reaches of the Mekong lead to critical changes in the Delta. “Extreme flood events occur more frequently, drinking water availability is increasingly limited, soils show signs of salinization or acidification, species and complete habitats diminish” (WISDOM 2011). All these issues require an optimized, integrated resources management. For this purpose, detailed knowledge on hydrological, hydraulic, ecological and sociological factors must be available for interaction between institutions and organizations. Understanding these aforementioned issues, a bi-lateral project between Vietnam and German government has been established focusing on development and implementation of an innovative water-relate information system containing all the outcomes and results of the different research disciplines involved in the project (WISDOM 2011). Users of the WISDOM IS comprise researchers and decision-makers, with an individual background resulting in different knowledge and demands. A researcher understands how datasets are grouped into thematic groups (for example, water level dataset should be somewhere under hydrology group) and knows what is most relevant for the research. On the contrary, the decision-makers may be unfamiliar with geographic information systems and may also be a technically unskilled person. They cannot guess which one from the results returned from their query is the most

relevant. As a result it is necessary to have an appropriate approach to provide highly accurate result for users.

4.2. Collected Data in WISDOM / Data model in WISDOM

The acronym WISDOM stands for Water-related Information System for the Sustainable Development of the Mekong Delta, a bi-lateral project between Vietnamese and German government. This is a multi-disciplinary project associated with the principles of an Integrated Water Resources Management (IWRM) which is defined as the following *"Integrated water resources management is a process, which promotes the coordinated development and management of water, land and related resources in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems"* (Global Water Partnership 2000). It is very useful to have an information system which can integrate and share data or information from different subjects related to water resource management. The water-related information system is not only a very useful tool for IWRM, but it is also an integral part besides improving policies for water related fields (Hristov et al. 2006).

4.2.1. Fields of research

Following the concept of IWRM, the multidisciplinary project approach of WISDOM makes it necessary to have a system collecting and managing all relevant data. The information system developed in the context of the WISDOM project was designed and implemented applying internet infrastructure and related technologies for the Mekong Delta. It contains information from the fields of hydrology, sociology and earth observation (WISDOM 2011). The integrated data domains are shown in Figure 4.2.

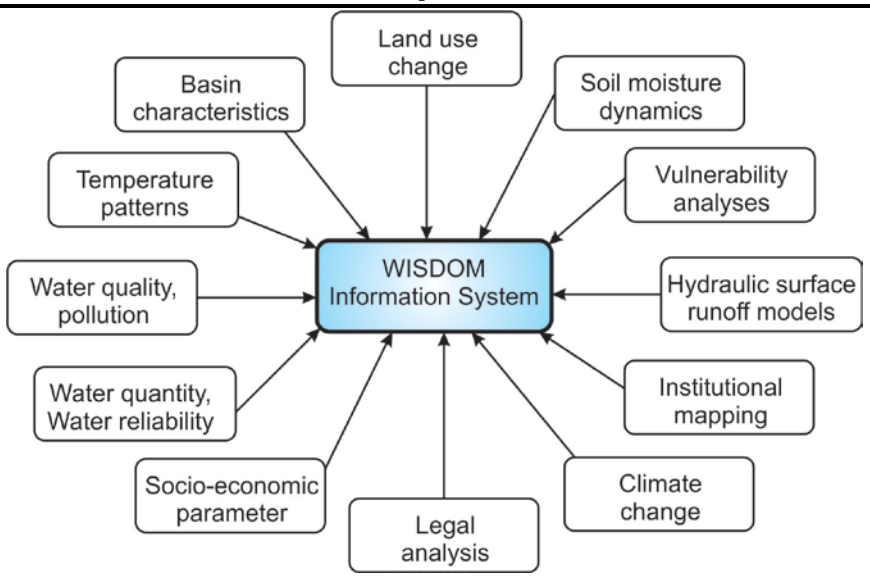


Figure 4.2: WISDOM research fields
(Source: (Klinger et al. 2010))

The system contains a huge amount of data in different formats from several research fields, i.e. land use change, soil moisture dynamics, vulnerability analyses, Hydraulic surface runoff models, institutional mapping, climate change, legal analysis, socio-economic parameter, water quantity, water reliability, water quality and pollution, temperature patterns and basin characteristics. The integration of such data enables the (end-) users of the system to perform analyses on specific questions transcending scientific disciplines; and thus supplies the (end-) users with a tool supporting regional management and planning activities by visualizing and disseminating data. The main idea behind the WISDOM Information System (WISDOM IS) is to provide an autonomous data management and data query system. This system uses an operational data flow to minimize requirements to end users' information technology skills and user driven errors.

In the context of the WISDOM project, data from various scientific disciplines are generated, which go along with the fact that scientists from different backgrounds have different requirements and concepts about facts to be collected in the real world. For example, a hydrologist refers to water level, water discharge or period of flooding, and a scientist from the social-economic domain considers flood in a way of the severeness affecting human life by calculating for example the hectares of agricultural production areas affected. Furthermore, data from different aspects vary not only in meaning, but also

in formats. Collected data from research fields use several formats to describe the same feature in different perspectives. The landscape of collected data can be depicted as follows.

- Vector data such as country boundaries, major cities, river networks on a national, provincial and district level, road networks residential areas, administrative boundaries with different temporal validities, etc.
- Raw remote sensing data from several sensors with different resolutions. Products from remote sensing data such as land cover classification, water turbidity, inundation mapping, precipitation data, water masks, soil moisture and others. These are stored in raster format.
- Temporal data from in-situ sensor measurements such as buoy and other sensors measuring water levels, water flow, salinity and temperatures and water quality indicators such as pesticide concentrations and endocrine disruptors in waterways. These are included field data or ground truth data
- Hydrological and hydraulic modeling results such as water levels, inundation areas. These data are stored in raster format.
- Statistical data on several topics for different years within the administrative levels (e.g. national, provincial, or even household level) stored in tabular format.
- Information of organizations in the water sector, especially in the Mekong Delta is stored including their issued documents.
- Literature and reports, which related to the Mekong Delta and water resource management in general, are also incorporated into the system.

The goal is to ensure users use the same data source and high quality data for their work. Therefore, it is demanding to standardize and map all these data into one thematic reference schema and to set up and describe relationships between these data in order to maximize the efficiency of such a system.

4.2.2. Data management model

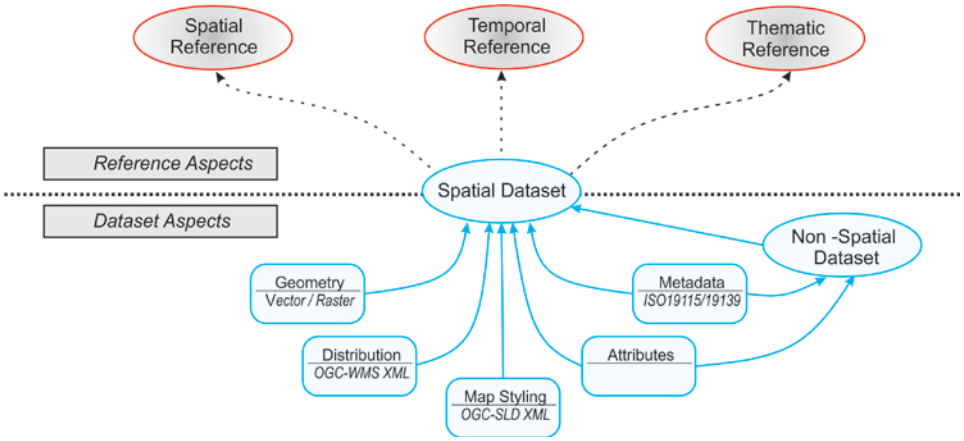


Figure 4.3: Aspects of data within WISDOM IS

This thesis interests only in the semantics of datasets, thus this section just describes the data management model covering semantic. It does not include the technique descriptions of the WISDOM IS. Since, WISDOM is a multidisciplinary project, the requirement for the information system is to develop methods to store data in an efficient way to be able to persistently manage all necessary aspects of data and provide the user with an easy way to retrieve all relevant data (Klinger et al. 2010). To manage and generate information from data provided by the various data producers, WISDOM IS applied a data management model which organizes dataset attributes into aspects. These aspects are divided into two groups to describe datasets in details and to define how data is stored. These groups are data aspects and reference aspects (Figure 4.3). Data aspects are divided into five as follows (Gebhardt et al. 2010a). Dataset are distinguished by UUID (a Universally Unique Identifier) and are described in several aspects depending on their original format.

- Geometric aspect: present the geographic extent of whether the data are raster or vector data. The geographic extents of data are stored as simple polygon geometry relating to corresponding dataset by their IDs (UUID).
- Data transfer aspect: describes parameters necessary to establish the transfer of datasets using OGC standard web mapping services (WMS),

- Data styling aspect: define the graphical representation of datasets following the OGC Styled Layer Descriptor specification which extends the WMS standard to symbolize and color geographic feature.
- Metadata aspect: Metadata contains information about datasets in detail including identification, the extent, the quality, the spatial and temporal reference scheme, and distribution of digital geographic data in an XML Extensible Markup Language file. This aspect contains the mandatory and the most important optional fields applied ISO 19115 and 19139. These fields present information such as author information, dataset point of contact, dataset identification and abstract, keywords for theme, region, discipline and temporal validity. All the datasets are described in this aspect.
- Data attributes describe interested or captured values of real-world objects storing in datasets. Every datasets have data attributes stored in several formats.

Data aspects act as metadata. They advanced data query and data distribution algorithms, as spatial datasets can be searched by ISO19115 and ISO 19139 metadata using OGC Geoservices (OGC 2012a) such as Geonetwork catalogue system (see more details in (WISDOM 2012)). A spatial dataset can be retrieved as a WMS layer via common web or desktop clients such as OpenLayers, Gaia or ESRI ArcMap.

Moreover, the reference aspects are designed to allow users to explore data using efficient search options by geographic, temporal and thematic search variables.

- Spatial reference aspect: this aspect presents a hierarchical structure of administrative areas, in which the observations or social data are collected, according to administrative level (in the case of WISDOM, administrative levels in Vietnam are country, region, province, district and commune). Every dataset is registered to at least one of these objects.
- Temporal references aspect: this aspect contains the instants value of time describing the valid period of datasets. It is necessary to describe data for a fast and sufficient access for both user defined and automatic data queries.

- Thematic references aspect: a list of themes is organized as hierarchical groupings, which increases the speed of retrieval and accessibility of hydrologic, environmental, or social data (Figure 4.5).

Spatial reference aspect: To provide a way to search for data in a quick and simple manner, datasets are tagged not only with plain coordinates (central point or extent) but also with administrative names, which are called spatial reference objects. The spatial reference schema is a hierarchy of administrative areas represented by their name, Figure 4.4 shows an example. The highest level in this case is the nation level which is subdivided into regional levels, which consist of province levels in the next lower level. The next level is district level, and the lowest is communal areas (Gebhardt et al. 2010a). This aspect represents the parent-child relationship of administrative object levels using n:1 relation storing in RDB. Table 4.1 shows an example from spatial reference table, datasets have code "w1c05cn42r08p12" and "w1c05cn42r08p13" assigned to "w1c05cn42r08" as a parent level, and then to "w1c05cn42" as a grand parents. Here 'w' stands for world, 'c' for continent, 'cn' for country, 'r' for region, 'p' for province, 'd' for district. Thus, the code carries the information to which level the administrative unit belongs to.

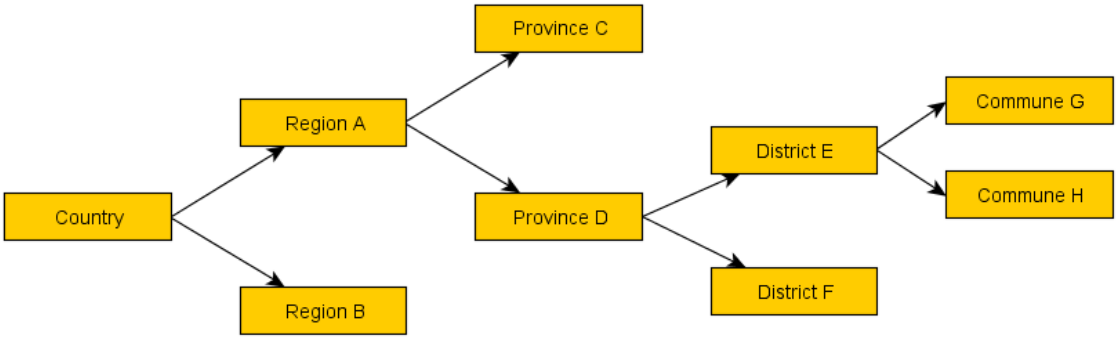


Figure 4.4: An example for spatial reference schema

code text	parentcode text	name text	level integer	date_from date	the_geom geometry
w1c05cn42	w1c05	Việt Nam	3	0001-01-01 BC	
w1c05cn42r08	w1c05cn42	Mekong Delta	4	0001-01-01 BC	
w1c05cn42r08p12	w1c05cn42r08	Cần Thơ	5	2004-01-01	0106000020E6
w1c05cn42r08p13	w1c05cn42r08	Hậu Giang	5	2004-01-01	0106000020E6
w1c05cn42r02p03	w1c05cn42r02	Lai Châu	5	2004-01-01	
w1c05cn42r06p04	w1c05cn42r06	Đắc Nông	5	2004-01-01	
w1c05cn42r06p05	w1c05cn42r06	Đắc Lắc	5	2004-01-01	
w1c05cn42r02p04	w1c05cn42r02	Điện Biên	5	2004-01-01	
w1c05cn42r08p07d04	w1c05cn42r08p07	Bình Tân	6	2004-01-01	0103000020E6

Table 4.1: Administrative units in Vietnam are stored in the spatial reference table

Dataset id	Dataset uuid	Reference entity id	Reference name
403	b33c503f-4038-44b7-9232-1c99065041de	843	H. Châu Phú
403	b33c503f-4038-44b7-9232-1c99065041de	333	Mekong Delta
403	b33c503f-4038-44b7-9232-1c99065041de	632	Bac Lieu
403	b33c503f-4038-44b7-9232-1c99065041de	631	An Giang
404	4ff0e48b-164a-4696-af16-21d8f6082e79	333	Mekong Delta
404	4ff0e48b-164a-4696-af16-21d8f6082e79	631	An Giang

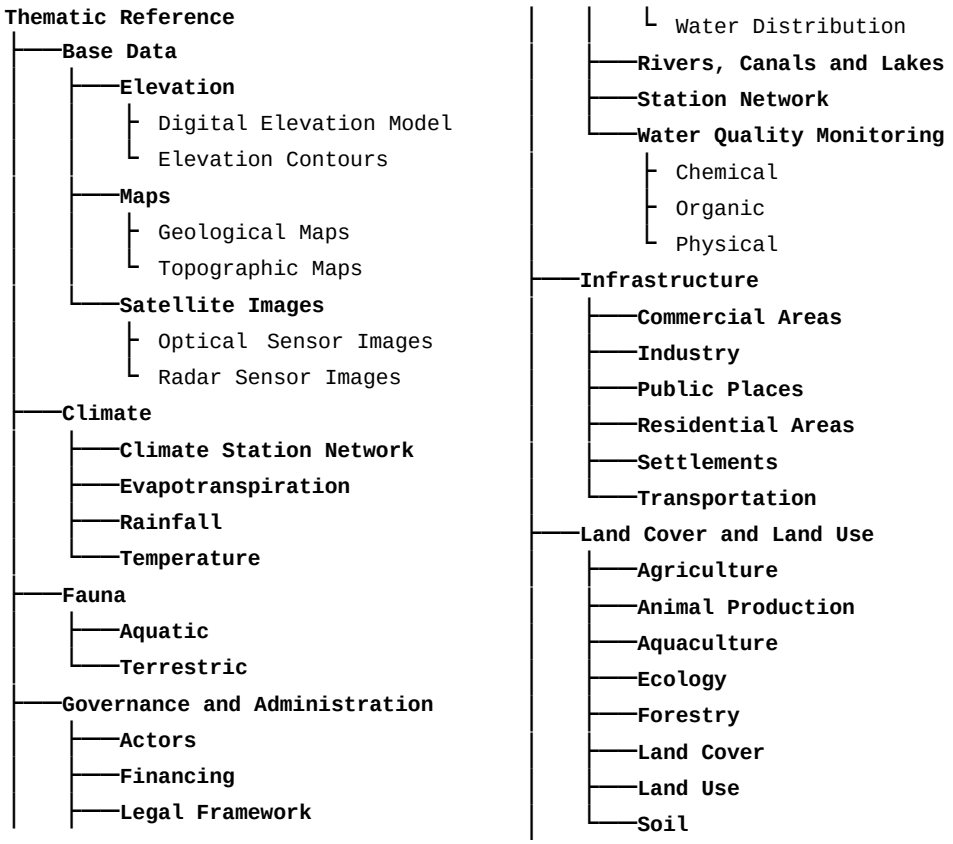
Table 4.2: Example of the spatial reference model in the WISDOM IS

By assigning a dataset to the spatial reference schema, the user can search data by giving a meaningful administrative name. For spatial datasets, the extents are intersected with the footprint of the administrative objects. Therefore, every dataset is registered to multiple objects of the various categories (e.g., administrative unit, catchment area). These relationships can be pre-calculated using GIS functions. As shown in Table 4.2, the dataset with id 403 refers to different administrative levels such as district, province and region via the ID and name of administrative objects. This enables a meaningful data search through the name of administrative areas. In the case of point observation data, the locations of collected data are assigned to administrative areas which contain these points. With non-spatial datasets, the spatial reference objects have to be defined manually either point or polygon objects, such as measurements to sensor point locations and census data to

administrative boundary polygons as shown in Figure 4.3. This aspect enables to search for every datasets by the administrative object names.

Temporal reference aspect: A dataset is defined by the date created and the time range over which it is valid. Time range is defined by start date and end date. It can be a period of time or just an instant of time. For example, the population census data is considered valid during the whole year, even for the whole period of time between two investigations (three or five years), while the satellite imageries and the observed data from buoy such as water level, water discharge are valid only for the time of acquisition, respectively of the observation for that specific moment only, therefore, the start date equals the end date. Another case, a geology or soil map may have infinite end date. The new observations are assumed that they will update the old.

Thematic reference aspect: To let user efficiently search and retrieve datasets, they are assigned to corresponding classes of a thematic reference, whether they are articles as pdf, a vector dataset, or statistics (Klinger et al. 2010).



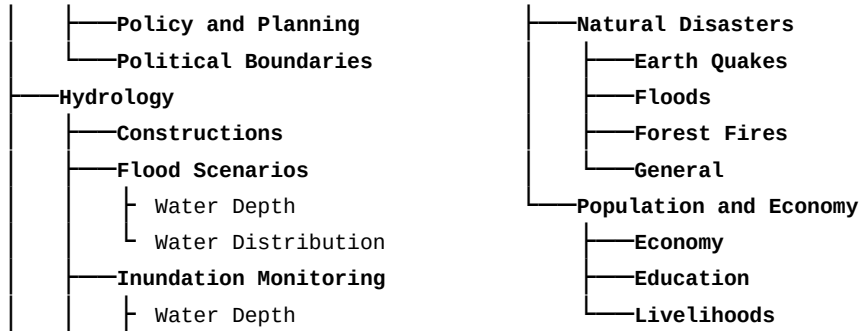


Figure 4.5: Thematic reference used to search data sets in WISDOM
(Source: (Klinger et al. 2010))

The thematic reference schema is managed in a hierarchical order as shown in Figure 4.5. Implicitly, datasets which are mapped to one level in thematic hierarchy are also belongs to all the parent levels along the branch. Thus, datasets are assigned to its respective themes at different levels to speed up the data access process. To be able to double register one dataset, a n:m relationship between thematic reference and product group was set up (Klinger et al. 2010). Product groups are groups of datasets which are created by the same processing method and/or describe the same object in the real world. They are described as sub classes of thematic reference groups and are very close to the meaning to the term "dataset", e.g. "watermask from optical sensors", "water chemical substance", "rainfall" and "soil moisture", etc. The product group list reduces the complexity of the system. Since, product groups have close meaning with data, so that it is also easy for data providers to define which product group the dataset belongs to.

As shown in Figure 4.6, firstly, datasets are tagged by product group using n:1 relationship, then each product group is mapped to a thematic reference group with a n:m relationship (Klinger et al. 2010). Table 4.3 shows an example on how a data assigned to thematic reference classes. A "watermask" belongs to "Environment" at the highest level, "Hydrology" in the next level and, finally to ‘‘Water level" at the lowest level. Also, spatial datasets are related to multiple themes within the same thematic level, e.g., the "River network" which belongs to "Environment", and "Infrastructure" at the highest level.

The thematic reference aspect adds thematic contextual information to a dataset using hierarchies. These relationships enable a meaningful search way through thematic groups.

Figure 4.6: The relationship between dataset and thematic reference via product group (Source:(Klinger et al. 2010))

Product Group ID	Product group name	Reference theme Id	Reference theme name	Reference theme level
16	River network	1	Environment	1
16	River network	2	Infrastructure	1
16	River network	8	Hydrology	1
16	River network	10	Transportation	2
17	Water mask	14	Environment	1
17	Water mask	28	Hydrology	2
17	Water mask	29	Water level	3

Table 4.3: Examples of "product-theme" entity relation model in WISDOM IS (Source: (Gebhardt et al. 2010a))

This study does not focus on the technique on how to visualize data on the internet browser; it focuses only on how to describe the semantics of data and how to retrieve relevant data for user search. Thus, data management model is the most interesting and important part of this thesis.

4.3. Conclusion

WISDOM data management model enables a contextual description of datasets and facilitates data query by defining meaningful search parameters such as the region administrative names, time ranges, and theme descriptions. To query for data, users have to define variables for thematic- (Figure 4.7), spatial- (Figure 4.8) and temporal reference (Figure 4.9). Then, the WISDOM IS returns a list of datasets in accordance with user demand, whether that is spatial data (vector or raster), reports or literatures. Using

WISDOM IS graphic user interface as shown in Figure 4.10, users can browse data on the map display or download for further process.

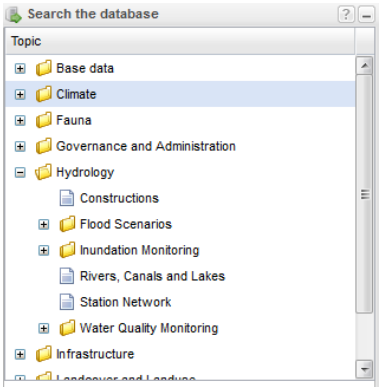


Figure 4.7: Thematic reference variable

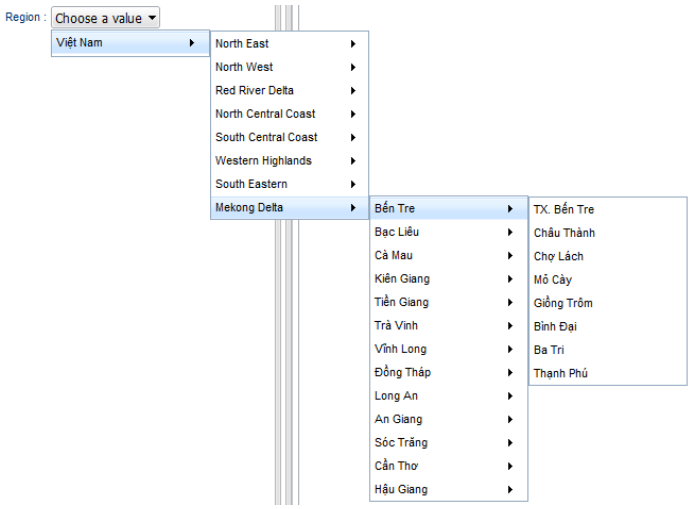


Figure 4.8: Spatial reference variable

Figure 4.9: Temporal reference variable

However, there are limitations of the model defined bellow:

- There are redundancies in the case of datasets related to different levels of the thematic reference schema. One thematic class connects to two or more higher level classes. One product group class links to two or more thematic classes (as

shown in Table 4.3). With this design, the extending of the thematic reference hierarchical structure may cause administration, maintenance and consistency problems (Gebhardt et al. 2010a), because the complexity of the RDB increases when the relationships increase. In fact, extending that structure is complicated.

- To retrieve all related data for user queries regarding to a certain region, datasets are assigned to different administrative levels. The dataset are assigned to areas which intersect with spatial extents of the datasets. At the same time, the parent, the grand parent and all the child level are also assigned to that dataset. Thus, with a certain query for a region at very low level such as district, the system retrieves datasets from very high level such as continental or country to district level. In contrary, every datasets within country are retrieved for a query with country level. Without ranking, the users find that it is very difficult for them to distinguish which datasets are appropriate for their need.

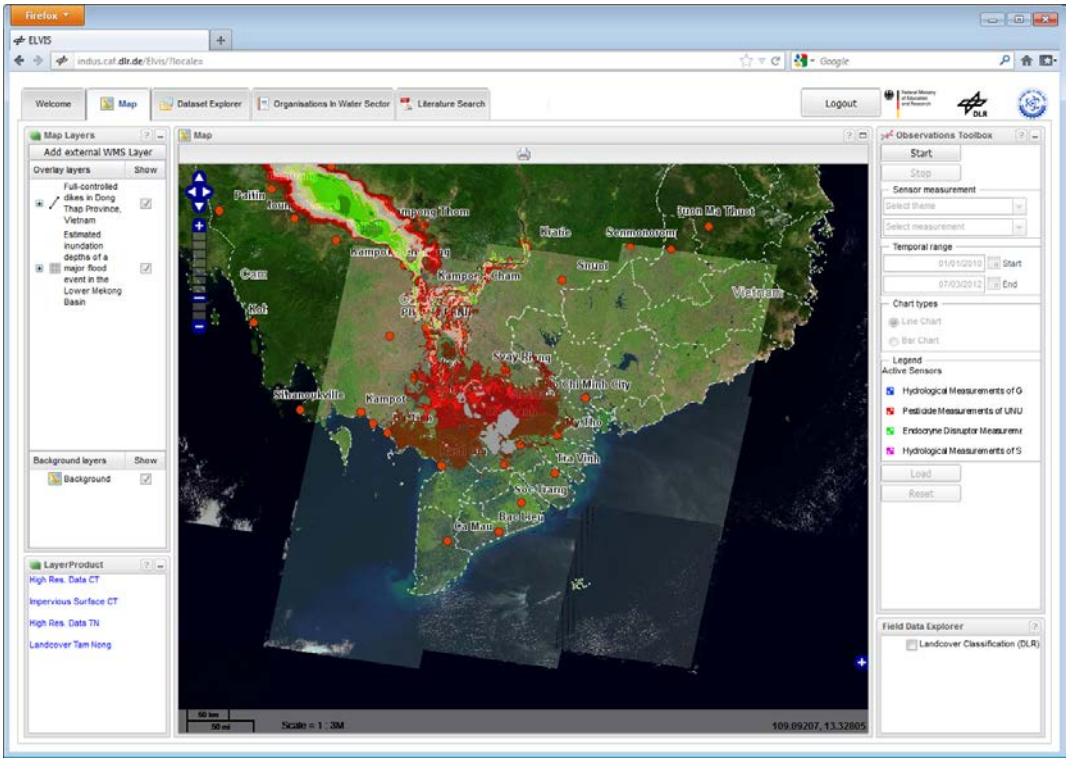


Figure 4.10: The GUI of WISDOM Information System

(Source: (WISDOM 2012))

- Furthermore, in the case of IWRM when users need relevant data to analyse the influence between real-world objects, or legal documents for water-related field, the current WISDOM IS cannot provide necessary data in one search. The users have to search several times and change thematic criteria by themselves. This is because the WISDOM IS with its cross-related structure cannot manage the relationships between thematic classes in a proper way. It cannot manage the relationships such as “relate”, “canObserve”, etc. For example, Land cover relates to land use, or the number of farms relates to agriculture production.

The insufficiency of cross-related data structure can be solved by applying an ontology based approach in order to add semantic descriptions to the database. Ontology can describe the semantics of data in a machine readable way, it provides an appropriate way to manage relationships between datasets and reference aspects. In addition, there are some software programs which can visualize the structure of ontology so that extensions and maintenance of the ontology gets easier.

5. IMPLEMENTATION OF PROTOTYPE

This chapter presents the steps to implement the prototype which apply the proposed method for the WISDOM Information System (IS). The flowchart of approach, the ontology domains in details, the applied tools and programing language are shown in details. The prototype proves the sufficiency of proposed approach. However, it is not a web-based application. It was only built to test the result returned from the proposed approach.

5.1. Proposed approach applied in the WISDOM Information System

A forward looking solution of providing all relevant data precisely for a specific query in WISDOM IS is to resolve the semantic heterogeneity of collected data. As mentioned before, data are described by several aspects which are arranged into two main groups, i.e. the data and reference aspects. However, as stated in the section conclusion of WISDOM Information System Context chapter (see section 4.3), the current data management model is insufficient to provide or retrieve relevant data for user search. Additionally, it is difficult to maintain and to extend the model. Therefore, one semantic layer is integrated into the existing system in order to provide a higher level data description in relation to the real world objects which are represented by datasets (Figure 5.1).

The semantic layer acts as an intermediate layer which describes the relationships between the datasets, the observed objects, the phenomena, and the user tasks. It consists of two sub layers, i.e. RDF layer and ontology layer. The ontology layer contains five domains, i.e. application-, observed object-, data-, spatial- and temporal domain. These domains connect to each other by properties which link concepts in domains or instants of concepts together. Applying the proposed approach, instead of assigning data to thematic schema using cross-relation tables in RDB, the values from the data aspects (see section 4.2.2) are mapped to RDF then comply with the constraints, rules and definitions predefined in semantic layer to

describe relationships between data and real world objects. That is the main key in this approach.

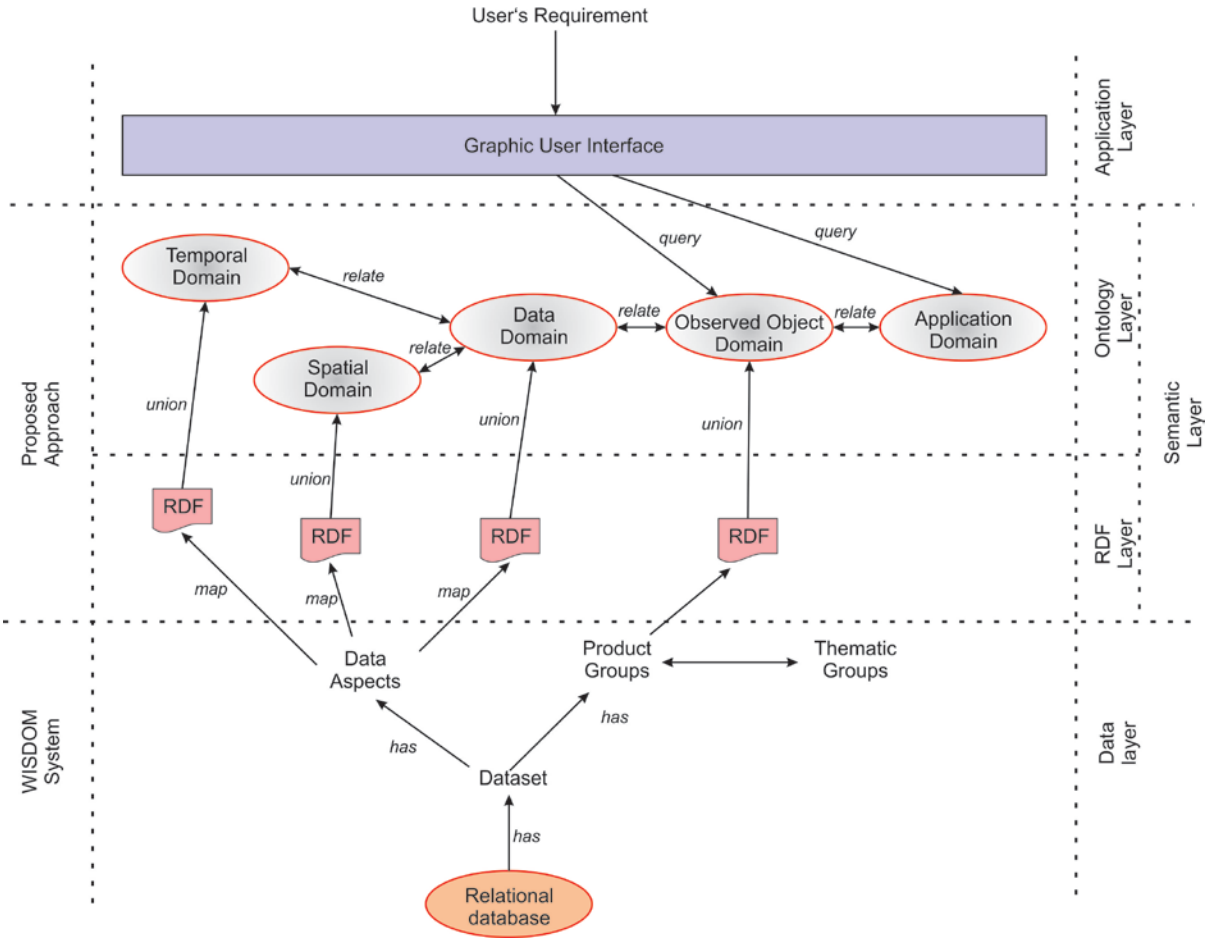


Figure 5.1: Integration of approach into existing system.

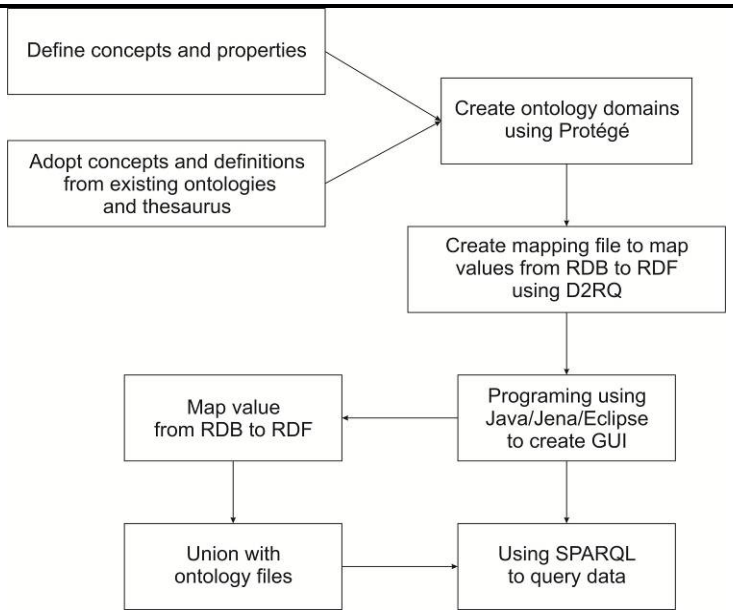


Figure 5.2: Flowchart of approach.

Figure 5.2 depicts the flowchart of the approach. The first step of flowchart is to define the concepts and the properties. This step includes two sub steps: after defining the concepts for the domains in the first step, the existing ontologies and thesauri are considered for reusing. Specifically, that is concepts adopted from SWEET ontology for the data aspects. However, the concepts of SWEET ontology are general; thus, adopted concepts are modified accordingly to WISDOM requirements (see section 5.2 for more details). Besides, existing thesaurus and definitions are also considered to construct other domains (see section 5.3 for more details). The existing concepts are rearranged in combination with new defined concepts to formulate domains.

The second step is to create domains using open source software Protégé. It is easy to construct, modify or maintain domains using Protégé, because it provides a user friendly graphical interface to create concept hierarchy and to define related properties of concepts. Rules and constraints are built easily using dialogues (Horridge 2011b). The ontologies can be checked with a reasoner and graphed with the visualization plugins. These functions help to ensure the consistency of domains.

For the next step, the mapping tool D2RQ is installed. The mapping file defines which values in RDB map to corresponding concepts of RDF. The mapping file uses N3 syntax - it is an

assertion and logic language which is a superset of RDF (W3C 2011). It can be created with any text editor program such as Notepad, WordPad or Notepad++.

Jena library, Eclipse software and Java programing language are used for the next step to build the prototype which has a GUI (Graphical User Interface) providing a user interactive environment for searching data. The mapping step assigns values from RDB to RDF, and then these RDF are integrated with pre-defined ontologies. Finally, user queries are transferred to the system by using SPARQL – the W3C standard query language for RDF (see section 2.2.3).

In the next sections, the steps of flowchart are presented in details. Domains are described with class hierarchy, constraints and rules. The way of mapping data from RDB to RDF using D2RQ is also presented. Jena, Eclipse, Pellet reasoner and query language SPARQL are mentioned as a tool to build a GUI for prototype.

5.2. Data domain

This domain consists of concepts which represent the characteristics of datasets, and additional information like data provider, contact person as well. The domain includes rules and constraints to ensure the consistency of the model. The concepts are defined based on WISDOM demands associated with the concepts adopted from existing ontologies and thesaurus.

The SWEET ontology has already defined concepts which describe the characteristics of datasets, which is called data ontology. The SWEET data ontology contains concepts presenting data model, data structure, data format, etc. However, these concepts are both redundant and lacking in comparison with WISDOM demands. For example, the SWEET data ontology has classes such as data structure classes present the way of storing and organizing data in a computer, or data representation classes present the way numbers are stored in a computer, i.e. 8, 16, 32 and 64 bit long (byte order class). On the other hand, WISDOM IS needs classes which describe the meaning of data, i.e. ProductGroup classes

(see section 4.2.2). Thus, the concepts of the data domain in this thesis are designed and arranged in a hierarchy as shown in Figure 5.3.

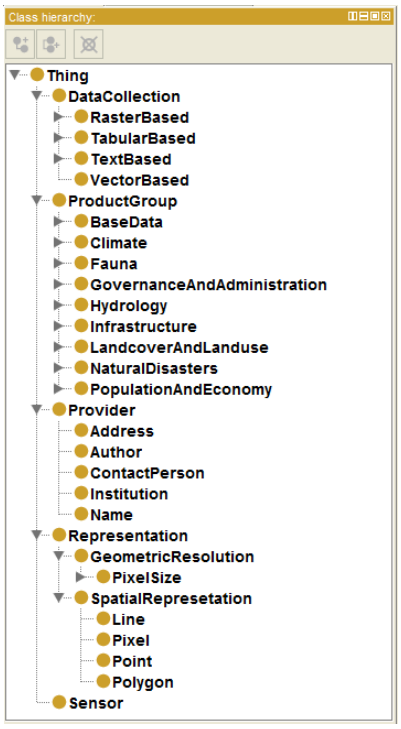


Figure 5.3: The abstract hierarchy of Data domain

In Figure 5.3, the concepts representing dataset attributes are organized into a hierarchy as follows.

- DataCollection: this class contains the list of datasets attributes, which are mapped from RDB. The datasets are indicated by ID and are grouped into sub classes regarding to their format type, such as rasterbased, vectorbased, tabularbased and textbased.
    - Rasterbased: data from flood scenario modeling, satellite imageries, scan map, field trip photographs are assigned to sub classes of rasterbased, i.e. SatelliteImage, ScanMap, Modeling and Photograph.
    - Vectorbased: contains data with vector format e.g. administrative boundary map, landuse map etc.
    - Tabularbased: collected data from observation stations and statistic works are divided into two sub classes of tabularbased, i.e. Observation and Statistic.

- Textbased: this class has two sub classes for literature and report data.

- ProductGroup (see section 4.2.2): consists of sub classes adopted from the WISDOM IS. They are organized at different levels (see in Appendix C) in a parent – child order. They all link to observed object classes in the observed object domain via the "canObserve" property.
- Provider: data provider's information is stored in classes for address, author, contact, institute and name.
- Representation: geometric resolution and spatial representation information of raster and vector data are mapped to sub classes of this class. Rasterbased data are grouped automatically into classes according to definition of classes. Figure 5.4 shows a definition of the high resolution class as a class holding raster based datasets that have pixel size larger than 2.5 meters and smaller than or equal 10 meters.
- Sensor: consists of individuals presenting sensors characteristics. These sensors have values which are used to distinguish each other (i.e. name, resolution).

The characteristics and relationship of concepts which are represented as classes in the domain are presented by properties. There are three property types (i.e. parent-child, data properties and object properties). Table 5.1 shows the other properties in details.

- Parent – child relationship is known as "is-a" property. It presents relations between classes and their subclasses. This property is expressed by relation of rasterbased and vectorbased sub classes with datacollection class as shown in Figure 5.3. With this relationship, the individuals of subclasses are inferred as individuals of all parent classes along the branch.
- Object properties present the relations of individuals values, in other words, they represent dataset attributes. For example, satellite imageries and data providers are individuals of rasterbased and provider class. They are assigned "hasProvider" property, which presents a sentence "satellite imagery A is provided by provider B" by structure "A – hasProvider - B". Object properties can also describe the relations between classes, such as the "canObserve" property presents the relations between water level and water mask like "water mask – canObserve – water level".4

- Data properties link a XML Schema Data Type value (e.g. integer, string, date time, etc.) to an individual of classes. That is, they describe relationships between an individual and data values. For example, dataset have name, this relation is shown with "hasName" property by structure "Dataset_hasName_X".

The figures in Appendix D show the relationships and the properties of the Data domain in more details.

Parent – child Property	
Is - a	This property presents the class hierarchical relationship. It enables to retrieve all relevant data in a same branch when users search for data in general concepts.
Object Properties	
Name	**Description**
bestFit	Indicates the suitability of raster or vector datasets for the administrative level area regarding to user query.
hasContactPerson	Links dataset to contact person.
hasCoverArea	Presents spatial region related to collected datasets.
hasSensor	Presents sensor's name of satellite imageries or products from satellite imageries.
hasProvider	Presents the organizations that provide data.
hasStartDate	Presents the start date of valid period of datasets.
hasEndDate	Presents the end date of valid period of datasets.
canObserve	Presents the relation between the product group classes and the observed object classes in observed object domain. This property links two domains together, i.e. data domain and observed object domain.
Data Properties	
Name	**Description**
hasAuthor	Presents the name of the author for literature or report datasets.
hasID	Presents the ID of the datasets which are used as name of individuals of classes.
hasResolution	Presents the resolution values of raster datasets.
hasScaleValue	Presents the scale of vector datasets.
hasUUID	Presents the UUID of datasets. These values are used to retrieve the datasets.

Table 5.1: Object properties and data properties in data domain

To ensure the integrity of the model during the mapping process, constraints are defined. The object properties can be used as constraints of concepts, such as data has only one format type; satellite imageries have only one resolution value. Properties can be combined together to build definitions as shown in Figure 5.4. The HighResolution class is defined as a raster based dataset, which are acquired by sensor with resolution of imageries larger than 2.5 meter and smaller than 10 meters. The definitions group datasets to classes automatically, they can help to discovery data in a fast and sufficient way.

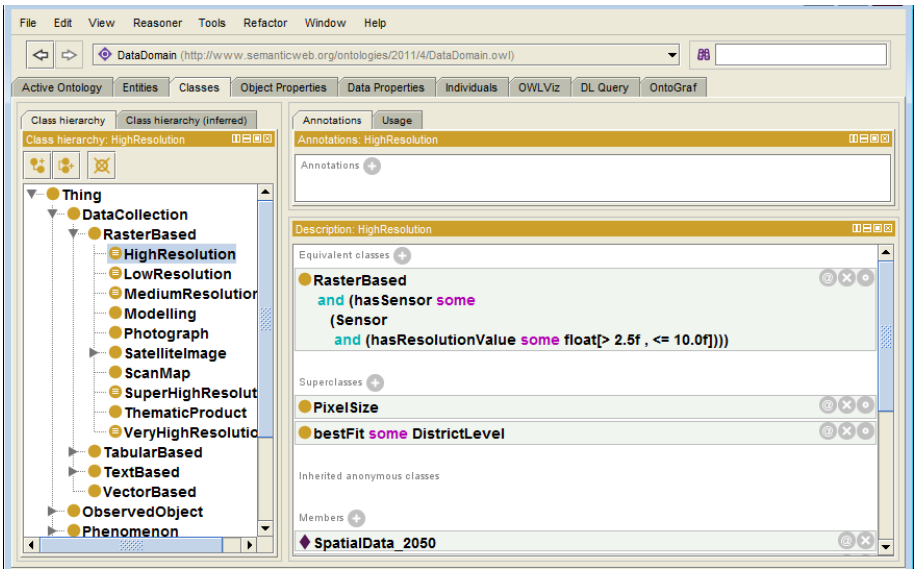


Figure 5.4: Properties are used as the definition for HighResolution class (Viewed by Protégé version 4.1.0 – Build 239).

Instead of assigning the datasets to administrative objects at different levels as in the WISDOM data management model (see section …), they are assigned to the largest administrative levels covered by them (e.g. province or district level) via the "hasCoverArea" property. A rule is defined as follows in order to refer all the relevant datasets.

$$\begin{cases} \forall AdministrativeObject(i,j), isPartOf(i,j) \\ \exists DataCollection(k), hasCoverArea(k,j) \end{cases} \rightarrow hasCoverArea(k,i)$$

This rule means, if there are two administrative objects (j) and (i); (i) is a part of (j) that means (i) is a lower level of (j) and the dataset (k) has relation with (j) as a cover area, then there is an inference that (i) are also covered by (k). Based on the hierarchy of administrative objects, the reasoner infers the higher and lower administrative level (if any), then retrieves all related datasets for a certain region for a query.

In summary, this data domain overcomes redundancy of cross-related structures of RDB as mentioned before in section 4.2.2. E.g. by using "is-a" and "hasCoverArea" properties, the reasoner can infer how datasets and administrative objects at difference levels relate to each other.

5.3. Observed object domain

The observed object domain consists of concepts representing the observed objects and the phenomena in relation to water. These concepts are organized in a hierarchy. This domain is created independently from the other domains and describes the relationships between observed objects and phenomena via the properties of these classes. To construct this domain, firstly, the observed object list and their relationships are specified. The second step defines the phenomena of interest. The influence between phenomena and observed objects are defined next.

The terms for observed objects are extracted mainly from AGROVOC the existing thesaurus created by Food and Agriculture Organization - FAO (FAO 2012) because most of the observed objects in this research relate to land cover and land use. The terms are also defined according to collected data of WISDOM. The relationships of observed objects were adopted from definitions of AGROVOC, for example, the term "agriculture" has the broader term "economic activities" and the related terms are "forestry" and "fisheries"[(3)]. In parallel, the list of phenomena is extracted from the SWEET ontology. The relationships between phenomenon and observed object are obtained from different definitions and dictionaries. The relationships can be updated anytime, whenever there are new consensuses from project partners or scientists. The changes in this domain do not compromise other domains.

[3] http://aims.fao.org/en/agrovoc-term-info?mytermcode=203&mylang_interface=en&myLanguage=EN

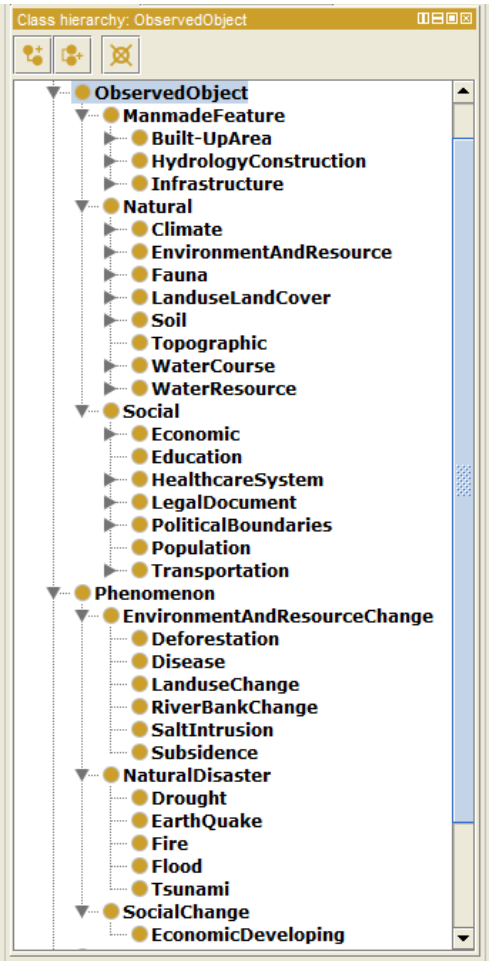


Figure 5.5: Outline of the classes hierarchy of the observed object domain (Viewed by Protégé version 4.1.0 – Build 239).

Figure 5.5 shows an abstract hierarchy of the observed object domain. The concepts presenting observed objects are grouped into three main classes – Infrastructure, Social and Natural basing on common concepts, i.e. human and nature (see section 1.1.2.1). This concept hierarchy is an abstract model of the real world and relates to water and collected data within WISDOM project. Under main classes, there are sub classes as shown in Figure 5.5.

- Infrastructure: consists of concepts presenting things constructed by human related to water (i.e. built-up area, hydrology construction and transportation).
- Natural: concepts of land use, land cover and environment.

- Social: concepts of human activities.

Phenomena are also grouped into three main classes. In this thesis, these groups are defined as follows.

- EnvironmentAndResourceChange: this class consists of sub classes related to environment and resource change, i.e. deforestation, disease, landuse change, riverbank change, salt intrusion and subsidence.
- NaturalDisaster: this class consists of subclasses for drought, earth quake, fire, flood and tsunami.
- SocialChange: economic developing is the subclass of SocialChange.

The influences of phenomena on observed objects are acquired from several paper and definitions, e.g. study guide for disaster management of Schramm (Schramm et al. 1986), or resource change in (Gorte et al. 2010). The figures in Appendix E show the relationships and the properties of the Observed object domain in more details.

Environmental Effects	
Effects	Inundation
Consequences	Damages structures, forces evacuation, erodes topsoil, may change course of streams, rivers; destroys most crops; deposits silt in some downstream areas that may not be beneficial
Effects of Natural Hazards	
On Land	Erosion, mudslides, silting
Structures	Undercuts foundations, buries structures
Agriculture	Destroys crops, changes cropping patterns, localized crops losses, improves soil
Tree	Reduces forests, localized timber losses

Table 5.2: An example for flood's effect from Schramm (Schramm et al. 1986)

Because this domain does not focus on the value of individual of classes (see section 0 for definition of individual), it focuses on defining the properties of classes on how they are related to each other. Thus, there are no data properties for this domain. The object properties

are “canObserve”, “affectedBy” with sub properties: “cause”, “change”, “decrease”, “destroy”, and “increase”. The classes in different branches of the observed object list may associate with each other by “relate” property.

Because the influence of observed objects and phenomena are very complex, this domain describes common relations which have widely been agreed upon such as AGROVOC thesaurus from FAO, study guide of Schramm, and Gorte (Schramm et al. 1986; Gorte et al. 2010; FAO 2012). Moreover, the relationships of the concepts are complicate. One concept may have many relation with others, thus, the constraint should not be assigned as “only”, but “some”. The “Only” constraint is known as universal quantifier or “allValuesFrom” restrictions that mean the set of individuals that, for a given property, *only* have relationships to other individuals of a specific class. They do not have any relationships along that property to any individual in other class. On the contrary, the “Some” constraint is known as existential quantifier or “someValuesFrom” restrictions. With this constraint, the set of individuals have at least one relationship to individuals of a specific class, and may have relationship with the individuals of other class (Horridge 2011a).

With defined properties and relationships, this domain can resolve the limitation of the cross related structural in managing the relationships of ProductGroup and thematic reference schema.

5.4. Application domain

This domain describes the user tasks in terms of the sufficient data needed to carry out the tasks, not in the way how to perform the tasks. The user tasks related to water domain are defined in this domain, e.g. monitoring and rescue task. The monitoring task is based on the concept of Suter (Suter 1993) who defined monitoring as “measurement of environmental characteristics over an extended period of time to determine status or trend in some aspect of environmental quality”. The rescue task is divided into sub tasks, i.e. Search and find, first aid, food and transport.

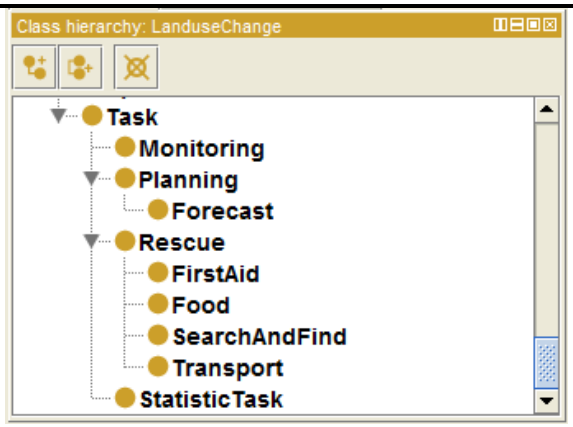


Figure 5.6: Classes hierarchy of application domain
(Viewed by Protégé version 4.1.0 – Build 239).

The main property in this domain is "hasMainConcern" which describes the relation of user tasks and observed object classes. It defines which kinds of dataset are sufficient for carry out a certain task.

Two types of relationship are defined as below.

- Direct relationship: where the property is assigned directly to class(es), for example, the task rescue has a sub task FirstAid as shown in Figure 5.6, FirstAid task has the main concern about the healthcare system which is an observed object class. So, it is assigned a property like "FirstAid - hasMainConcern - HealthcareSystem" (Figure 5.7).

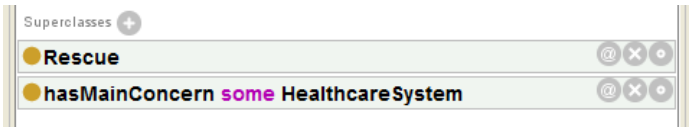


Figure 5.7: An example of a direct relationship of FirstAid task
(Viewed by Protégé version 4.1.0 – Build 239).

- Indirect relationship: where the property is assigned via a definition of class(es). For example, the task Monitoring is defined as a task which has the main concern in some classes which can observe a phenomenon, however, these classes do not determine which phenomenon. That means any observed objects, which can observe any phenomenon, should be the main concern of monitoring task. So the monitoring task

is assigned like "Monitoring – hasMainConcern – (ObservedObject – canObserve - Phenomenon)" as shown in Figure 5.8.

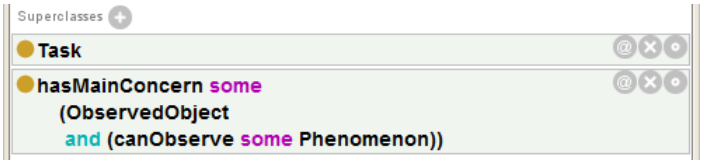


Figure 5.8: An example of indirect relationship of Monitoring task
(Viewed by Protégé version 4.1.0 – Build 239).

The task definitions in this domain are based on the common vocabularies and existing knowledge mentioned above, and they can be changed or updated during the life cycle of the system. With a certain query, a certain task can be combines with a phenomenon, and then via properties of phenomena and observed object, observed object classes are determines. Finally, individuals of these classes (datasets) are discovered and provided to the user. The figures in Appendix F show the relationships and the properties in more details.

5.5. Spatial and Temporal domain

These two domains present information of the temporal and spatial reference aspects of the WISDOM data management model (see section 4.2.2). The spatial domain consists of classes presenting the administrative levels in a hierarchical structure. The SWEET spatial ontology is applied for spatial domain in this study. The administrative objects link to the datasets by the "hasCoverArea" property. Similarly, the temporal domain also adopts the concepts from the temporal ontology recommended by W3C (W3C 2006) provided by Feng Pan 2005(Pan et al. 2005). The valid period of time of the datasets are mapped to the temporal domain as individuals of the "Instant" class which link to datasets by properties "hasStartDate" and "hasEndDate".

5.6. Implementation of a prototypical Graphical User Interface

This section presents the tools and software used to build the GUI and the functions of the GUI.

5.6.1. The used tools and software.

The prototype is built based on the Java programming language using Eclipse software, which is an integrated development environment (IDE). Eclipse provides a software development environment (Eclipse 2012). The prototype has a convenient graphical user interface with functions as shown below.

- The prototype can connect to WISDOM database.
- Data from RDB are mapped to RDF, and then the RDF files are integrated with ontology file.
- User queries are translated to SPARQL to query RDF. The list of datasets returned from the system is shown with the ranking level and the relationships with phenomena.

D2RQ is applied to map the values from RDB to the corresponding defined concepts in RDF. Jena, a Java framework for building Semantic Web applications is also used. Its libraries are added to Eclipse user library to read, write and merge RDF to the ontology files.

Pellet reasoner is used to infer implicit information in the ontology file. It runs whenever there is new dataset imported into the database in order to infer new results that are stored in the memory or written to the file. The user queries are translated to SPARQL syntax within java code.

The returned result is ranked following the ranking method proposed in chapter 3. And then the final result is shown in the GUI. The appendix G shows some main parts of the source code of the prototype.

5.6.2. The Graphical User Interface

The prototype offers two tabs, which are marked in boxes in Figure 5.10 to search for data. In the first tab the users can search for observed objects and in the second one the users can search for phenomena and user tasks.

To search for the observed objects, the users have to define the object they are interested in; the administrative area they want to search for and the period of time they focus on (Figure 5.10). The users can indicate which criterion is the most important for their queries by adjusting the slider. The ranking points help the users to define which datasets are most appropriate to their queries. The users can reduce the list of datasets by choosing the format type of data which they prefer.

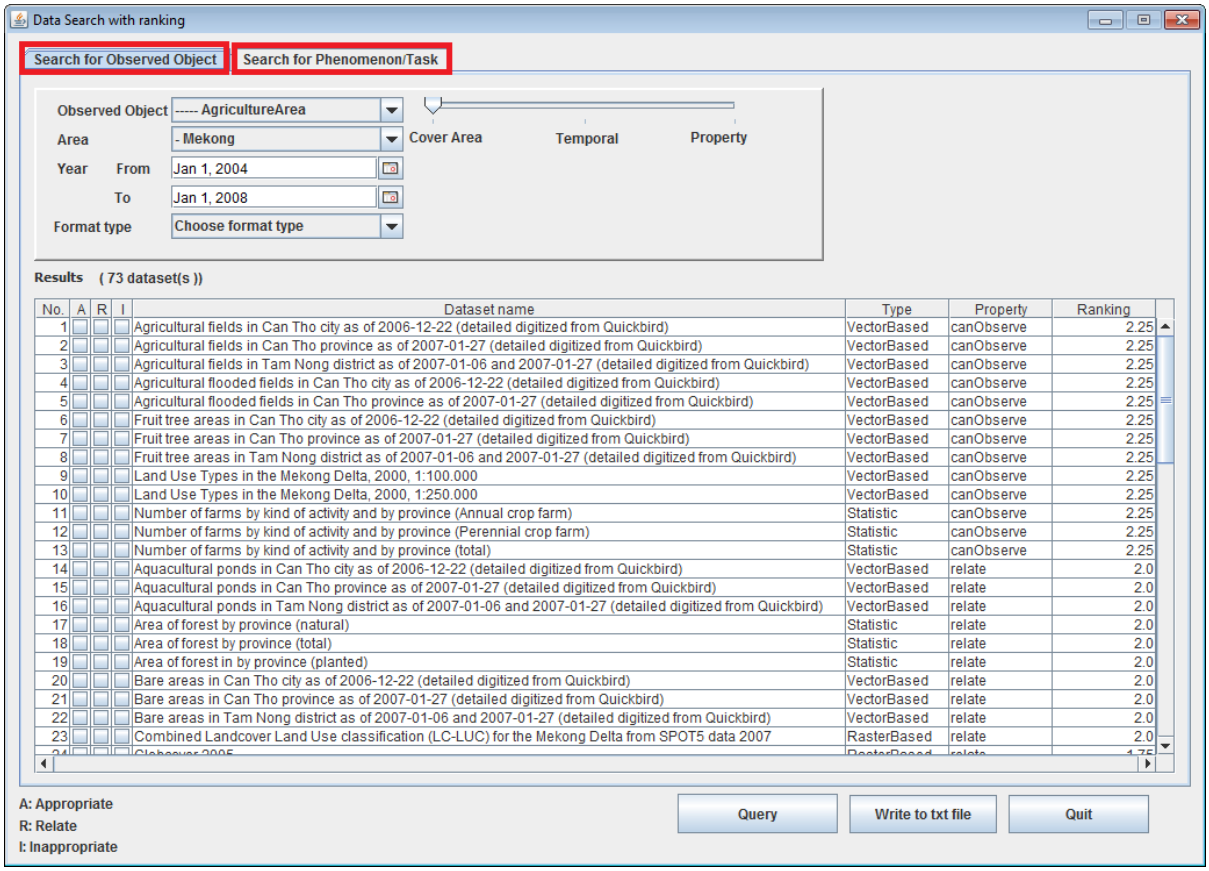


Figure 5.9: The GUI of the prototype for Observed Object search

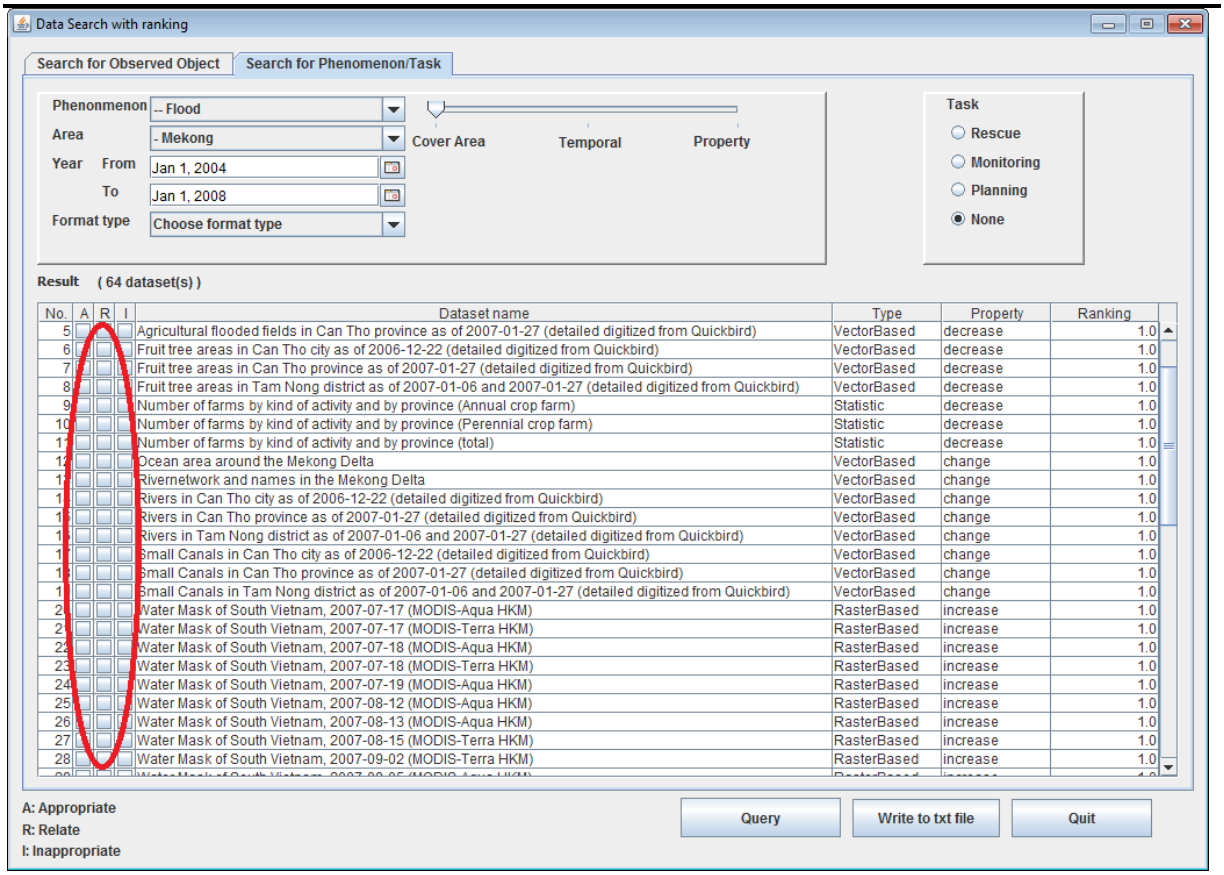


Figure 5.10: The GUI of the prototype for phenomena search.

The users can also search for a certain phenomenon or can combine a phenomenon with a task. As shown in Figure 5.10, the criteria to search for a phenomenon are similar to the first tab. To combine a phenomenon with a task, the users have to define the phenomenon and then choose a predefined task in the GUI. The result shows the ranked list of the properties which show the relationship between datasets with the phenomenon.

The prototype is also used to test the datasets whether they are sufficient for the user query. The testers choose one of the check boxes, which are shown in the ellipse in Figure 5.10, to mark which datasets are sufficient, relate or insufficient to their queries. The next chapter describes the testing scenarios and evaluation step in more details.

5.7. Conclusion

This chapter provides a detailed view of the most interesting parts of the implementation of the prototype. A semantic layer is built on top of existing system. It has two sub layer, i.e. RDF and ontology layer. The RDF layer maps the value from existing RDB to RDF and then implies with the ontology files in the ontology layer in order to present the semantic of the datasets. The prototype provides a tool to search for data with a user friendly and simple graphical interface. The prototype provides a facility way to search for data. With one search, the users can retrieve all relevant data they need. The result is ranked based on several criteria showing the list of datasets in order from the most appropriate to the least in comparison with the user's query.

Furthermore, the proposed approach overcomes the limitations of cross relational structure, which are mentioned in section 4.3. It is also easy to maintain and to extend. The next chapter evaluates the query result from the prototype in comparison with the user demands.

6. EVALUATION

This chapter assesses the feasibility of the proposed approach in this study. The feasibility assessment applies the widely known precision and recall criteria (Raghavan et al. 1989; Buckland et al. 1994; NIST 2001; Zhu 2004; Goutte et al. 2005; Webber 2010). In the information retrieval context, precision and recall are calculated based on the ratio of the expected results and the effective correspondences which are relevant for the user queries and which are not. Specifically, the precision measures the ratio of number of retrieved relevant datasets over total number of retrieved datasets that indicates the degree of correctness of the system. Meanwhile, the recall measures the ratio of number of retrieved relevant datasets over total number of relevant datasets which should be retrieved. The recall logically measures the missing relevant datasets which should be provided to the users. Furthermore, the ranking method is also evaluated with average precision criterion. Average precision considers the position of relevant datasets in the retrieved list. It can combine precision, relevant ranking and recall in a single value (Zhu 2004).

The next sections present the evaluation of the approach in more detail. This evaluation does not include the performance of the system in terms of the speed of the search process.

6.1. Precision and recall

As mentioned above, the precision and the recall are taken into account to estimate the viability of the proposed approach. They are common criteria to measure the quality of a searching method and are calculated by the functions shown below (Raghavan et al. 1989; Buckland et al. 1994; NIST 2001; Zhu 2004; Goutte et al. 2005; Webber 2010).

The functions (1) and (2) show how to calculate precision and recall.

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}} \quad (1)$$

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}} \quad (2)$$

Regarding to the function (1), the precision indicates how many percent of the returned result relates to the user query. The recall, which is shown in function (2), indicates the ratio of the relevant datasets retrieved to the total number of relevant datasets in the database. The values of precision and recall are from zero to one. Figure 6.1 shows, what the precision and the recall are in an intuitive way. Although, they are the most common measures of search performance, there is always a contradiction between recall and precision (Cleverdon 1972).

- If we have 100% recall, that means all relevant documents were retrieved, but maybe also many non-relevant ones.
- If we have 100% precision, that means all retrieved documents were relevant, but maybe not all relevant documents were retrieved. (Raghavan et al. 1989)

In the past, there were many discussions and researches about the inverse relationship between precision and recall, but now, it is generally accepted. In this study, the inverse relationship has not been discussed, it is adopted based on the research of (Cleverdon 1972; Heine 1973; Jones 1981).

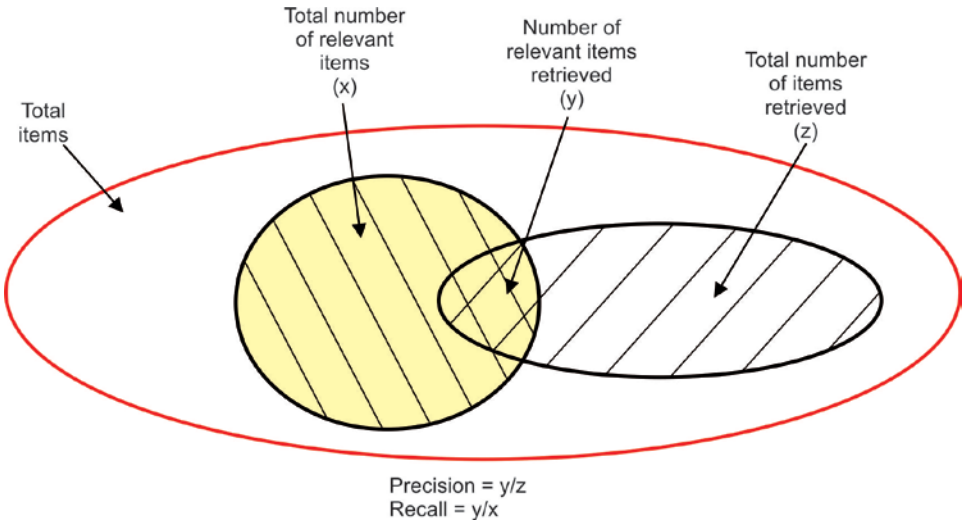


Figure 6.1: An example of precision and recall

To calculate these two criteria, it is vital to specify which datasets are relevant for a certain query. That is really difficult; because the classification of a dataset to be relevant or not differs depending on the view point of the discipline from which a user looks on the datasets. Actually, the notion "relevance" is definitely subjective to the tester. It depends on the individual perception: "what is relevant to one person may not be relevant to another" (Malik 2006). In fact, the more test cases are done, the better evaluation is because the subjectivity is reduced. However, because of the limitation of time and budget, there are fourteen test cases which have been done by ten testers (Table 6.1) for the proposed approach.

To evaluate the proposed approach, the returned results are compared with the user estimation. As mentioned above in the section 5.6.2, the prototype is also used for testing. Ten testers have been chosen from different knowledge levels and disciplines. They have different backgrounds, experience and interests as well. The testers have to mark one of three check boxes as shown in Figure 5.10 that indicate which datasets are appropriate, related or inappropriate for a particular query. The test cases have been chosen as below.

- Two observed objects with different administrative levels.
- Four phenomena with different administrative levels.
- One phenomenon combined with a task.

The Table 6.1 shows the list of the test cases which have been done in this study.

No.	**Observed object / Phenomenon / Phenomenon and task**		**Region**	**Period of time**
1	Observed Object	Agriculture area	Mekong Delta	01-01-2004 to 01-01-2008
2		Agriculture area	Cần Thơ	01-01-2004 to 01-01-2008
3		Healthcare system	Mekong Delta	01-01-2004 to 01-01-2008
4		Healthcare system	Cần Thơ	01-01-2004 to 01-01-2008
5	Phenomenon	Landuse change	Mekong Delta	01-01-2004 to 01-01-2008
6		Landuse change	Cần Thơ	01-01-2004 to 01-01-2008
7		River bank change	Mekong Delta	01-01-2004 to 01-01-2008
8		River bank change	Cần Thơ	01-01-2004 to 01-01-2008
9		Flood	Mekong Delta	01-01-2004 to 01-01-2008
10		Flood	Cần Thơ	01-01-2004 to 01-01-2008
11		Drought	Mekong Delta	01-01-2004 to 01-01-2008
12		Drought	Cần Thơ	01-01-2004 to 01-01-2008

13	Phenomenon and task	Flood and Rescue	Mekong Delta	01-01-2004 to 01-01-2008
14		Flood and Rescue	Cần Thơ	01-01-2004 to 01-01-2008

Table 6.1: The list of test cases

The Table 6.2 shows the precision values of the test cases which have done by the testers (from the tester T1 to T10). In this table, the precision is calculated by both the appropriate and related values. They are summed and then divided by the total number of datasets retrieved.

No.	**Search criteria**	**Tester**									
		T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
1	Agriculture Area Mekong Delta 01-01-2004 to 01-01-2008	1.00	1.00	0.75	0.67	0.54	0.71	0.88	0.75	0.63	1.00
2	Agriculture Area Cần Thơ 01-01-2004 to 01-01-2008	0.79	0.75	0.79	0.58	0.58	0.54	1.00	0.58	0.58	0.88
3	Healthcare System Mekong Delta 01-01-2004 to 01-01-2008	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	Healthcare System Cần Thơ 01-01-2004 to 01-01-2008	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	Landuse change Mekong Delta 01-01-2004 to 01-01-2008	1.00	0.95	1.00	0.80	1.00	1.00	0.68	0.68	0.10	0.90
6	Landuse change Cần Thơ 01-01-2004 to 01-01-2008	0.83	0.72	0.78	0.50	0.83	0.78	0.65	0.65	0.05	0.78
7	River bank change Mekong Delta 01-01-2004 to 01-01-2008	1.00	1.00	1.00	1.00	0.88	1.00	0.88	0.88	1.00	1.00
8	River bank change Cần Thơ 01-01-2004 to 01-01-2008	0.63	0.75	0.75	0.75	0.63	0.75	0.63	0.63	0.75	0.75
9	Flood Mekong Delta 01-01-2004 to 01-01-2008	0.99	0.76	0.74	0.78	0.72	0.96	0.95	0.72	0.58	1.00
10	Flood Cần Thơ 01-01-2004 to 01-01-2008	0.82	0.69	0.69	0.65	0.65	0.86	0.70	0.65	0.22	0.91
11	Drought Mekong Delta 01-01-2004 to 01-01-2008	1.00	0.88	0.87	0.95	0.92	0.96	0.99	0.87	0.87	1.00
12	Drought Cần Thơ 01-01-2004 to 01-01-2008	0.99	0.88	0.87	0.88	0.90	0.94	0.87	0.87	0.87	0.98

13	Flood – Rescue Mekong Delta 01-01-2004 to 01-01-2008	0.99	0.98	0.79	0.65	0.85	0.24	0.85	0.85	0.81	1.00
14	Flood – Rescue Cần Thơ 01-01-2004 to 01-01-2008	0.87	0.90	0.87	0.63	0.72	0.22	0.76	0.76	0.73	0.89

Table 6.2: The precision of the test cases have been done by testers

A graphical presentation of the precisions is shown in Figure 6.2.

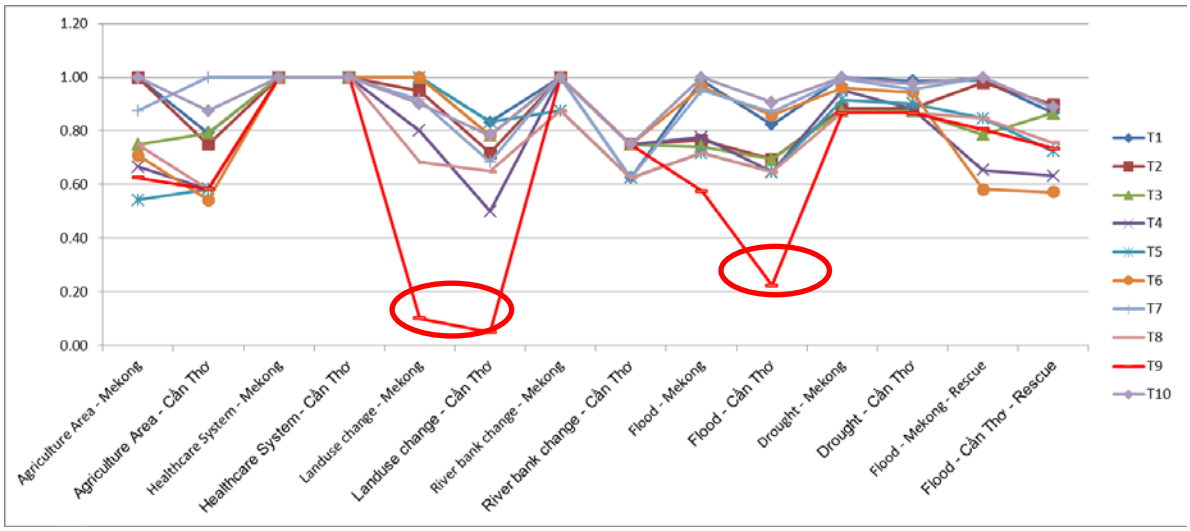


Figure 6.2: The precision (1)

There are some test cases that have low precision such as the cases in the red ellipse in Figure 6.2. These test cases are the searching for phenomenon. In fact, the influents of a phenomena over observed object are sophisticated, moreover, the testers make relevance judgments based on their knowledge, needs, and others factors specific to them. That is the reason why the precision is low in some test cases. However, in general, the test cases have good results as shown in Figure 6.3.

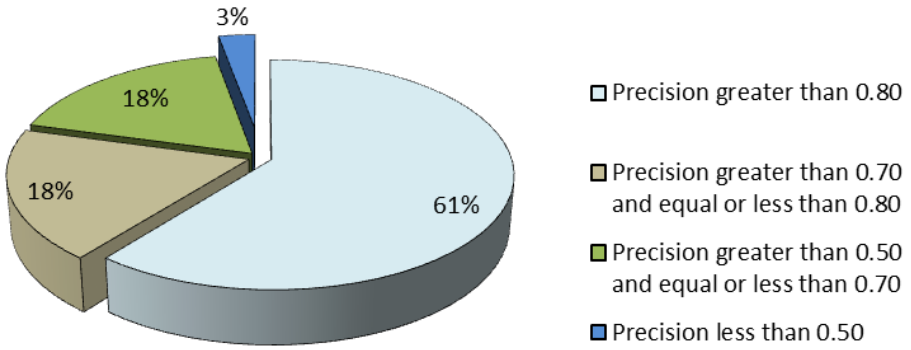


Figure 6.3: The precision (2)

The statistic shows that most of the test cases have precision greater than 0.50. The Figure 6.3 shows that the test cases which have precision less than 0.50 are 3 percent. Meanwhile, 61 percent of the test cases have precision greater than 0.80. And, 97 percent of them have precision greater than 0.50. That proves the returned results are sufficient for the user demands.

To assess the recall, it is also a time consuming task to scan the whole list of available datasets in the database. At the time of this test, the system consists of nearly one thousand six hundred datasets. Therefore, there are only four test cases, which have been done. These test cases have same search criteria as the last test (see in the Table 6.3 and Table 6.1). Every test case has been done with two testers, who have different background and experience to reduce the subjectivity. During the test, the representative users will provide a list of datasets which they require from the list of available datasets, that list will be compared with results from the system. The analysis has been done to evaluate the viability of the approach.

The Table 6.3 shows the values of the recall for the four test cases. The relevant retrieved values are calculated in combination with the result from the last tests for precision criteria.

No.	**Search Criteria**	**Tester**	**Relevant**	**Not relevant**	**Total dataset**	**Relevant retrieved**	**Recall** (relevant retrieved / Relevant)
1	Agriculture area Cần Thơ 01-01-2004 – 01-01-2008	T1	46	1535	1581	43	**0.93**
		T2	42	1539	1581	30	**0.71**

2	Landuse change Mekong Delta 01-01-2004 – 01-01-2008	T1	67	1514	1581	59	**0.88**
		T5	147	1434	1581	58	**0.39**
3	Flood and rescue Cần Thơ 01-01-2004 – 01-01-2008	T2	150	1431	1581	85	**0.57**
		T5	350	1231	1581	70	**0.20**
4	Drought Mekong Delta 01-01-2004 – 01-01-2008	T2	322	1259	1581	263	**0.82**
		T3	260	1321	1581	253	**0.97**

Table 6.3: The recall of the four test cases

According to (Xie 2005), for recall values, if the value is greater than 0.50, it is acceptable. The Figure 6.4 shows that in most of the test cases, the recall is higher than 0.50, i.e. 50 percent of test cases is greater than 0.80; 12 percent is between 0.70 and 0.80; 13 percent is between 0.50 and 0.70; and only 25 percent test cases is less than 0.50.

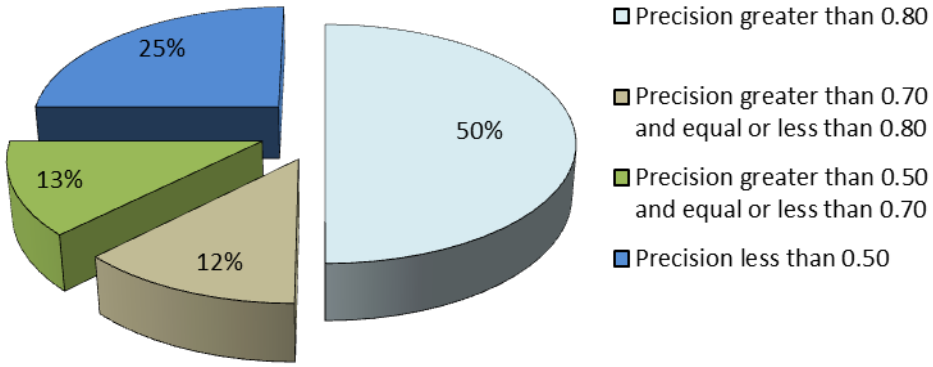


Figure 6.4: The recall values

Although, precision and recall are popular criteria for evaluating the feasibility of a search method, they are measures for the entire retrieved list of datasets. They cannot estimate the quality of the ranking method. In fact, users want the retrieved documents to be ranked according to their relevance level (Webber 2010). The most relevant datasets must be in the top of the showing list. The next section presents the average precision criterion which can assess the ranking method.

6.2. Average precision

This section evaluates the ranking method proposed approach using the average precision (AP) which is defined as "the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved" (Buckley et al. 2000). That means, average precision is the sum of the precision at each relevant document retrieved in the result divided by the total number of relevant documents. It can be calculated in two different ways as shown in the next sections.

6.2.1. Average precision at seen relevant documents

The average precision is the most widely adopted criterion to evaluate the information system in terms of the relevant retrieved items (Webber 2010). In the case of the search result which returns a ranked sequence of datasets, it is important to consider the order in which the returned datasets are presented. According to (Webber 2010), the average precision (AP) criterion calculates the precision in combination with the position of the datasets in the showing list. Thus, it is also appropriate to estimate the ranking method. The function (3) and (4) shows how to calculate average precision at seen relevant documents (Campos 2007). It is calculated based on the position of the relevant documents which are retrieved for a certain query.

$$AP = \frac{\sum_{k=1}^{n} \left(P(k)\ x\ rel(k)\right)}{number\ of\ relevant\ datasets\ retrieved} \qquad (3)$$

$$P(k) = \frac{Number\ of\ relevant\ items\ retrieved\ at\ k\ position}{Total\ number\ of\ items\ retrieved\ at\ k\ position} \qquad (4)$$

Where

k: the rank position in the sequence of retrieved datasets

n: the number of retrieved datasets

P(k): the precision at cut-off k position in the list

rel(k): an indicator function. Equal 1 if the item at rank *k* is a relevant document, and zero (0) otherwise.

No.	Dataset Retrieved from system I	Relevant items retrieved from system I	Dataset Retrieved from system II	Relevant items retrieved from system II
1	A		E	x
2	B		G	x
3	C		H	x
4	D		I	x
5	E	x	A	
6	F		B	
7	G	x	C	
8	H	x	D	
9	I	x	F	
10	J	x	M	
Precision:		0.50		0.40
Average Precision at seen documents:		0.36		1.00

Table 6.4: Example of average precision from two different systems

$$AP\ for\ system\ I = \frac{\frac{1}{5}x\,1 + \frac{2}{7}x1 + \frac{3}{8}x1 + \frac{4}{9}x1 + \frac{5}{10}x1}{5} = 0.36 \qquad (5)$$

$$AP\ for\ system\ II = \frac{\frac{1}{1}x\,1 + \frac{2}{2}x1 + \frac{3}{3}x1 + \frac{4}{4}x1}{4} = 1.00 \qquad (6)$$

For example, there are two different search engines which work on the same database. The system I retrieves 5 relevant documents and the system II retrieves 4 relevant documents. The results are ranked as shown in Table 6.4. The average precision values for two systems are

calculated as (5) and (6). Table 6.4 shows the results return from two systems, with ranking order, and the precision and the average precision. The assessment is:

- System I has higher precision than system II
- However, the ranking method of the system I is worse than system II, so the AP of the system I is lower than the system II.
- As a consequence, the ranking method of the system II is assessed better than the system I.

Average precision is the most widely used evaluation metric for the systems which rank the returned list based on the relevance of the datasets to the user queries (Webber 2010). The example shows that AP is a sufficient criterion to evaluate the ranking method of the proposed approach in this study.

The Table 6.5 shows the average precision at seen relevant documents for each test case in this study. These values are compared with the value of 1.00, 0.80 and 0.50. The Table 6.6 and Figure 6.5 show that 29 percent test cases have AP equal 1.00; 20 percent test cases have AP between 0.8 and 1.00; 29 percent between 0.50 and 0.80. There are only 22 percent less than 0.50.

No.	**Search Criteria**	**Tester**									
		T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
1	Agriculture Area Mekong Delta 01-01-2004 to 01-01-2008	1.00	1.00	0.82	0.82	0.56	0.76	0.84	0.82	0.45	1.00
2	Agriculture Area Cần Thơ 01-01-2004 to 01-01-2008	0.70	0.64	0.65	0.49	0.41	0.51	1.00	0.56	0.37	0.73
3	Healthcare System Mekong Delta 01-01-2004 to 01-01-2008	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	Healthcare System Cần Thơ 01-01-2004 to 01-01-2008	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	Landuse Change Mekong Delta 01-01-2004 to 01-01-2008	1.00	0.94	1.00	0.73	1.00	1.00	0.84	0.55	0.02	0.85
6	Landuse change Cần Thơ 01-01-2004 to 01-01-2008	0.73	0.58	0.66	0.36	0.73	0.66	0.47	0.46	0.01	0.66
7	River bank change	1.00	1.00	1.00	1.00	0.66	1.00	1.00	0.66	1.00	1.00

	Mekong Delta 01-01-2004 to 01-01-2008										
8	River bank change Cần Thơ 01-01-2004 to 01-01-2008	0.47	0.81	0.81	0.81	0.47	0.81	0.81	0.47	0.81	0.81
9	Flood Mekong Delta 01-01-2004 to 01-01-2008	0.97	0.45	0.41	0.50	0.38	0.91	0.85	0.38	0.22	1.00
10	Flood Cần Thơ 01-01-2004 to 01-01-2008	0.63	0.36	0.36	0.30	0.30	0.68	0.70	0.30	0.04	0.80
11	Drought Mekong Delta 01-01-2004 to 01-01-2008	1.00	0.66	0.64	0.87	0.75	0.91	0.98	0.63	0.62	1.00
12	Drought Cần Thơ 01-01-2004 to 01-01-2008	0.94	0.66	0.64	0.65	0.72	0.84	0.85	0.62	0.62	0.92
13	Flood – Rescue Mekong Delta 01-01-2004 to 01-01-2008	0.96	0.95	0.51	0.31	0.62	0.24	1.00	0.63	0.59	1.00
14	Flood – Rescue Cần Thơ 01-01-2004 to 01-01-2008	0.72	0.81	0.68	0.29	0.42	0.22	0.77	0.47	0.46	0.77

Table 6.5: The average precision of the test cases

No.	**Search Criteria**	**Test cases have AP equal 1.00**	**Test cases have AP equal or greater than 0.80 and less than 1.00**	**Test cases have AP equal or greater than 0.50 and less than 0.80**	**Test cases have AP equal or less than 0.50**
1	Agriculture Area Mekong Delta 01-01-2004 to 01-01-2008	30	40	20	10
2	Agriculture Area Cần Thơ 01-01-2004 to 01-01-2008	10	00	60	30
3	Healthcare System Mekong Delta 01-01-2004 to 01-01-2008	100	00	00	00
4	Healthcare System Cần Thơ 01-01-2004 to 01-01-2008	100	00	00	00
5	Landuse change Mekong Delta 01-01-2004 to 01-01-2008	40	30	20	10
6	Landuse change Cần Thơ 01-01-2004 to 01-01-2008	00	00	60	40
7	River bank change Mekong Delta 01-01-2004 to 01-01-2008	80	00	20	00
8	River bank change	00	70	00	30

	Cần Thơ 01-01-2004 to 01-01-2008				
9	Flood Mekong Delta 01-01-2004 to 01-01-2008	10	30	10	50
10	Flood Cần Thơ 01-01-2004 to 01-01-2008	00	10	30	60
11	Drought Mekong Delta 01-01-2004 to 01-01-2008	20	30	50	00
12	Drought Cần Thơ 01-01-2004 to 01-01-2008	00	40	60	00
13	Flood – Rescue Mekong Delta 01-01-2004 to 01-01-2008	20	20	40	20
14	Flood – Rescue Cần Thơ 01-01-2004 to 01-01-2008	00	10	40	50

Table 6.6: The AP at seen relevant documents compare with 1.00, 0.80 and 0.50 (14 test cases with 10 testers and calculate by %)

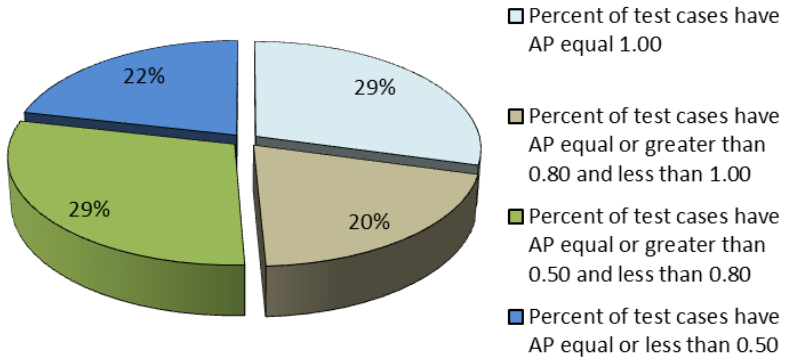


Figure 6.5: The AP at seen relevant documents compares with the value of 0.50, 0.80 and 1.00 for all test cases

The values of average precision at seen relevant documents shows that this criterion is good to assess the ranking method, but it does not present the recall value. It does not show how good the descriptions are in retrieving the related documents. For example, there are cases that system I retrieves nine relevant dataset which are ranked from position 2 to 9, and system II retrieves only one relevant document ranked at first position. So, in that cases, the value of AP

at seen relevant documents of the system I is (1/2+2/3+3/4+4/5+5/6+6/7+7/8+8/9)/9 equal 0.69, while, the value of the system II is 1. So, although, the value of the system II is greater than the system I, but the system I should be assessed better than the system II in terms of the total relevant documents retrieved. Therefore, there is another criterion, that can combine precision, ranking values and recall into one value. It is presented in the next section.

6.2.2. Average precision in combination with recall

According to (Zhu 2004), The AP can combine precision, ranking result and recall into a single score, which is known as single value summary. To do so, the average precision is calculated based on the function followed (function (7)).

$$AP = \frac{\sum_{k=1}^{n} \big(P(k)\ x\ rel(k)\big)}{total\ number\ of\ relevant\ datasets} \qquad (7)$$

Where

k: the rank position in the sequence of retrieved datasets

n: the number of retrieved datasets

P(k): the precision at cut-off k position in the list (use the function (4))

rel(k): an indicator function. Equal 1 if the item at rank *k* is a relevant document, and zero (0) otherwise.

By dividing the numerator by the total existing relevant datasets in database, the constraint of the AP at seen relevant documents mentioned above is solved (Rijsbergen 1979). The Table 6.7 shows the values AP in combination with recall for test cases in this study.

No.	Search Criteria	Tester	Average Precision
1	Agriculture area Cần Thơ 01-01-2004 – 01-01-2008	T1	0.74
		T2	0.53
2	Landuse change	T1	0.90

	Mekong Delta 01-01-2004 – 01-01-2008	T5	0.41
3	Flood and rescue Cần Thơ 01-01-2004 – 01-01-2008	T2	0.53
		T5	0.12
4	Drought Mekong Delta 01-01-2004 – 01-01-2008	T2	0.59
		T3	0.70

Table 6.7: The AP in combination with recall for test cases

The result of the evaluation shows that most of the values of AP in combination with recall for test cases are high. There are six test cases have value higher than 0.5. The evaluation shows that the proposed approach is good under the assumption that the test cases and the testers are representative to evaluate the search mechanism.

The main drawback of this function is, that it cannot be calculated in cases where the database is too large or not static. As recall needs to take all relevant datasets of the data basis into consideration, it can only be calculated for closed, static databases, that are not updated randomly. E.g. it is not possible to calculate the average precision, as described by (Zhu 2004), for search engines like google, yahoo or bing, as the data basis is changing rapidly.

6.3. Conclusion

The purpose of this study, in fact, is not to obtain the perfect descriptions which present the relations between datasets and observed objects or phenomena and user tasks. The proposed approach aims to provide relevant datasets for user requirements. These descriptions are based on the common knowledge that is widely agreed upon (see chapter 5) to provide more relevant datasets and their possible relations for a certain query.

The test cases show that the precision and the recall depend on many factors:

- The values of recall and precision depend very much on the knowledge, experience and discipline of the testers.

- The estimations of the influents between phenomenon and observed objects are different between disciplines. This does also affect the recall and precision values.
- As data are assigned to product group which, in turn, link to observed objects and phenomena and user tasks via relationships, the homogeneity of datasets in product group has also an effect on the recall and precision values. The datasets in a product group do not have the same properties in 100 percent, thus, one relevant document can be accompanied by non-relevant documents. That is one of the reasons of the inverse relationship between precision and recall.

The evaluation shows that the results have high precision and acceptable average precision as well. Moreover, the recall values are also acceptable. The evaluation also proves that the ranking method is good. The evaluation proves that the proposed approach is good and has high ability to apply in practice. Furthermore, the proposed approach overcomes the limitations of the cross relational structure on describing the relationship between datasets and thematic reference schema (see section 4.2.2).

7. CONCLUSION

Data heterogeneity is the primary issues in achieving good data integration from data of different disciplines into one information system. That makes finding and accessing appropriate data or information to answer scientific questions not straightforward. Moreover, it is difficult and time consuming for users to discover and collect all the relevant data for their works. In fact, they have to search many times as a trial and error process to find sufficient data for their demands.

This study provides an innovative approach which applies ontology to resolve the semantic heterogeneity of the collected datasets in an already existing water-related information system for the sustainable development of the Mekong Delta (WISDOM). Within this new approach, all datasets are described by linkages to observed objects, phenomena and user task so that all relevant datasets of different research fields are provided by only one search. The users can search for data at three different levels, i.e. Observed object, phenomenon or phenomenon in combination with a particular task. All in all, the ontology approach facilitates user search for data being more precise and suitable for their demands.

This chapter summarizes what has been done, and presents the main findings of this study. They are followed by the conclusion and recommendation sections.

7.1. Summary of findings

In most of the current systems, data are stored in a cross relational structured database and arranged into three aspects: thematic, spatial and temporal aspect. A case analysis on WISDOM project shows that the database with a cross relational structure is not able to manage semantic heterogeneity issues of collected data from different research fields.

Ontology is applied in this study to resolve the constraints of existing structure on describing the relationships between datasets and thematic reference groups. Ontology, which is defined as a "formal, explicit specification of a shared conceptualization", uses RDF, RDFs and OWL to describe the semantic of data in an explicitly way. RDF is a basic data model that identifies objects ("resources") and their relations to allow information to be exchanged between applications while conserving meaning. RDFs is a semantic extension of RDF, it describes the properties of generalization-hierarchies and class of RDF. OWL is on top of RDF and RDFs to add vocabulary to explicitly represent the meaning of terms and their classes' relationships.

The research questions have been answered with the proposed method, they are:

- **How to apply ontology to describe semantic of data sources?** The Data domain was designed to present the semantics of a data source. It contains many classes which present the properties of datasets, e.g. format type; geometric resolution – pixel size; spatial representation – line, point, polygon or pixel; and spatial relation - which area the datasets relate to; and thematic reference classes of datasets. Datasets in the RDB are assigned as individuals to datacollection class and have relation to corresponding thematic classes and another class (see section 3.2). This domain also has some rules and constraints to ensure the consistency of the model.
- **How to describe data in relationships of observed objects and phenomena?** The Observed object domain consists of classes that describe physical and non-physical objects related to the water subject, i.e. "man-made feature", "natural" and "social" which are called observed objects (see section 1.1.2.1). Phenomena are also presented concerning observed objects. The relationships in this domain are described independently from other domains. Therefore, the defined concepts in this domain are easy to combine with any tasks defined in application domain. The observed object, the phenomena and the relationships between them are adopted from the common dictionaries and definitions.
- **How to improve user search for data in the context to their task?** The Application domain describes the user's tasks divided into types, e.g. response task, monitoring task, etc. The user tasks are described in relation to observed objects which are the main concerns of these tasks. The task acts like constraints to limit the returned result regarding to a certain phenomenon (see section 3.4).

Since the proposed approach can retrieve all relevant datasets for user query, a ranking method is presented in order to rank the returned result based on their relevance level. The ranking method orders datasets based on their semantics, i.e. dataset cover areas, temporal entities and the level of relations between datasets; between datasets and observed objects.

A prototype has been built to evaluate and prove the feasibility of the proposed approach. It was built using JAVA, Jena, Eclipse, D2RQ and SPARQL. An evaluation has been done with particular test cases based on common criteria, i.e. precision, recall and average precision. These criteria assess the feasibility of proposed approach regarding the relevance of data for a certain query. The result of the evaluation proves the proposed approach is good and has high ability to be applied in practice.

7.2. Conclusion

Based on the results and discussion in the previous chapters, major conclusions can be drawn from this research as followed.

- Ontology is a good solution to overcome the constraints of the cross relational structure on describing the semantics of the collected data. It can provide relevant documents in one search more precise.
- Using ontology to describe the semantics of data have been applied in many existing research, however, describing the semantics of datasets in relation to observed objects, phenomena and user tasks is an innovative approach (see chapter 3).
- The descriptions are based on common knowledge and dictionaries, however, knowledge is not stable, it is constantly changing, developing and, sometimes, the new knowledge refutes the old one (Schön 1983; Swann 2010). Thus, the descriptions have to be checked and updated during the life cycle of the system in case there is any new definition coming from scientific community. That makes sure that the systems always apply the up to date definitions to describe the relationships and the properties of datasets. It ensures result returned from the system more precise and appropriate for user search.

- The ontology hybrid approach (see section 2.2.2), which is applied in this study, ensures for scalability and transferability to other research fields rather than water related. The hybrid approach describes the system into several domains, i.e. the data domain, the observed object domain and the application domain. Thus, whenever this proposed approach is applied for other research field, the class hierarchy and properties between classes in the domains can be changed to suit the purpose of the research field.
- In fact, it is impossible to create perfect descriptions which are accepted by everybody. So, this study tries to adopt common knowledge which is widely agreed upon to provide possible relevant datasets for a certain query. The evaluation shows that most of the test users accept the result returned from the proposed approach (see chapter 6). The evaluation of this study uses the widely adopted criteria, i.e. precision, recall and average precision. The precision and the recall are calculated based on the ratio of the expected results and the effective correspondences which are relevant for the user queries and which are not. The ranking method also assesses by average precision criterion. All the evaluation results show the proposed approach has high ability to be applied in practice (see chapter 6).

7.3. Recommendation

The following aspects are recommended for further works based on the results and discussion in the previous chapters and the conclusions of this study.

- This study proposes an innovative method and proves the correctness in terms of the methodology. However, as there are several definitions and models, it is necessary to have an investigation for concepts and their relation from experts in the fields of water related research. There should be more specific research and analysis for the user tasks.
- There should be to build a procedure to track and assess the user's choice in order to update the rules and the relationships between concepts during the life cycle of the system. The procedure records which datasets are the most choice for certain query,

and then the records are analysed to assess the relationship between datasets and the query criteria. For example, the dataset A are the most choice for the query B, the analyst defines which product group dataset A is belong to, and what is the properties from product group to observed object or phenomenon in the query B. That will help to improve the relationship between concepts of the domains, and the ranking method as well.

- Existing ontology should be considered in order to reuse definition about relations between administrative areas such as “nearby”, “neighbour”, etc. The GeoNames is a recommendation, which is a geographical database that covers all countries in the world and contains over eight million place names (GeoNames 2012). It should be integrated into observed objects domain to reuse the existing definitions for geospatial relationships between different administrative level areas, specifically the properties “nearby” and “neighbour”. It helps the user search data in more meaningful way, e.g. find rice fields near city A.
- Finally, the descriptions of observed object domain can be extended to the reason why and when a phenomenon happens. It is useful for users who want to analyse the reason why a phenomenon occurs. Moreover, these descriptions can be combined with the processing domain, for example, to become early warning system. The processing domain contains processes which can analyse the real time values such as water level from buoy or rainfall, and then combine with descriptions of the phenomena in order to predict or warn where and when a phenomenon can happen.

As mentioned in previous section, this study does not evaluate the speed of the search process. Thus, it should be tested how good the search mechanism works in terms of speed. The solution to improve the response time of the system has to be found, so that the new search functionality can be applied for the current WISDOM system.

Appendices

A. List of ISO/TC 211 Standards

STANDARDS THAT SPECIFY THE INFRASTRUCTURE FOR GEOSPATIAL STANDARDIZATION	
ISO 19101 Geographic information	Reference model
ISO/TS 19103 Geographic information	Conceptual schema language
ISO/TS 19104 Geographic information	Terminology
ISO 19105 Geographic information	Conformance and testing
ISO 19106 Geographic information	Profiles
STANDARDS THAT DESCRIBE DATA MODELS FOR GEOGRAPHIC INFORMATION	
ISO 19109 Geographic information	Rules for application schema
ISO 19107 Geographic information	Spatial schema
ISO 19137 Geographic information	Core profile of the spatial schema
ISO 19123 Geographic information	Schema for coverage geometry and functions
ISO 19108 Geographic information	Temporal schema
ISO 19141 Geographic information	Schema for moving features
ISO 19111 Geographic information	Spatial referencing by coordinates
ISO 19112 Geographic information	Spatial referencing by geographic identifiers
STANDARDS FOR GEOGRAPHIC INFORMATION MANAGEMENT	
ISO 19110 Geographic information	Methodology for feature cataloguing
ISO 19115 Geographic information	Metadata
ISO 19113 Geographic information	Quality principles
ISO 19114 Geographic information	Quality evaluation procedures
ISO 19131 Geographic information	Data product specifications
ISO 19135 Geographic information	Procedures for item registration
ISO/TS 19127 Geographic information	Geodetic codes and parameters
ISO/TS 19138 Geographic information	Data quality measures
STANDARDS FOR GEOGRAPHIC INFORMATION SERVICES	
ISO 19119 Geographic information	Services
ISO 19116 Geographic information	Positioning services
ISO 19117 Geographic information	Portrayal
ISO 19125-1 Geographic information	Simple feature access — Part 1: Common architecture
ISO 19125-2 Geographic information	Simple feature access — Part 2: SQL option
ISO 19128 Geographic information	Web map server interface
ISO 19132 Geographic information	Location based services — Reference model
ISO 19133 Geographic information	Location based services — Tracking and navigation
ISO 19134 Geographic information	Location base services — Multimodal routing and navigation
STANDARDS FOR ENCODING OF GEOGRAPHIC INFORMATION	
ISO 19118 Geographic information	Encoding
ISO 6709 Standard representation of	

geographic point location by coordinates	
ISO 19136 Geographic information	Geography Markup Language (GML)
ISO/TS 19139 Geographic information	Metadata — XML schema implementation
STANDARDS FOR SPECIFIC THEMATIC AREAS	
ISO/TS 19101-2 Geographic information	Reference model — Part 2: Imagery
ISO 19115-2 Geographic information	Metadata — Part 2: Extensions for imagery and gridded data

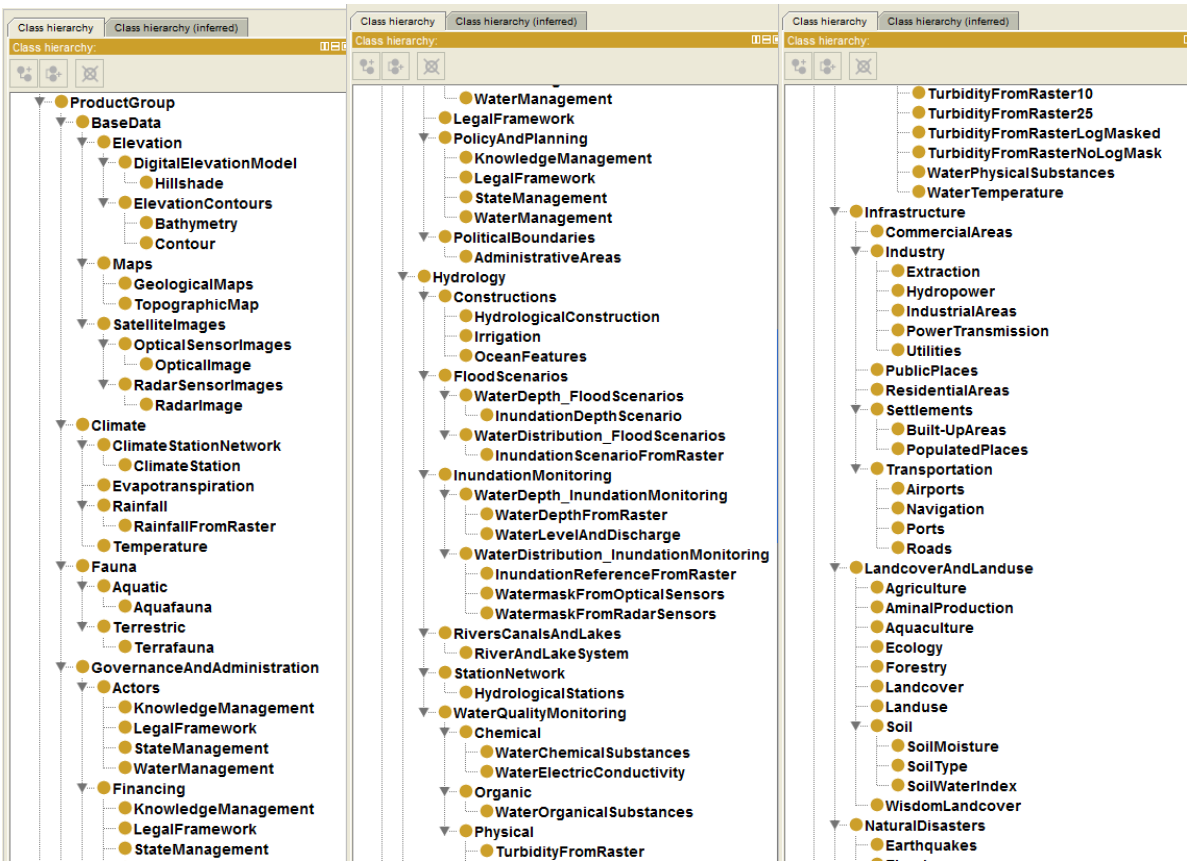
B. ISO 19115:2003

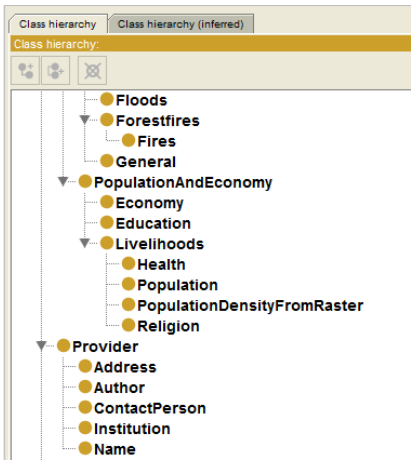
The ISO 19115:2003 defines abstract guidelines of metadata for geospatial data product in fourteen parts (adopted from (ISO 2003; Zhao et al. 2006)):

1. Metadata entity set (MD_Metadata) is the mandatory part that includes identification, content, portrayal catalog, distribution, metadata extension, and application schema.
2. Identification (MD_identification) uniquely identifies the data by defining format, graphic overview, specific uses, constraints, keywords, maintenance and aggregate information.
3. Constraint (MD_Constraints) defines the restrictions placed on the data.
4. Data quality (DQ_DataQuality) contains quality of the dataset and information about the source and production processes.
5. Maintenance (MD_MaintenanceInformation) describes the scope and frequency of updating.
6. Spatial representation (MD_SpatialRepresentation) points out the mechanism to represent spatial information.
7. Reference system (MD_ReferenceSystem) describes spatial and temporal reference system.
8. Content (MD_ContentInformation) identifies the feature catalog.
9. Portrayal catalog (MD_PortrayalCatalogReference) gives the type for displaying data.
10. Distribution (MD_Distribution) describes the distributor of the data.

11. Metadata extension (MD_MetadataExtentionInformation) is for user-specified extensions.
12. Application schema (MD_ApplicationSchemaInformation) is for the schema used to build a dataset.
13. Extent (EX_Extent) describes the spatial and temporal extent.
14. Citation and responsible party (CI_Citation)

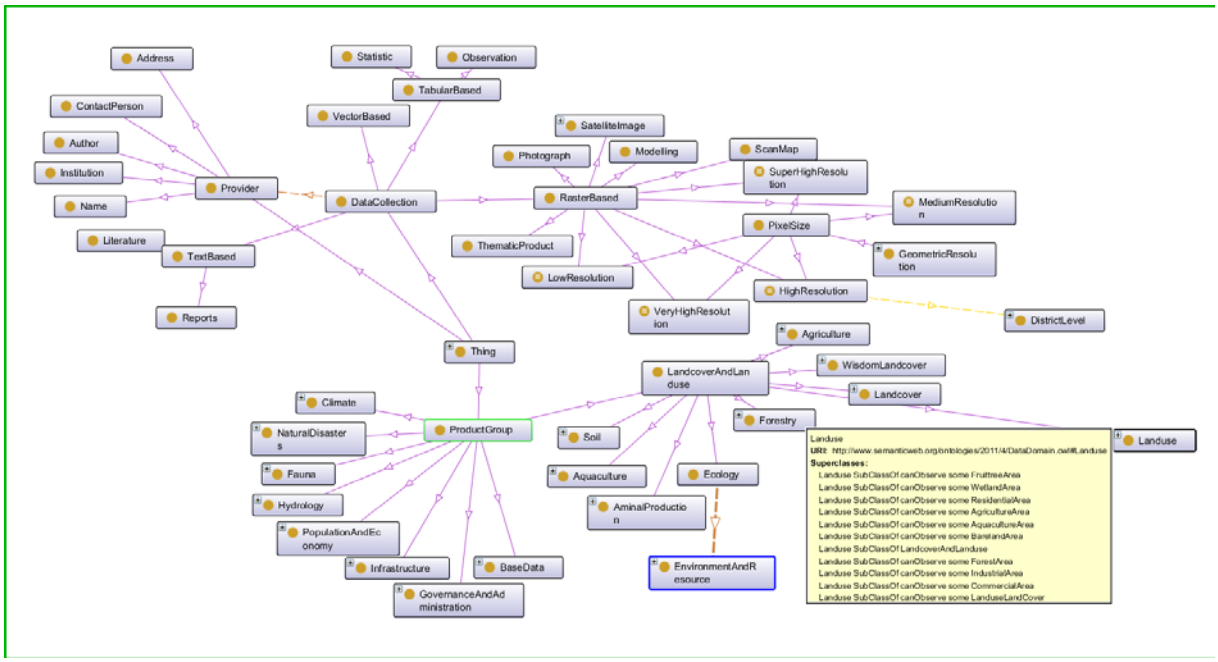
C. List of ProductGroup

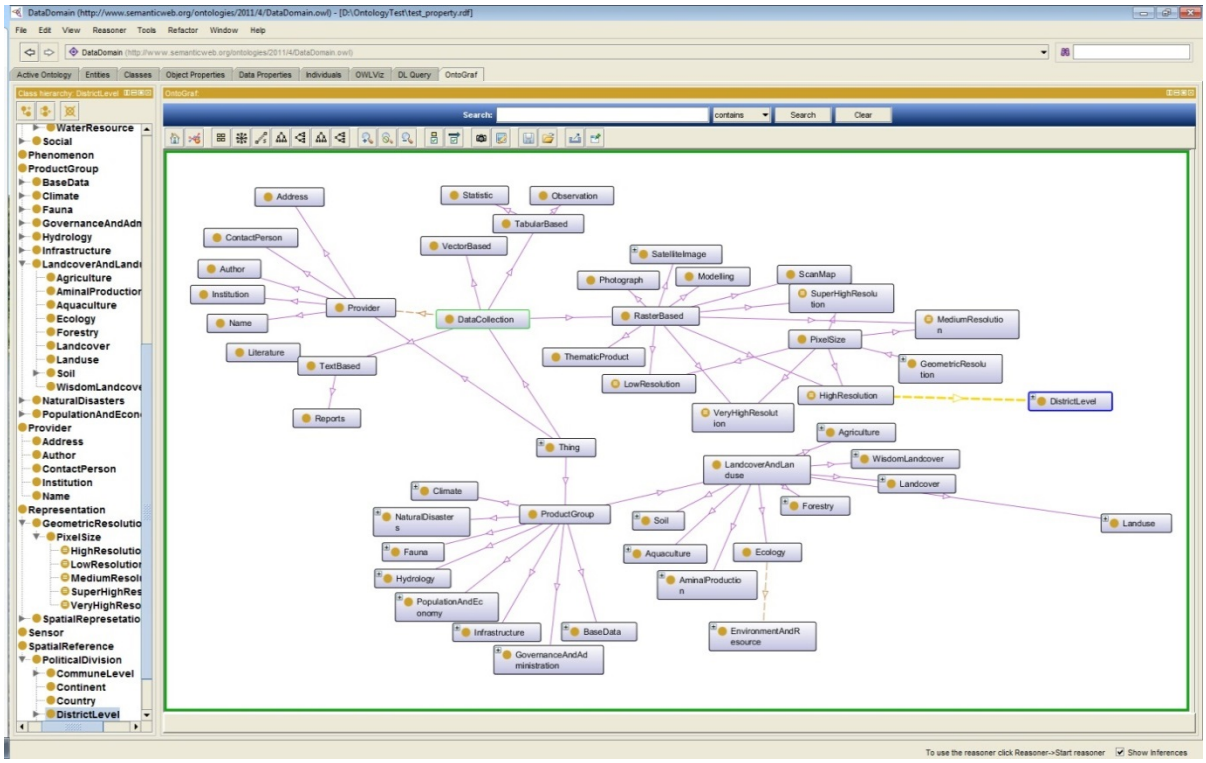


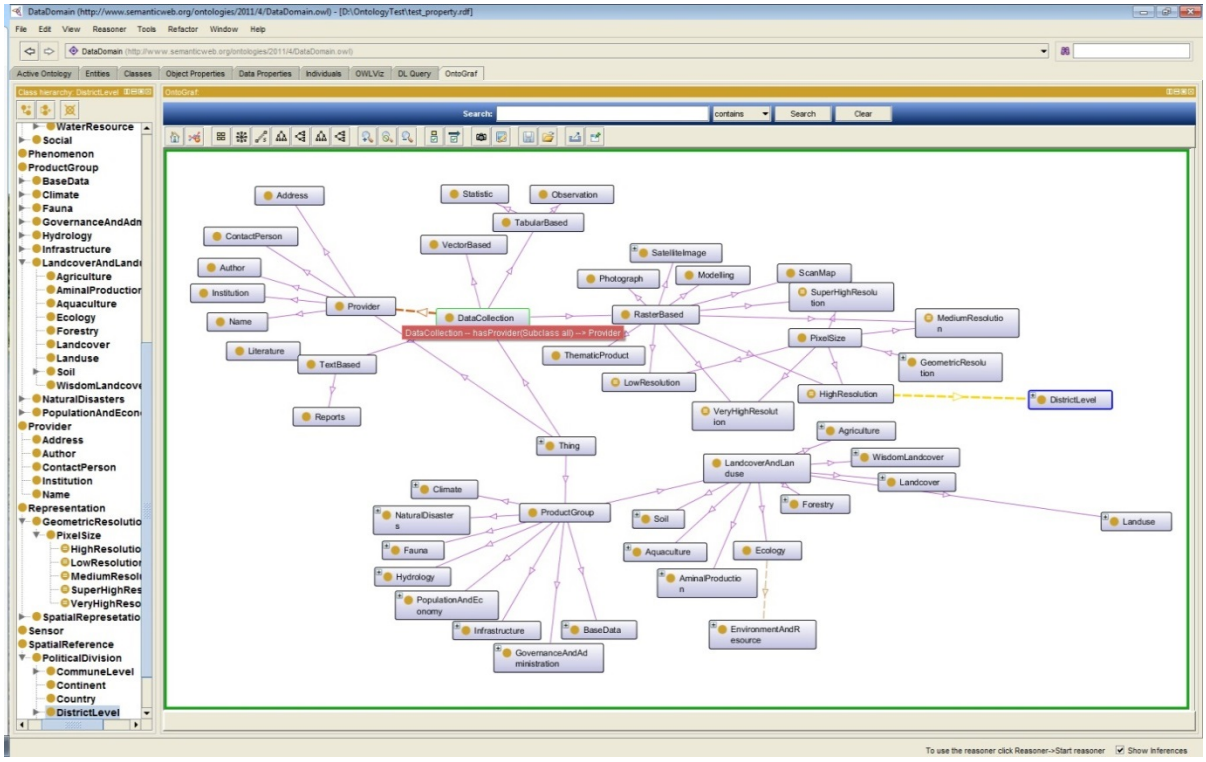


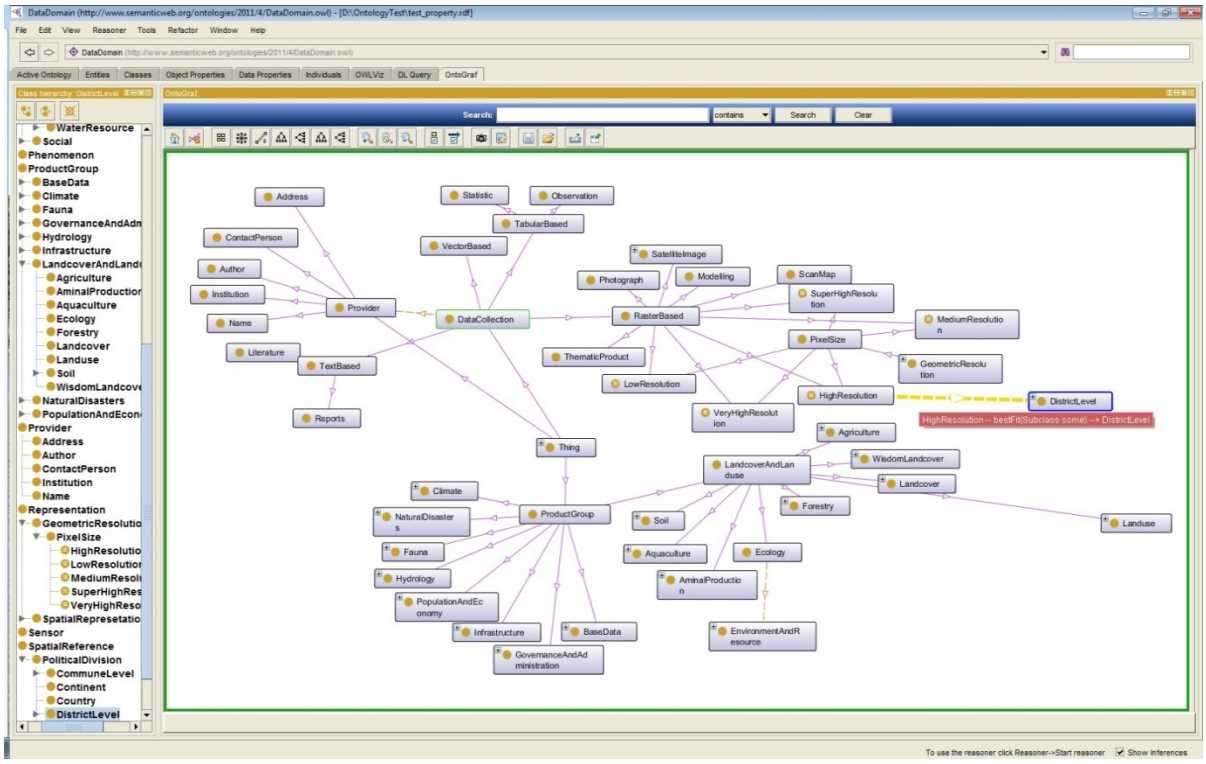
List of ProductGroup (viewed by Protégé version 4.1.0 – Build 239)

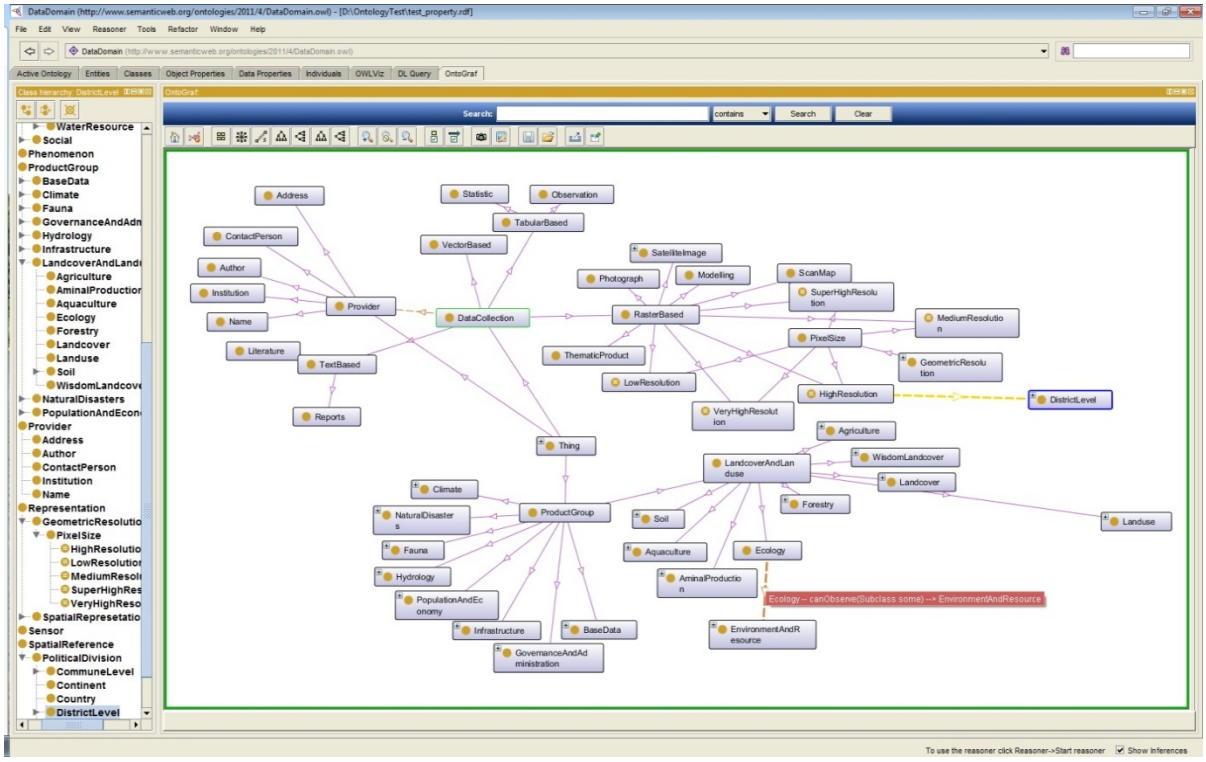
D. The Relationships and Properties in Data Domain



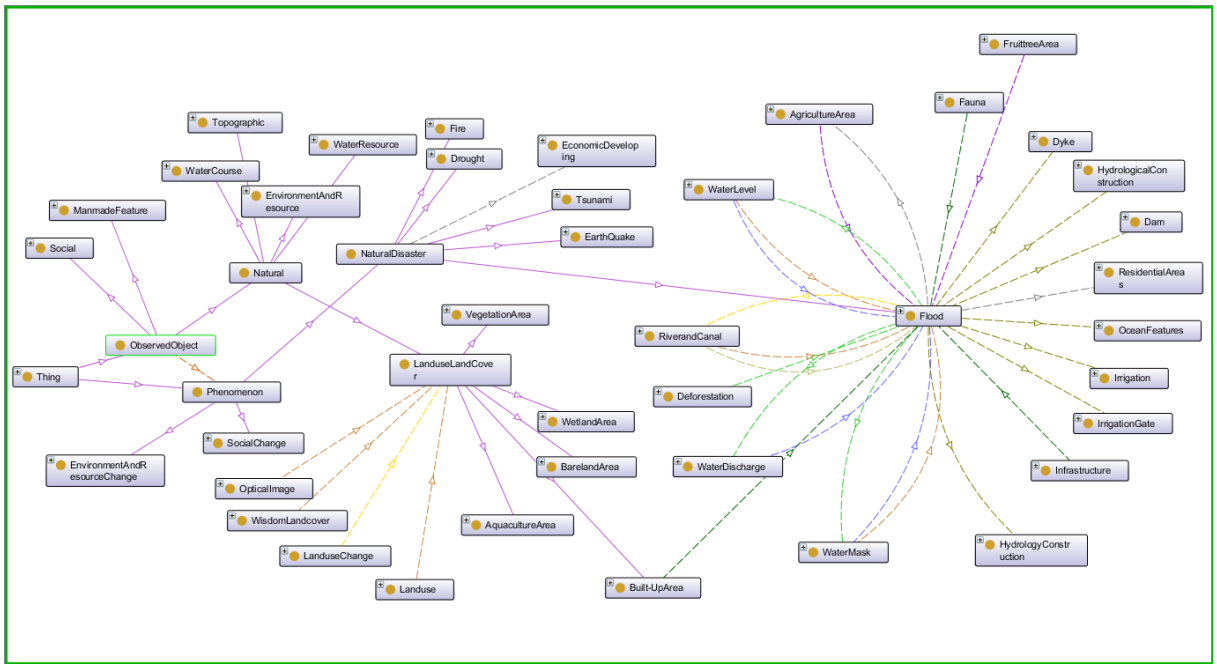

DataDomain (http://www.semanticweb.org/ontologies/2011/4/DataDomain.owl) - [D:\OntologyTest\test_property.rdf]
File Edit View Reasoner Tools Refactor Window Help
Active Ontology Entities Classes Object Properties Data Properties Individuals OWLViz DL Query OntoGraf
Class hierarchy: DistrictLevel
OntoGraf:
Search: contains Search Clear
WaterResource
Social
Phenomenon
ProductGroup
BaseData
Climate
Fauna
GovernanceAndAdm
Hydrology
Infrastructure
LandcoverAndLand
Agriculture
AminalProductior
Aquaculture
Ecology
Forestry
Landcover
Landuse
Soil
WisdomLandcove
NaturalDisasters
PopulationAndEcon
Provider
Address
Author
ContactPerson
Institution
Name
Representation
GeometricResolutio
PixelSize
HighResolutio
LowResolutior
MediumResol
SuperHighRes
VeryHighReso
SpatialRepresetatio
Sensor
SpatialReference
PoliticalDivision
CommuneLevel
Continent
Country
DistrictLevel
Address
Statistic
Observation
TabularBased
ContactPerson
VectorBased
SatelliteImage
Author
Photograph
Modelling
ScanMap
Institution
Provider
DataCollection
RasterBased
SuperHighResolu tion
Name
MediumResolutio n
Literature
TextBased
ThematicProduct
PixelSize
GeometricResolu tion
LowResolution
HighResolution
DistrictLevel
Reports
VeryHighResolut ion
Agriculture
Thing
WisdomLandcover
LandcoverAndLan duse
Landcover
Climate
Forestry
NaturaDisaster s
ProductGroup
Soil
Landuse
Fauna
Aquaculture
Ecology
Hydrology
AminalProductio n
PopulationAndEc onomy
Infrastructure
BaseData
EnvironmentAndR esource
GovernanceAndAd ministration
To use the reasoner click Reasoner->Start reasoner
Show Inferences

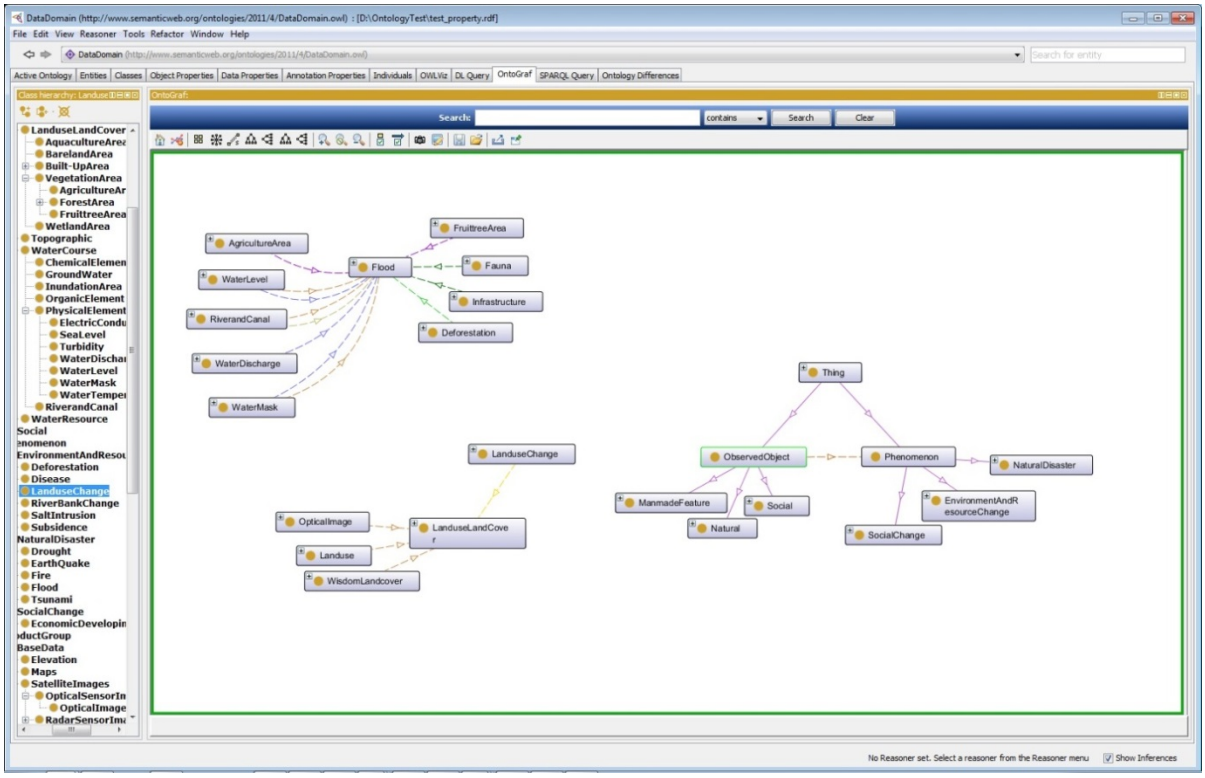

DataDomain (http://www.semanticweb.org/ontologies/2011/4/DataDomain.owl) - [D:\OntologyTest\test_property.rdf]
File Edit View Reasoner Tools Refactor Window Help
Active Ontology Entities Classes Object Properties Data Properties Individuals OWLViz DL Query OntoGraf
Class hierarchy: DistrictLevel
OntoGraf:
Search: contains Search Clear
DataCollection -- hasProvider(Subclass all) --> Provider
To use the reasoner click Reasoner->Start reasoner
Show Inferences

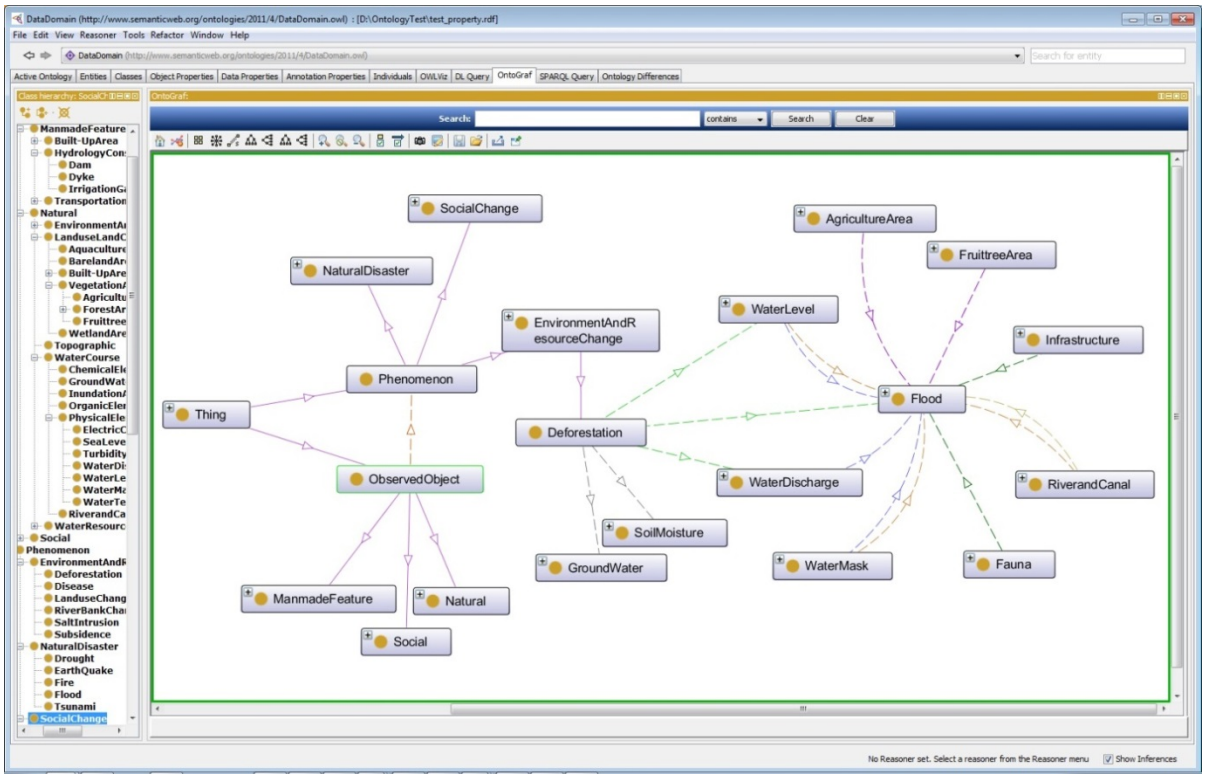




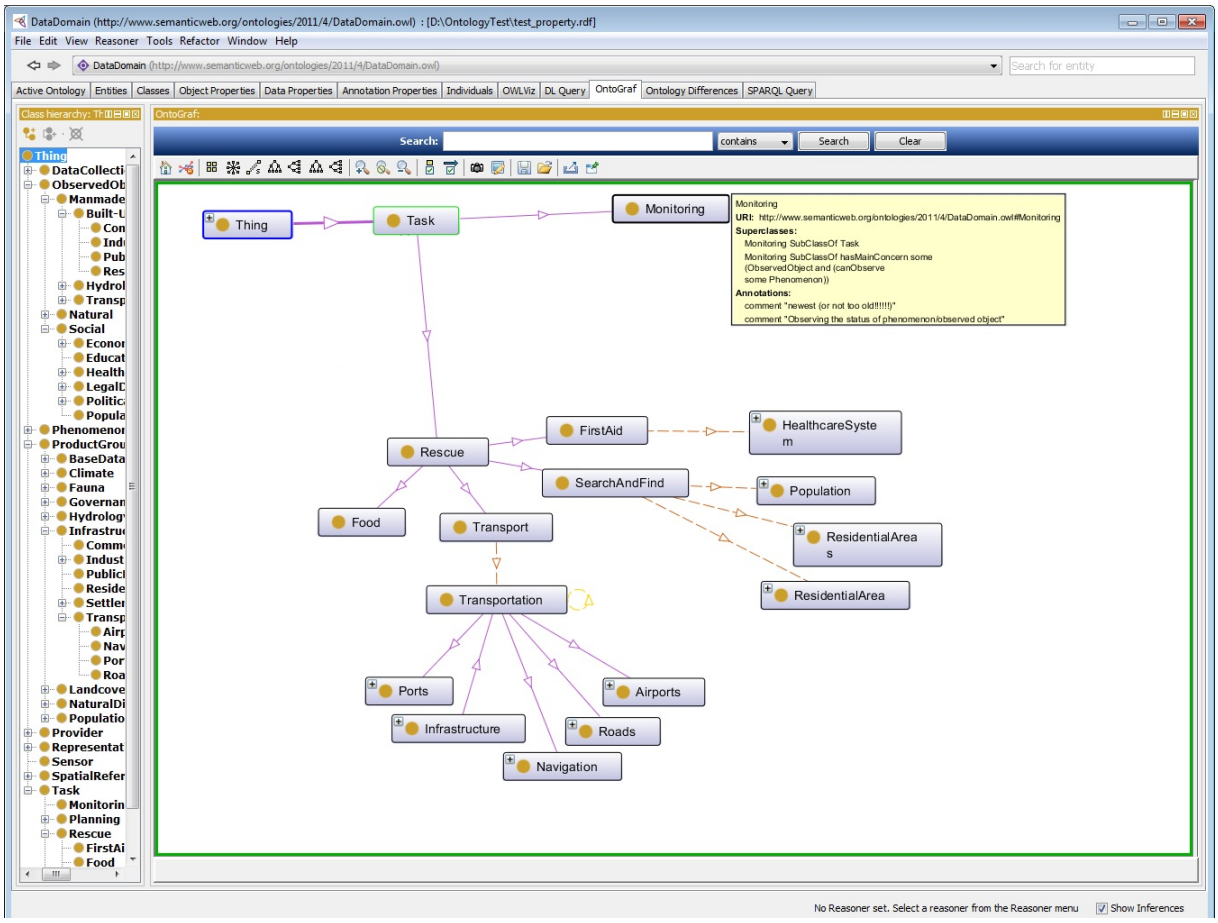
E. The Relationships and Properties in Observed Object Domain

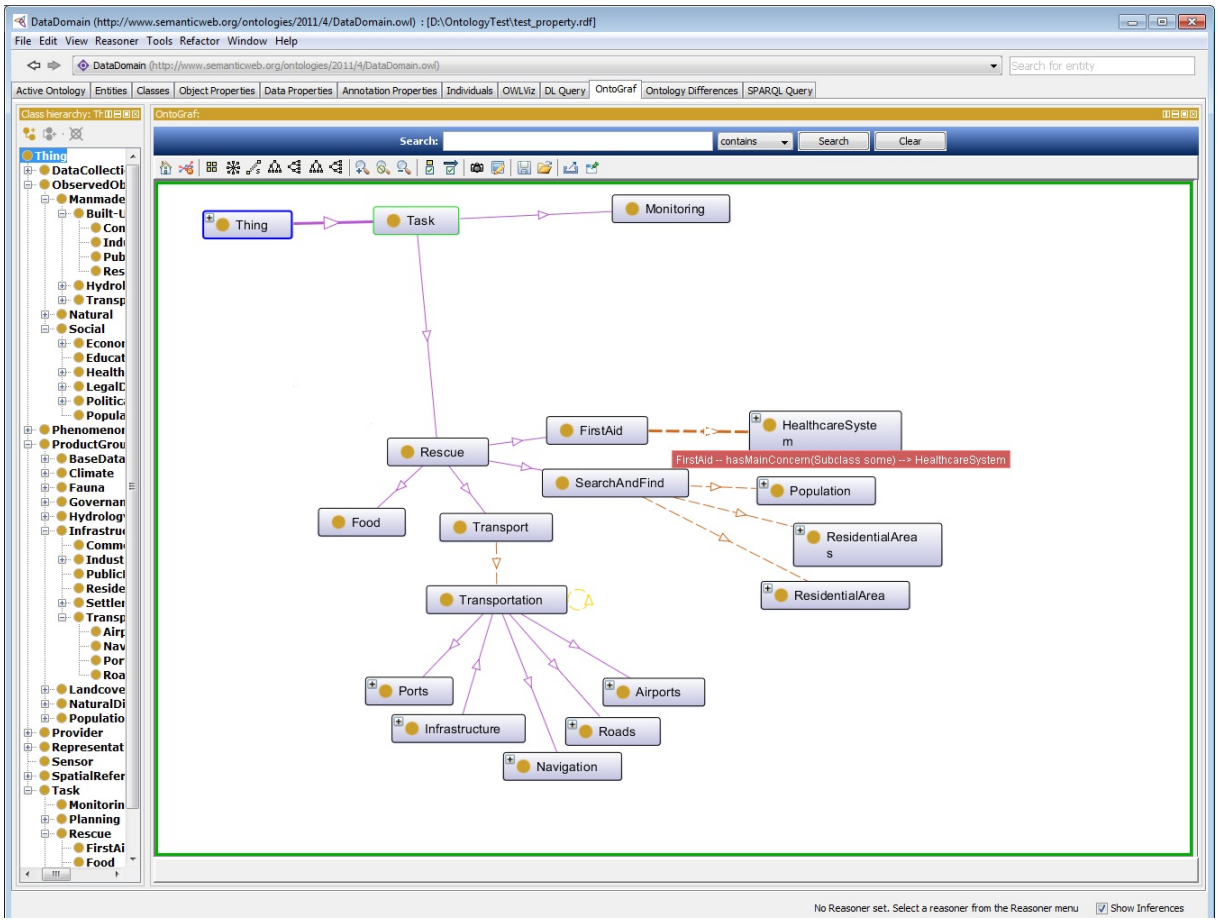

FruittreeArea
Fauna
AgricultureArea
Dyke
Topographic
WaterResource
Fire
EconomicDevelop
ing
HydrologicalCon
struction
WaterCourse
Drought
WaterLevel
EnvironmentAndR
esource
Tsunami
Dam
ManmadeFeature
EarthQuake
Social
NaturalDisaster
ResidentialArea
s
Natural
Flood
VegetationArea
OceanFeatures
RiverandCanal
ObservedObject
LanduseLandCove
r
Irrigation
Thing
Deforestation
Phenomenon
IrrigationGate
WetlandArea
SocialChange
Infrastructure
EnvironmentAndR
esourceChange
BarelandArea
WaterDischarge
OpticalImage
WisdomLandcover
AquacultureArea
WaterMask
HydrologyConstr
uction
LanduseChange
Landuse
Built-UpArea


DataDomain (http://www.semanticweb.org/ontologies/2011/4/DataDomain.owl) : [D:\OntologyTest\test_property.rdf]
File Edit View Reasoner Tools Refactor Window Help
Active Ontology Entities Classes Object Properties Data Properties Annotation Properties Individuals OWLViz DL Query OntoGraf SPARQL Query Ontology Differences
OntoGraf:
Search:
contains
Search
Clear
AgricultureArea
FruittreeArea
Flood
Fauna
WaterLevel
Infrastructure
RiverandCanal
Deforestation
WaterDischarge
WaterMask
LanduseChange
Thing
ObservedObject
Phenomenon
NaturalDisaster
ManmadeFeature
Social
EnvironmentAndR
esourceChange
Natural
SocialChange
OpticalImage
LanduseLandCove
r
Landuse
WisdomLandcover
No Reasoner set. Select a reasoner from the Reasoner menu
Show Inferences



F. The Relationships and Properties in Application Domain


DataDomain (http://www.semanticweb.org/ontologies/2011/4/DataDomain.owl) : [D:\OntologyTest\test_property.rdf]
File Edit View Reasoner Tools Refactor Window Help
Active Ontology Entities Classes Object Properties Data Properties Annotation Properties Individuals OWLViz DL Query OntoGraf Ontology Differences SPARQL Query
Class hierarchy:
OntoGraf:
Search: contains Search Clear
Thing
Task
Monitoring
Monitoring
URI: http://www.semanticweb.org/ontologies/2011/4/DataDomain.owl#Monitoring
Superclasses:
Monitoring SubClassOf Task
Monitoring SubClassOf hasMainConcern some (ObservedObject and (canObserve some Phenomenon))
Annotations:
comment "newest (or not too old!!!!!)"
comment "Observing the status of phenomenon/observed object"
FirstAid
HealthcareSyste m
Rescue
SearchAndFind
Population
Food
Transport
ResidentialArea s
ResidentialArea
Transportation
Ports
Airports
Infrastructure
Roads
Navigation
No Reasoner set. Select a reasoner from the Reasoner menu
Show Inferences



G. JAVA Code

Mapping RDB to RDF, and then merging to ontology file

```
private void prepare() {
        Model mappingModel = new ModelD2RQ(
                        "D:\\OntologyTest\\mapping_ranking.n3");
        Model model1 = ModelFactory.createDefaultModel();
        InputStream in1 = FileManager.get().open(
                        "D:\\OntologyTest\\test_property.rdf");
        model1.read(in1, null);
        model = mappingModel.union(model1);
        modelcheck = ModelFactory.createOntologyModel(
                        org.mindswap.pellet.jena.PelletReasonerFactory.THE_SPEC, model);
        // Write to file system
        try {
                FileOutputStream fout = new FileOutputStream(
                        "D:\\OntologyTest\\test_property_jena.rdf");
                modelcheck.write(fout);
        } catch (IOException e) {
        }
```

Query for Observed Object

```
private void executeQueryObservedObject() {
        System.out.print("\n");
        System.out.print(queryString);
        System.out.print("\n");
        table.revalidate();
        Vector rowData;
        DefaultTableModel m = (DefaultTableModel) table.getModel();

        table.revalidate();
        Query query = QueryFactory.create(queryString);

        // Execute the query and obtain results
        QueryExecution qe = QueryExecutionFactory.create(query, modelcheck);
        ResultSet results = qe.execSelect();

        int idDataset = 1;
        String property = "";
        while (results.hasNext()) {

                QuerySolution resultItem = results.nextSolution();
                String id = resultItem.getResource("data").getLocalName();
                String datasetname = resultItem.getLiteral("name").getString();
                String type = resultItem.getLiteral("type").getString();
                if (resultItem.getResource("property") != null) {
                        property = resultItem.getResource("property").getLocalName();
                } else {
                        property = "---";
                }

                //check duplicate row
                dup = "no";
                for (int ii = 0; ii < table.getModel().getRowCount(); ii++) {
                        String mm = table.getModel().getValueAt(ii, 4).toString().trim();
                        if (mm.equals(id)){
                                dup = "yes";
                        break;
                        }
                }

                if (dup.equals("no")) {
                String[] tempSD = resultItem.getLiteral("StartDate").getString().split("T");
                String[] tempED = resultItem.getLiteral("EndDate").getString().split("T");
                String datasetSD = tempSD[0];
                String datasetED = tempED[0];
                DateFormat testSD = new SimpleDateFormat("yyyy-MM-DD");
                Date theSD = null;
                Date theED = null;
                try {
                        theSD = testSD.parse(datasetSD);
                } catch (ParseException e) {
                        // TODO Auto-generated catch block
                        e.printStackTrace();
                }
                try {
                        theED = testSD.parse(datasetED);
```

```
                        } catch (ParseException e) {
                                // TODO Auto-generated catch block
                                e.printStackTrace();
                        }

                        //Query Date
                        Date querySD = null;
                        Date queryED = null;
                        try {
                                querySD = testSD.parse(ngaythangnamFrom1);
                        } catch (ParseException e) {
                                // TODO Auto-generated catch block
                                e.printStackTrace();
                        }
                        try {
                                queryED = testSD.parse(ngaythangnamTo1);
                        } catch (ParseException e) {
                                // TODO Auto-generated catch block
                                e.printStackTrace();
                        }
                        //SD, ED to UNIX
                        long theSDUNIX = theSD.getTime();
                        long theEDUNIX = theED.getTime();
                        //Query date to UNIX
                        long startTimeQueryUNIX = querySD.getTime();
                        long endTimeQueryUNIX = queryED.getTime();

                        //Compare SD, ED by Query Date
                        //(5)
                        if ((theSDUNIX <= startTimeQueryUNIX) & (endTimeQueryUNIX <=
theEDUNIX)) {
                                T = 1;
                        }
                        //(4)
                        if ((startTimeQueryUNIX <= theSDUNIX) & (theEDUNIX <=
endTimeQueryUNIX) & (theSDUNIX == theEDUNIX)) {
                                T = 1;
                        }
                        //(3)
                        if ((startTimeQueryUNIX <= theSDUNIX) & (endTimeQueryUNIX <=
theEDUNIX) & (theSDUNIX < endTimeQueryUNIX)) {
                                T = (endTimeQueryUNIX - theSDUNIX)/(endTimeQueryUNIX -
startTimeQueryUNIX);
                        }
                        //(2)
                        if ((theSDUNIX <= startTimeQueryUNIX) & (theEDUNIX <=
endTimeQueryUNIX) & (startTimeQueryUNIX < theEDUNIX)) {
                                T = (theEDUNIX - startTimeQueryUNIX)/(endTimeQueryUNIX -
startTimeQueryUNIX);
                        }
                        //(1)
                        if ((startTimeQueryUNIX <= theSDUNIX) & (endTimeQueryUNIX <=
theEDUNIX) & (theSDUNIX < theEDUNIX)) {
                                T = (theEDUNIX - theSDUNIX)/(endTimeQueryUNIX -
startTimeQueryUNIX);
                        }
                        weight = (Wca * CA) + (Wt * T) + (Wth * Th);
                        rowData = new Vector();
```

```
                rowData.add(idDataset);
                rowData.add(new Boolean(false));
                rowData.add(new Boolean(false));
                rowData.add(new Boolean(false));
                rowData.add(id);
                rowData.add(datasetname);
                rowData.add(type);
                rowData.add(property);
                rowData.add(weight);
                m.addRow(rowData);
                idDataset = idDataset + 1;
                }
        }
        qe.close();
}
```

H. SPARQL

A SPARQL query comprises, in order (adopted from (Lin 2011)):

- Prefix declarations: Abbreviating for URIs
- Dataset definition: Defining which RDF graph(s) are being queried
- A result clause: Identifying what information will return from the query
- The query pattern: Specifying what will be queried in the underlying dataset
- Query modifiers: slicing, ordering, and otherwise rearranging query results, e.g. ORDER BY, LIMIT etc.

An example is shown below

 # prefix declarations

PREFIX foo: <http://example.com/resources/>

...

 # dataset definition

FROM ...

 # result clause

SELECT ...

 # query pattern

```
WHERE {
...
}
# query modifiers
ORDER BY ...
```

References

52North. (2012). "Sensor Observation Service." November 2012, from http://52north.org/communities/sensorweb/sos/index.html.

Abolhassani, H., B. B. Hariri and S. H. Haeri (2006). On Ontology Alignment Experiments. Webology. **3**.

Amirian, P. and A. A. Alesheikh (2008). "Publishing Geospatial Data through Geospatial Web Service and XML Database System." American Journal of Applied Sciences **5**(10): 1358-1368.

Antoniou, G. and F. v. Harmelen (2008). "A Semantic Web Primer." MIT Press, USA, 2004. ISBN: 0-262-01210-3

Arens, Y., C.-N. Hsu and C. A. Knoblock (1996). "Query Processing in the SIMS Information Mediator." In The AAAI Press, May 1996.

Athanasis, N., K. Kalabokidis, M. Vaitis and N. Soulakellis (2009). "Towards a semantics-based approach in the development of geographic portals." Computers&Geosciences **35**: 301-308.

Bacharach, S. (2008). "About the Open Geospatial Consortium, Inc (OGC)." geoNews, the rmDATA newsletter.

Becker, S., V. Walter and D. Fritsch (2012). Integrated Management Of Heterogeneous Geodata With A Hybrid 3d Geoinformation System. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,. Melbourne, Australia. **I-2**.

Bernard, L., U. Einspanier, S. Haubrock, S. Hübner, E. Klien, W. Kuhn and R. Lessing (2004). "Ontology-Based Discovery and Retrieval of Geographic Information in Spatial Data Infrastructures." Geotechnologien Science Report **4**.

Bernard, L., U. Einspanier, S. Haubrock, S. Hübner, W. Kuhn, R. Lessing, M. Lutz and U. Visser (2003). "Ontologies for Intelligent Search and Semantic Translation in Spatial Data Infrastructures." Photogrammetrie - Fernerkundung - Geoinformation **2003**(6): 451-462.

Bernard, L., I. Kanellopoulos, A. Annoni and P. Smits (2005). "The European geoportal—one step towards the establishment of a European Spatial Data Infrastructure." Computers, Environment and Urban Systems **29**(1): 15-31.

Bernstein, A. and M. Klein (2002). "Towards High-Precision Service Retrieval." In: Horrocks I., Hendler J. (eds.) The Semantic Web-First International Semantic Web Conference (ISWC 2002) **84-101**.

Bishr, Y. (1998). "Overcoming the semantic and other barriers to GIS interoperability." International Journal of Geographical Information Science **12**(4): 299-314.

Buccella, A., A. Cechich and P. Fillottrani (2009). "Ontology-driven geographic information integration: A survey of current approaches." Computers & Geosciences **35**: 710-723.

Buckland, M. and F. Gey (1994). "The Relationship between Recall and Precision." Journal of the American Society for Information Science and Technology **45**(1): 12-19.

Buckley, C. and E. M. Voorhees (2000). "Evaluating Evaluation Measure Stability."

Cai, G. (2002). "GeoVSM: An Integrated Retrieval Model for Gerographical Information." In Egenhofer M J and Mark D M (eds) Proceedings of the Second International

Geographic Information Science Conference. Berlin, Springer Lecture Notes in Computer Science **No 2478**: 65–79.

Campos, L. F. d. B. (2007). "Increase of Precision on the Top of the List of Retrieved Web Documents Using Global and Local Link Analysis." Webology **4**(3).

Chalupsky, H. (2000). "OntoMorph: A Translation System for Symbolic Knowledge." In: Proc. 17th Intl. Conf. on Principles of Knowledge Representation and Reasoning KR'2000 Colorado,USA: (April 2000) 471--482.

Choi, N., I.-Y. Song and H. Han (2006). "A Survey on Ontology Mapping." ACM SIGMOD Record **35**(3).

Cleverdon, C. W. (1972). "On the Inverse Relationship of Recall and Precision." Journal of Documentation **28**(3): 195-201.

Coppock, J. T. and D. W. Rhind (1991). The History of GIS. D. T. Maguire, M. F. Goodchild and D. W. Rhind. **1:** 291-300.

Cruz, I. F. and H. Xiao (2003). "Using a Layered Approach for Interoperability on the SemanticWeb." In Proceedings of the 4th International Conference on Web Information Systems Engineering (WISE 2003). Rome, Italy, December 2003: 221-232.

Cruz, I. F. and H. Xiao (2005). "The Role of Ontologies in Data Integration." Jounal of Engineering Intelligent Systems **13**(4).

Curé, O. (2005). "Mapping Databases To Ontologies To Design And Maintain Data In A Semantic Web Environment." Computer and Information Science **4**(4): 52-57.

D2RQ. (2012). "The D2RQ Mapping Language." Retrieved May 5th, 2012, from http://d2rq.org/d2rq-language.

Dentler, K., R. Cornet, A. t. Teije and N. d. Keizer (2011). "Comparison of Reasoners for large Ontologies in the OWL2 EL Profile." Semantic Web **2**(2): 71-87.

Durbha, S. S., R. L. King, V. P. Shah and N. H. Younan (2009). "A Fraework for Semantic Reconciliation of Disparate Earth Observation Thematic Data." Computers & Geosciences **35**(4): 761-773.

Eclipse. (2012). "Eclipse." Retrieved July, 2012, from http://www.eclipse.org/org/.

ESRI (2010). "Spatial Data Infrastructure (SDI)."

Fahad, M., M. A. Qadir and S. A. Hussain (2008). "Evaluation of Ontologies and DL Reasoners." Intelligent Information Processing IV, Volume 288/2008, Springer, 1571-5736 (Print) 1861-2288 (Online): 17-27.

FAO. (2012). "AGROVOC Multilingual agricultural thesaurus." Retrieved July, 2012, from http://aims.fao.org/standards/agrovoc/about.

Fensel, D. (2001). Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Berlin, Springer-Verlag.

Fonseca, F. T. and M. J. Egenhofer (1999). "Ontology-Driven Geographic Information Systems." 7th ACM Symposium on Advances in Geographic Information Systems Kansas City, MO.

Fonseca, F. T., M. J. Egenhofer, P. Agouris and G. Camara (2002). "Using Ontologies for Integrated Geographic Information Systems." Transactions in GIS **6**(3): 231-257.

Freitas, R. A. P. and J. C. Ramalho (2011). "Using Ontologies to Abstract Relational Databases Conceptual Model." EPIA'2011.

Friis-Christensen, A., S. Schade and S. Peedell (2005). "Approaches to solve schema heterogeneity at the European level." Proceedings of the 11th EC-GIS Workshop (June 2005).

Gebhardt, S., T. Wehrmann, V. Klinger, I. Schettler, J. Huth, C. Künzer and S. Dech (2010a). "Improving Data Management and Dissemination in Web based Information System by Semantic Enrichment of Description Data Aspects." Computers & Geosciences **36**: 1362-1373.

Gebhardt, S., T. Wehrmann, I. Schettler, J. Huth, C. Künzer, M. Schmidt and S. Dech (2010b). "A Water-related Information System for the Mekong Delta: Data modelling and data management." Computers &Geosciences.

GeoNames. (2012). "GeoNames." Retrieved 12th June, 2012, from http://www.geonames.org/.

GIS.com. (2012). "What is GIS?" Retrieved 13th June, 2012, from http://www.gis.com/content/what-gis.

Giunchiglia, F., P. Shvaiko and M. Yatkevich (2007). "Semantic Matching: Algorithms and Implementation." Journal on Data Semantics **1**.

Global Water Partnership (2000). http://www.gwp.org/.

Goodchild, M., M. Egenhofer, R. Fegeas and C. Kottman (1999). "Interoperating Geographic Information Systems." The International Series in Engineering and Computer Science **Vol. 495**.

Gorte, R. W. and P. A. Sheikh (2010). "Deforestation and Climate Change." Congreeional Research Service.

Goutte, C. and E. Gaussier (2005). A Probablilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation. Proceedings of the European Colloquium on IR Resarch (ECIR'05), Springer.

Gruber, T. (2008). "Ontology." in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag. Retrieved from http://tomgruber.org/writing/ontology-definition-2007.htm.

Gruber, T. R. (1995). "Toward Principles for the Design of Ontologies Used for Knowledge Sharing." International Journal Human-Computer Studies **43**: 907-928.

Guarino, N. (1998). "Formal Ontology and Information Systems." In: Formal Ontologies in Information Systems, N. Guarino (Ed.), IOS Press: 3-15.

Gwenzi, J. (2010). Enhancing Spatial Web Search with Semantic Web Technology and Metadata Visualisation. Master, Internation Institute for Geo-Information Science and Earth Observation (ITC), Enschede, Netherlands.

Hall, M. M. and P. Mandl (2006). "Spatially Extended Ontologies for a Semantic Model of Harmonised Landuse and Landcover Information."

Han, J. and M. Kamber (2006). Data Mining: Concepts and Techniques, second edition. San Francisco, CA, Morgan Kaufmann.

Heine, M. H. (1973). "The Inverse Relationship of Precision and Recall in terms of the Swets Model." Journal of Documentation **29**: 181-198.

Hert, M., G. Reif and H. C. Gall (2011). "A Comparison of RDB-to-RDF Mapping Languages." Proceedings of the 7th International Conference on Semantic Systems.

Hochmair, H. H. (2005). "Ontology Matching for Spatial Data Retrieval from Internet Portals." In Rodríguez M A, Cruz I F, Egenhofer M J, and Levashkin S (eds) Proceedings of the First International Geospatial Semantics Conference. Berlin, Springer Lecture Notes in Computer Science **No 3799: 166–82**.

Horridge, M. (2011a). A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools Edition 1.3, The University Of Manchester.

Horridge, M. (2011b). A Practical Guide to Building OWL Ontologies Using Protégé 4 and CO-Ode Tools Edition 1.3, The University Of Manchester.

Hristov, T. and V. Ioncheva (2006). "The Geographical Water Resources Information and Assessment System – an Element of the IWRM." International conference on Water Observation and Information System for Decision Support – BALWOIS, Ohrid, Republic of Macedonia.

Ikeda, M., K. Seta, O. Kakusho and R. Mizuguchi (2009). "Task ontology: Ontology for building conceptual problem solving models." Proceedings of ECAI98 Workshop on Applications of ontologies and problem-solving model 126-133. .

Ikeda, M., K. Seta and R. Mizoguchi (1997). "Task Ontology Makes It Easier To Use Authoring Tools." Proceedings of the 15th international joint conference on Artifical intelligence, August 23-29, 1997, Nagoya, Japan 342-347.

ISO (2003). "ISO 19115:2003."

ISO (2009). Standards Guide: ISO/TC 211 Geographic Information / Geomatics.

ISO. (2012). "About ISO." November 2012, from http://www.iso.org/iso/home/about.htm.

Java. (2012). "Java." Retrieved July, 2012, from http://www.java.com/en/download/faq/whatis_java.xml.

Jones, C. B., A. I. Abdelmoty, D. Finch, G. Fu and S. Vaid (2004). "The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing." GIScience 2004: Third International Conference on Geographic Information Science **3234**: 125-139.

Jones, K. S. (1981). Information Retrieval Experiment, Butterworth-Heinemann Newton, MA, USA.

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

Khaled, R., L. M. Tayeb and S. Servigne (2010). Geospatila Web Service Semantic Discovery Approach Using Quality. JCIT.

Klien, E., U. Einspanier, M. Lutz and S. Hübner (2004). "An Architecture for Ontology-Based Discovery and Retrieval of Geographic Information." In the proceedings of AGILE Int. Conf. on Geographic Information Science.

Klien, E., M. Lutz and W. Kuhn (2006). "Ontology-based discovery of geographic information services—An application in disaster management." Computers, Environment and Urban Systems **30**: 102-123.

Klinger, V., T. Wehrmann, S. Gebhardt and C. Künzer (2010). A Water-Related Web-based Information System for the Sustainable Development of the Mekong Delta. The mekong Delta System. Interdisciplinary Analyses of a River Delta. C. Künzer and F. G. Renaud. Springer Netherlands, Springer Environmental Science and Engineering. **XV:** 423-444.

KnowledgeWeb (2005). "D2.2.1 Specification of a Common Framwork for Characterizing Alignment." Project Deliverable. KWEB/2004/D2.2.1/v2.0. KnowledgeWeb Consortium (IST Project IST-2004-507482).

Laclavík, M. (2006). "RDB2Onto: Relational Database Data to Ontology Individuals Mapping." Tools for Acquisition, Organisation and Presenting of Information and Knowledge: 86-99.

Le, A. T. (2010). Impacts of climate change and sea level rise to the integrated agriculture-aquaculture system in the Mekong River Basin - A case study in the Lower Mekong River Delta in Vietnam. The International Workshop on the "Climate Change Responses for Asia International Rivers: Opportunities and Challenges". China, 26-28 February, 2010.

Lenzerini, M. (2002). "DataIntegration: A Theoretical Perspective." Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA. ACM 2002.

Li, W., C. Yang and R. Raskin (2008). "A Semantic Enhanced Model for Searching in Spatial Web Portals." In: AAAI'08 Workshop on Scientific Semantic Knowledge Integration, Palo Alto, CA.

Lin, K. (2011). Data Discovery and Federation. The Cyberinfrastructure Summer Institute for Geoscientist (CSIG'11).

Lutz, M., J.Sprado, E.Klien, C.Schubert and I.Christ (2009). "Overcoming semantic heterogeneity in spatial data infrastructures." Computers &Geosciences **35**: 739-752.

Lutz, M. and E. Klien (2006). "Ontology-based retrieval of geographic information." International Journal of Geographic Information **20**(3): 233-260.

Madhavan, J., P. A. Bernstein, P. Domingos and A. Y.Halevy (2002). "Representing and Reasoning about Mappings between Domain Models." 18th National Conference on Artificial Models (AAAI 2002).

Malik, R. (2006). CONAN: Text Mining in the Biomedical Domain. PhD, Utrecht University.

Mavratza, O., D. Sarafidis and I. Paraschakis (2007). Design of an ISO 19115 compliant profile for documenting spatial datasets and series for the Hellenic Cadastre. Proceedings of the Intenational Conference "Spatial Data Quality" (ISSDQ 2007), ITC, Enschede, the Netherlands.

Mena, E., V. Kashyap, A. Illarramendi and A. Sheth (1998). Domain Specific Ontologies for Semantic Information Brokering on the Global Information Infrastructure. Proceedings, International Conference on Formal Ontology in Information Systems (FOIS '98), Trento.

Mena, E., V. Kashyap, A. P. Sheth and A. Illarramendi (1996). "OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies." In Proceedings of the 1st IFCIS International Conference on Cooperative Information Systems (CoopIS 1996): 14–25.

MRC. (2012). "Mekong River Commission for Sustainable Development." Retrieved 02 July, 2012, from http://www.mrcmekong.org/

MSDN. (2012). "What is SPARQL." last access March, 2012, from http://msdn.microsoft.com/en-us/library/aa303673.aspx.

Musen, M. A. (1992). "Dimension of Knowledge Sharing and Reuse." Computer and Biomedical Research **25**(5): 435-467.

Navarrete, T. (2006). "Semantic integration of thematic geographic information in a multimedia context." PhD Thesis, Doctorate in Computer Science and Communication Department of Technology, Universitat Pompeu Fabra, Barcelona.

Nigro, H. O., S. E. G. Cisaro and D. H. Xodo (2008). Data mining with ontologies: Implementations, Findings and Frameworks. Hershey, New York, Information Science Reference.

NISO (2004). Understanding Metadata, National Information Standards Organization Press.

NIST (2001). Common Evaluation Measures. National Institute of Standards and Technology (NIST) Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001) Gaithersburg, Maryland, USA.

Noy, N. F. (2009). "Ontology mapping." In Staab, Steffen and Studer, Rudi, eds., Handbook on ontologies, 2 nd ed. International handbooks on information systems. Berlin: Springer: 573-590.

Noy, N. F. and D. L. McGuinness (2001). "Ontology Development 101: A Guide to Creating Your First Ontology." Stanford Knowledge Systems Laboratory Technical Report

KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, Retrieved from http://protege.stanford.edu/publications/ontology_development/ontology101.pdf.
OGC. (2004). "OpenGIS® Web Map Server Cookbook." November 2012, from http://www.opengeospatial.org/standards/wms.
OGC. (2005). "The Open Geospatial Consortium's Web Map Service (WMS) Approved as International Organization for Standardization (ISO) Standard." November 2012, from http://www.opengeospatial.org/node/436.
OGC. (2008). "OpenGIS Location Services (OpenLS): Core Services." November 2012, from http://www.opengeospatial.org/standards/ols.
OGC. (2010a). "OGC® WCS 2.0 Interface Standard - Core." November 2012, from http://www.opengeospatial.org/standards/wcs.
OGC. (2010b). "OpenGIS Web Feature Service 2.0 Interface Standard." November 2012, from http://www.opengeospatial.org/standards/wfs.
OGC. (2012a). "GeoServices Rest SWG " Retrieved 15th July, 2012, from http://www.opengeospatial.org/projects/groups/gservrestswg.
OGC. (2012b). "Glossary of Terms - C." November 2012, from http://www.opengeospatial.org/ogc/glossary/c.
OGC. (2012c). "OGC® Sensor Observation Service Interface Standard." November 2012, from http://www.opengeospatial.org/standards/sos.
OGC. (2012d). "OGC® Standards and Supporting Documents " Retrieved 28 May, 2012, from http://www.opengeospatial.org/standards.
OGC. (2013a). "Geographic Features." Retrieved May, 2013, from http://ormdev.opengeospatial.org/node/94.
OGC. (2013b). "Spatiotemporal Geometry and Topology." Retrieved May, 2013, from http://ormdev.opengeospatial.org/node/95.
Orchestra. (2007). "The Value of Standards." November 2012, from http://www.eu-orchestra.org/TUs/Standards/en/text/Standards.pdf.
Ostensen, O. M. and P. C. Smits (2002). "ISO/TC211: Standardisation of geographic information and geo-informatics." Geoscience and Remote Sensing Symposium, IGARSS '02 **IEEE International. v1**: 261-263.
Palmer, S. B. (2001). "The Semantic Web: An Introduction." Retrieved 30 May, 2012, from http://infomesh.net/2001/swintro/#whatIsSw.
Pan, F. and J. R. Hobbs (2005). "Time in OWL-S." In Proceedings of the AAAI Spring Symposium on Semantic Web Services, Stanford Universizy. CA, AAAI Press: 29-36.
Pan, Z. (2005). "Benchmarking DL Reasoners Using Realistic Ontologies." In Proc. of the OWL: Experiences and Directions Workshop, 2005.
Paul, M. and S. K. Ghosh (2006). "An Approach for Service Oriented Discovery and Retrieval of Spatial Data." In Proceedings of IWSOSE'06: 88-94.
Podwyszynski, M. (2009). "Knowledge-based search for Earth Observation products." Diploma Thesis, University of Passau.
Powers, S. (2003). "Practical RDF." O'Reilly 2003.
Protégé (2012). "Protégé." http://protege.stanford.edu/.
Raghavan, V., P. Bollmann and G. S. Jung (1989). "A critical investigation of recall and precision as measures of retrieval system performance." ACM Transactions on Information Systems **7**(3): 205-229.

Ramanujam, S., A. Gupta, L. Khan, S. Seida and B. Thuraisingham (2009). "R2D: A Bridge between the Semantic Web and Relational Visualization Tools." In International Conference on Semantic Computing, Berkeley, California: 303-311.

Raskin, R. (2005). "Guide to SWEET Ontologies." Retrieved 13th June, 2012, from http://sweet.jpl.nasa.gov/guide.doc.

Raskin, R. G. and M. J. Pan (2005). "Knowledge Representation in the Semantic Web for Earth and Environment Terminology (SWEET)." Computers & Geosciences **31**: 1119-1125.

Reitz, T. (2008). "Geospatial Interoperability Issues." BOSS4GMES Workshop on Interoperability, Toulouse, France.

Ren, Y., J. Lemcke, T. Rahmani, A. Friesen, S. Zivkovic, B. Gregorcic, A. Bartho, Y. Zhao and J. Z. Pan (2010). "Task Representation and Retrieval in an Ontology-Guided Modelling System." CEUR Workshop Proceedings **529**.

Riedemann, C. and C. Timm (2003). "Service for Data Integration." Data Science Journal (Spatial Data Usability Special Section) **2**.

Rijsbergen, C. J. v. (1979). Information Retrieveal. Online book, Butterworth-Heinemann.

Roset, R., M. Lurgi, M. Croitoru, B. Hu, M. L. i. Ariet and P. Lewis (2008). "A Visual Mapping Tool for Database Interoperability: the HealthAgents case." In: Proceeding of the 3rd CS-TIW Workshop.

Sahoo, S. S., W. Halb, S. Hellmann, K. Idehen, T. T. Jr, S. Auer, J. Sequeda and A. Ezzat (2009). "A Survey of Current Approaches for Mapping of Relational databases to RDF." W3C RDB2RDF Incubator Group.

Schön, D. (1983). The Reflective Practitioner: How Professionals Think in Action, Basic Books, New York.

Schramm, D. and R. Dries (1986). Natural Hazards: Causes and Effects - Study Guide for Disaster Management, Disaster Management Center - University of Wisconsin-Madison.

SiliconPress. (2002). "HTTP." November 2012, from http://www.silicon-press.com/briefs/brief.http/brief.pdf.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur and Y. Katz (2007). "Pellet: A Practical OWL-DL Reasoner." Journal of Web Semantics **5**(2): 51-53.

Sriphaisal, W. and A. K.Pujari (2006). "A Comparative Assessment of Giservices Architectures." Map Asia.

Stopper, R., I. I. Enescu, S. Wiesmann and O. Schnabel. (2011). "Open Geospatial Consortium (OGC) and Web Services (WMS, WFS)." November 2012, from http://www.elml.uzh.ch/preview/cartouche/webservice/en/html/index.html.

Stuckenschmidt, H. (2003). "Ontology-Based Information Sharing in Weakly Structured Environments." PhD Thesis, Vrije Universiteit Amsterdam, Amsterdam.

Suter, G. W. (1993). Ecological Risk Assessment. Chelsea, MI, Lewis Publishers**:** 505.

Swann, J. (2010). "A Dialogic Approach to Online Facilitation." Australasian Journal of Education Technology **26**(1): 50-62.

SWEET. (2012). "Semantic Web for Earth and Environmental Terminology." Retrieved 14th June, 2012, from http://sweet.jpl.nasa.gov/ontology/.

Timpf, S. (2002). "The need for task ontologies in interoperable GIS." In University of Zürich, Department of Geography, Retrieved from http://e-collection.library.ethz.ch/view/eth:25486.

Tobler, W. R. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." Economic Geography **46**: 234-240.

Tran, T. B., W. Thilo, G. Steffen, Verena, Klinger, H. Juliane, Q. T. Vo and K. Claudia (2010). "Ontology Based Approach for Geospatial Semantic Web." In: Proceedings of the 31st Asian Remote Sensing Conference, . 31st Asian Remote Sensing Conference.

Tran, V. X. and H. Tsuji (2007). "OWL-T: A Task Ontology Language for Automatic Service Composition." In: IEEE International Conference on Web Services. IEEE, Los Alamitos.

Tsarkov, D. and I. Horrocks (2006). "FaCT++ Description Logic Reasoner: System Description." Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006) **4130**: 292–297.

Vaccari, L., P. Shvaiko and M. Marchese (2009). "A geo-service semantic integration in Spatial Data Infrastructures." International Journal of Spatial Data Infrastructures Research **4**: 24-51.

Visser, P. R. S., D. M. Jones, T. J. M. bench-Capon and M. J. R. Shave (1997). "An Analysis of Ontology Mismatches; Heterogeneity versus Interoperability." Working notes of the AAAI 1997 Spring Symposium on Ontological Engineering, Stanford University, California, USA.

Visser, U. and C. Schlieder (2002). Modeling Real Estate Transactions: the potential Role of Ontologies. in The Ontology and Modelling of Real Estate Transactions. H. Stuckenschmidt, E. Stubjkaer and C. Schlieder, Ashgate**:** 99-113.

Vöegle, T., S. Huebner and G. Schuster (2003). "BUSTER - An Information Broker for the Semantic Web." Künstliche Intelligenz **3**: 31-34.

W3C "Extensible Markup Language (XML)." http://www.w3.org/XML/.

W3C. (2000). "A Little History of the World Wide Web." Retrieved 16th July, 2012, from http://www.w3.org/History.html.

W3C. (2006). "Time Ontology in OWL." Retrieved 24th July, 2012, from http://www.w3.org/TR/owl-time/.

W3C (2009). "OWL Web Ontology Language Guide." http://www.w3.org/TR/2004/REC-owl-guide-20040210/#BasicDefinitions.

W3C. (2010a). "Semantic Web Activity, http://www.w3.org/2001/sw/." Retrieved 15th August, 2010.

W3C (2010b). "W3C, Resource Description Framework (RDF), http://www.w3.org/RDF/ , last accessed September 10th 2010."

W3C. (2011). "Notation3 (N3): A readable RDF syntax." Retrieved 19th June, 2012, from http://www.w3.org/TeamSubmission/n3/#intro.

W3C. (2012). "OWL Web Ontology Language." Retrieved 12th June, 2012, from http://www.w3.org/TR/owl-features/.

Wache, H., T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hübner (2001). "Ontology-Based Integration of Information - A Survey of Existing Approaches." In Proceedings of the IJCAI-01Workshop on Ontologies and Information Sharing,.

Washington, N. and S. Lewis (2008). "Ontologies: Scientific Data Sharing Made Easy." Nature Education **1**(3).

Webber, W. E. (2010). Measurement in Information Retrieval Evaluation. PhD, The University of Melbourne.

Welie, M. v. (2001). "Task-Based User Interface Design." PhD Thesis, Vrije Universiteit Amsterdam, Amsterdam.

Wiegand, N. and C. Garcia (2007). "A Task-Based Ontology Approach to Automate Geospatial Data Retrieval." Transactions in GIS **11**(3): 355-376.

WISDOM. (2011). "Water-Related Information System for the Sustainable development of the Mekong Delta." Retrieved 25, 2011, from http://wisdom.eoc.dlr.de/en/content/objectives-wisdom-project.
WISDOM. (2012). "WISDOM Information System." Retrieved 17th July, 2012, from http://wisdom.eoc.dlr.de/Elvis/.
WorldWideWebSize. (2012). "The size of the World Wide Web (The Internet)." Retrieved 07 May, 2012, from http://www.worldwidewebsize.com.
Wu, Z., H. Chen, H. Wang, Y. Wang, Y. Mao, J. Tang and C. Zhou (2006). "Dartgrid: a Semantic Web Toolkit for Integrating Heterogeneous Relational Databases."
Xiao, H. (2006). Query Processing for Heterogeneous Data Integration Using Ontologies. PhD Thesis, University of Illinois at Chicago, USA.
Xie, H. (2005). User Model Driven Architecture for Information Retrievel in Construction Project Management. PhD, University of Florida.
Yan, X., Y. Peng, J. Meng, J. Ruzante, P. M. Fratamico, L. Huang, V. Juneja and D. S. Needleman (2011). "From Ontology Selection and Semantic Web to an Integrated Information System for Food-borne Diseases and Food Safety." Advances in experimental medicine and biology **696**: 741-750.
Yanfeng, S., Z. Jack Fan and Z. Xiaofang (2006). A Grid-Enabled Architecture for Geospatial Data Sharing. Services Computing, 2006. APSCC '06. IEEE Asia-Pacific Conference on.
Yuan, M. (1997). "Development of a Global Conceptual Schema for Interoperable Geographic Information." Paper presented in INTEROP'97 Conference, Santa Barbara, December 1997. http://www.ncgia.ucsb.edu/conf/interop97/program/papers/yuan/yuan.html.
Zhan, Q., X. Zhang and D. Li (2008). "Ontology-Based Semantic Description Model for Discovery and Retrieval of Geo Spatial Information." Computer and Information Science **XXXVII**: 2-7.
Zhang, C. and W. Li (2005). "The Roles of Web Feature and Web Map Services in Real-time Geospatial Data Sharing for Time-critical Applications." Cartography and Geographic Information Science **32**(4): 269-283.
Zhao, J.-s., X. Li, Y. Zhao, T. Xu and X. Fu (2005). "Methods and Implementation of the Geospatial Databases Integration and Update towards E-Government." ISPRS Workshop on Service and Application of Spatial Data Infrastructure, XXXVI(4/W6),Oct.14-16,Hangzhou, China.
Zhao, P., G. Yu and L. Di (2006). "Geospatial Web Service." Emerging Spatial Information System and Application.
Zhu, M. (2004). "Recall, Precision and Average Precision." Working paper 2004-09, Department of Statistics & Actuarial Science, University of Waterloo, Octorber 2012, from http://sas.uwaterloo.ca/stats_navigation/techreports/04WorkingPapers/2004-09.pdf.