Masters
Program
in **Geospatial Technologies**

Microtheories for SDI - Accounting for Diversity of Local Conceptualizations at a Global Level

Stephanie Duce

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

UNIVERSIDADE NOVA ISEGI
UNIVERSITAT JAUME·I
WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

Master Thesis

Microtheories for SDI - Accounting for Diversity of Local Conceptualizations at a Global Level

Stephanie Duce

25 February 2010

Dissertation Supervised by:

Rafa Berlanga Llavori PhD

Dept. Lenguajes y Sistemas Informaticos, Universitat Juame I, Spain

Krzysztof Janowicz PhD

Department of Geography, Pennsylvania State University, USA

Werner Kuhn PhD

Institute for Geoinformatics, Westfälische Wilhelms Universität,
Münster, Germany

Acknowledgements

I would like to sincerely thank my supervisors Rafa Berlanga, Krzyztof Janowicz and Werner Kuhn for all their help, guidance and their great ideas. It is very much appreciated. I am also grateful to the Erasmus Mundus Association for giving me the opportunity to undertake this program and all the experiences it has brought with it. I would like to thank the professors and staff at each of the three institutes who have helped, taught and inspired me along the way. I am incredibly grateful to my parents and family and friends in Australia for their great help, encouragement and faith in me. Equally, I would like to thank my classmates who have become my friends and family over the last 18 months.

Microtheories for SDI - Accounting for Diversity of Local Conceptualizations at a Global Level

Abstract

The categorization and conceptualization of geographic features is fundamental to cartography, geographic information retrieval, routing applications, spatial decision support and data sharing in general. However, there is no standard conceptualization of the world. Humans conceptualize features based on numerous factors including cultural background, knowledge, motivation and particularly space and time. Thus, geographic features are prone to multiple, context-dependent conceptualizations reflecting local conditions. This creates semantic heterogeneity and undermines interoperability. Standardization of a shared definition is often employed to overcome semantic heterogeneity. However, this approach loses important local diversity in feature conceptualizations and may result in feature definitions which are too broad or too specific. This work proposes the use of microtheories in Spatial Data Infrastructures, such as INSPIRE, to account for diversity of local conceptualizations while maintaining interoperability at a global level. It introduces a novel method of structuring microtheories based on space and time, represented by administrative boundaries, to reflect variations in feature conceptualization. A bottom-up approach, based on non-standard inference, is used to create an appropriate global-level feature definition from the local definitions. Conceptualizations of rivers, forests and estuaries throughout Europe are used to demonstrate how the approach can improve the INSPIRE data model and ease its adoption by European member states.

Contents

List of Figures

List of Acronyms

AI	Artificial Intelligence
BFO	Basic Formal Ontology
BOS	British Ordnance Survey
CC	Crown Cover
DL	Description Logics
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
EEA	European Environment Agency
EnvO	The Environmental Ontology
EU	European Union
FAO	Food and Agriculture Organization
GI	Geographic Information
GISci	Geographic Information Science
GIS	Geographic Information Systems
GCS	Good Common Subsumer
INSPIRE	Infrastructure for Spatial Information in the European Community
IFCD	INSPIRE Feature Concept Dictionary
KB	Knowledge Base
KYOTO	Knowledge Yielding Ontologies for Transition-based Organization
MT	microtheory
NCS	Nearest Common Superclass
OWL	Web Ontology Language
SDI	Spatial Data Infrastructure
SUMO	Suggested Upper Merged Ontology
UML	Unified Modeling Language
UN	United Nations
W3C	World Wide Web Consortium
WFD	Water Framework Directive

Chapter 1

Introduction

A great volume of data is available today. The majority of this data is directly or indirectly spatially referenced (Deng (2007)). Geospatial information can be used to integrate data from heterogeneous sources and is valuable for almost infinite purposes (Hart and Dolbear (2007)). There is a push to globalise this information making it easily accessible via the semantic web.

The ability to discover and share data and information from different sources has become crucial to successful and efficient planning, management and research and avoids wasteful duplication of effort in data collection and manipulation (Granell et al. (2008)). The seamless sharing of resources requires interoperability and the harmonization of spatial data sets and services (Bishr (1998)). Among the numerous barriers to interoperability (refer to Bishr (1998)), semantic heterogeneity is one of the most difficult to overcome, requiring robust and detailed definitions of geographic entities and the relations between them (Bennett et al. (2008)).

Semantics is the relationship between words and the real world things to which the words refer i.e. the meaning of words or terms. Semantics defines the relationships between computer representations and the real world entities to which they correspond in a certain context (Bishr (1998)). Semantic heterogeneity occurs when a word has multiple meanings or can be interpreted differently by people from different domains or backgrounds. Geographic features are particularly prone to semantic heterogeneity as their definitions are often vague and multiple-conceptualizations exist (Smith and Mark (2003)). Making local conceptualizations formal and unambiguous is fundamental to achieving semantic interoperability.

1.0.1 Background and Motivation

To successfully share data and information between different systems and from different sources it is imperative that the intended meaning and application (i.e. semantics) of the data is clear. The fundamental importance of overcoming semantic heterogeneity has long been recognised and is traditionally achieved by attempting to impose common definitions for geographic feature types (standardization)[1]. Spatial Data Infrastructures (SDIs) are designed to achieve optimal interoperability and data sharing though the co-ordination of terminology, technology standards, institutional arrangements and policies (Nebert (2004)).

In May 2007 the European Union (EU) launched the Infrastructure for Spatial Information in the European Community (INSPIRE). The INSPIRE initiative aims to create a Europe-wide SDI supporting cross-scale, cross-language and cross-border interoperability and accessibility for spatial data and information[2]. This involves the development of spatial data themes, web services, agreements on data and service sharing, coordination and monitoring mechanisms and, particularly relevant to this work, common metadata standards and geographic feature (object) type catalogs.

The EU is composed of a diverse range of ecosystems, climatic and physical conditions, cultures, languages and administrative systems. This makes the definition of a shared conceptualizations difficult. Requiring these unique member states to agree upon a shared conceptualization of geographic feature types is unrealistic and likely to result in a compromise that satisfies no one. If the INSPIRE guidelines are too generic and fail to sufficiently restrict possible interpretations, interoperability will not be achieved without manual, case by case adjustments. Conversely, overly specific guidelines could hinder implementation and reduce the usability of provided data. Another danger when creating such broad systems to ensure overarching interoperability is that important nuances in local and contextual terminology and conceptualization will be lost (Mallenby (2006)).

For the INSPIRE initiative to be effective, efficient and successful, all parties should be free to define feature types in a manner most suited to their unique environment though still consistent at an all-encompassing upper level. This introduces a struggle to create, integrate and maintain conceptualizations at a local and European level.

The importance of employing local conceptualizations for geographic features is well

[1]U.S. Geological Survey, *What is SDTS?*, `http://mcmcweb.er.usgs.gov/sdts/whatsdts.html`, 2003. Accessed 05–11–2009

[2]European Commission, *INSPIRE Directive.* `http://inspire.jrc.ec.europa.eu/index.cfm`, 2009. Accessed 25–10–2009

developed throughout the literature. Their vague boundaries (Smith and Mark (2003)), vague adjective-based definitions (Mark (1993); Mallenby (2007); Bennett et al. (2008)), meso-scale (Smith and Mark (1998)) and temporal dynamics (Frank (2003a)) mean that human perception and social agreement play a strong role in the local conceptualization of geographic features (Kuhn (2005); Hart and Dolbear (2007); Bennett et al. (2008); Mallenby (2006)). They are also prone to multiple conceptualizations depending on perspective (Egenhofer and Mark (1995)) and different conceptualizations may be conflicting. For example, a forest may simultaneously be a protected area, recreational area, plantation, agricultural area and so forth. These different perspectives give rise to potential socio-economic conflicts and hinder the classification, and hence, mining and retrieval of data. In addition, as geographic features (like forests) do not stop at borders, a forest in Spain may be regarded as meadowland in France. The categorization of a feature may also have legal and economic consequences as is the case with deforestation.

1.0.1.1 Motivating Scenario

Let us assume that an international service exists allowing national management agencies to search for and share geospatial data and information from other nations. The data is stored according to a common conceptual model of geographic feature types (as would be the case in INSPIRE). The following simple scenario demonstrates the practical need for a well established conceptual framework which takes into account local conditions and conceptualizations and allows effective querying and retrieval.

The Turkish Water Management Agency is creating a management plan for the Kizilirmak River in southern Turkey. They want to find data from similar rivers and determine which management techniques have been successful. However, when querying the online service to find data about 'rivers', similar Spanish rivers are not recognized because the formal definition of 'river' employed by the system requires that they contain *flowing water*. Thus, Spanish rivers, which may be dry much of the year, are not identified as 'rivers' in the data model. Thus, the results returned are incomplete and do not meet the needs of the user.

1.0.1.2 Providing Semantic Interoperability to SDI

Ontology has been suggested to provide semantic interoperability in the geospatial domain. Ontology is a broad and diverse field with roots in philosophy and now in computer science. In the context of this project, an ontology can be best defined as a shared vocabulary plus a specification (characterization) of its intended meaning (Guarino (1998)). Ontologies help to structure knowledge and improve our understanding of concepts of the world by clearly defining the entities of a domain, how these entities relate to each other and how they relate to entities from other domains (Gruber (1993)). By defining entities and their relations, ontologies help to overcome the problems of semantic heterogeneity described above. Uschold (2000) discusses issues arising from the creation, integration and maintenance of local ontologies, which represent the semantics of entities in a certain context, and global ontologies, which aim to allow interoperability between different local ontologies so a shared understanding can be achieved.

This challenge has been the subject of much research in the Artificial Intelligence (AI) domain. A promising approach is the use of microtheories. Microtheories are an internally consistent set of facts, similar to small ontologies. Separate microtheories can hold information about the same concept but hold incompatible facts. Microtheories have been employed in AI to facilitate modularity in large knowledge bases. However, their use in the field of SDI is yet to be explored.

This work presents and explores a novel method of structuring SDI data models using local microtheories. This allows the diversity of different geographic feature type conceptualizations across Europe to be preserved, while creating and maintaining a consistent global ontology at a European scale to support interoperability and overcome semantic heterogeneity.

1.0.2 Aims and Objectives

The overarching aim of this thesis is to demonstrate how microtheories, structured by space and time, can be used to overcome semantic heterogeneity caused by multiple local conceptualizations of geographic features. It adopts a bottom-up approach, based on non-standard inference, to create an appropriate upper-level microtheory which allows interoperability between local microtheories. The work examines novel methods of structuring microtheories and different approaches to defining geographic features ontologically. Of the many possible applications for this work, INSPIRE is treated as a use case throughout.

The following 6 objectives were identified to fulfil the overarching aim:

- Examine microtheory structuring principles, comparing geographic, administrative, cultural and linguistic structures, to determine which is most effective for this purpose.

- Discuss challenges in defining geographic features ontologically. Test the use of ecological functions and services as roles to formally define features in an intuitive manner.
- Define geographic features (rivers, forests and estuaries) from different European countries, in natural language, then create conceptual models and formalize these models into local microtheories.
- Compute appropriate feature definitions at a European-level based on reasoning upon the local microtheories.
- Compare the outcomes with the geographic feature definitions put forward by INSPIRE or other published geo-ontologies.
- Identify challenges and relevant directions for future work.

The key aim is not to present perfect, locally correct and logically consistent geographic feature definitions. This is best done in consultation with experts in the field from each country. This thesis aims to present and test the suitability of microtheories for SDI at a conceptual level. Application of the principles to the example scenarios of rivers, forests and estuaries is intended to reveal challenges in the methodology that warrant future investigation.

1.0.3 Thesis Structure

The following chapter (Background and Related Work) outlines the importance of the study and discusses relevant literature and related work. Chapter 3 (Methodology) evaluates different structuring methods for microtheories and suggests the most appropriate for the INSPIRE use-case. It then introduces the geographic feature scenarios to be treated and describes the methodologies employed to define and reason upon them. Chapter 4 (Application and Results) documents the application of the methodology to each feature type and presents the results of this work. Chapter 5 (Discussion) is dedicated to the discussion of results and the challenges they reveal in general and with respect to INSPIRE. It also identifies necessary avenues for future research. Chapter 6 (Conclusion) summarizes the work.

Chapter 2

Background and Related Work

A great deal of work towards achieving semantic interoperability in geospatial data sets has been undertaken in recent years (Kuhn (2005)). This section provides a brief review of the notions of INSPIRE, ontology development and similarity reasoning. It documents the semantic challenges associated with geospatial data and the need for local conceptualizations of geographic features. The benefits and challenges of ontological modularity and matching to achieve this are reviewed. The concept of microtheories is introduced and its potential application to the field of SDI is discussed.

2.1 SDI and INSPIRE

Efficient and successful management and planning requires the use of relevant, accurate and appropriately processed spatial data. The lack of quality, organization, accessibility and sharing of spatial information presents challenges across all levels of public authority in Europe [1]. Spatial Data Infrastructures (SDI) employ self-contained, specialized and interoperable web services to perform the five major activities needed for successful research, planning and management: discovering, accessing, updating, processing and visualizing spatial data (Nebert (2004)). Semantic interoperability is fundamental to each of these tasks (Janowicz et al. (forthcoming)) as users must understand the wider meaning of the data and information (Comber and Fisher (2005)). To achieve interoperability for data and information across the EU, the INSPIRE initiative aims to provide a system-independent infrastructure that is operational across multiple languages, professional and legal practices (Craglia (2006)).

[1] INSPIRE, *D2.8.1.8 INSPIRE Data Specification on Hydrography - Guidelines*, http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_HY_v3.0.pdf, 2009, Accessed 06–02–2010

INSPIRE requires member states to create and maintain infrastructures for spatial information which include: metadata, spatial data themes, spatial data services; network services and technologies; agreements on data and service sharing, access and use; coordination and monitoring mechanisms, processes and procedures (refer to Appendix Figure 7.1 to see the INSPIRE technical architecture overview)[2]. The INSPIRE Data Specifications aim to provide a description of data sets and additional information enabling it to be created, supplied to and used by another party. The Specifications are structured according to 34 Spatial Data Themes[3]. To introduce semantics to the INSPIRE Data Specifications a common or interoperable terminology and conceptual model for geospatial features is required. Thus, agreement on common terminology and standards must be achieved.

The 'Consolidated INSPIRE UML Model' is used to maintain the INSPIRE application schemas based on the Generic Conceptual Model. This is created based on the existing data holdings of Member States. Kuhn (2001) suggests that human activities and user needs should play a stronger role in the creation of such schemas and models. The INSPIRE Feature Concept Dictionary (IFCD)[4] contains terms and definitions for spatial object types in the INSPIRE data specification themes. Its purpose is to identify conflicts between specifications in different themes and ease the harmonization effort. While the INSPIRE guidelines concede that the Data Specifications need not be applied at the national level, Member States are required to transform existing data specifications into the INSPIRE specifications. For reasons discussed below, this will be difficult and may result in the loss of important semantic nuances.

2.2 Providing Semantic Interoperability to Geospatial Data

As discussed above, providing semantic interoperability to geospatial data is inherent to the INSPIRE initiative. In INSPIRE this is currently undertaken using the Unified Modeling Language (UML) to define a Generic Conceptual Model. Formal ontologies can also be used achieve semantic interoperability.

[2]INSPIRE, *Drafting Team 'Data Specifications' Deliverable D2.3: Definition of Annex Themes and Scope*, http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/inspireDataspecD2_3v2.0.pdf, 2007, Accessed 20–11–2009

[3]INSPIRE, *Drafting Team 'Data Specifications' Methodology for the development of data specifications* D2.6, v3.0. 2008 http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/D2.6_v3.0.pdf Accessed 07–02–10

[4]http://inspire-registry.jrc.ec.europa.eu/registers/FCD, Last updated 19–01–2010. Accessed 9–02–10

2.2.1 Ontology for Semantic Interoperability

Ontologies improve the understanding of the world and help facilitate interoperability (Smith and Mark (1998)). They define concepts through their relationships with other concepts thus capturing their semantics (Rodriguez and Egenhofer (2003)). When these relationships are formalized (for example in a logical, computer language such as OWL), computers are also able to 'understand' and reason about entities and phenomena. This ability to reason allows implicit relations and semantic similarity to be discovered and makes querying more efficient. Despite the potential for a healthy symbiosis between ontology and SDI the two are rarely mixed in the literature.

Ontologies vary markedly in level of detail and logic from 'lightweight ontologies' (e.g. thesauri and conceptual vocabularies) (Uschold (2000)) to more formal, axiomatic ontologies. The categories of these ontologies are defined in logic or some language that can be automatically translated into logic (Sowa (2000)).

2.2.2 Ontology in the Geospatial Domain

The field of 'geo–ontology' has arisen from the application of ontology to the geographic domain (Kun et al. (2005)). Kuhn (2002) outlines geo-ontologies as semantic reference systems for Geographic Information (GI) and Geographic Information Systems (GIS) at the type level. There are numerous advantages in the use of ontologies for the geographic domain. Mark (2009) claims that an expressive domain ontology 'is critical for semantic integration of information for machine-to-machine information services and for human information retrieval'. Ontologies provide a way to manage geographic data so that it can be queried with maximum efficiency and can enable semi-automatic processing of data (Bennett et al. (2008)). Smith and Mark (2003) suggest that ontologies can make GISystems friendly to untrained users. The creation of an ontology can also yield a better understanding of the structure of the geographic world (Smith and Mark (1998)).

Numerous existing ontologies deal with geographic features (e.g. British Ordnance Survey (BOS) Hydrology Ontology[5], gwsg[6] , The Environment Ontology (EnvO)[7] and EarthRealm[8]). These ontologies vary in their level of development, level of formalization (i.e. the logical language used), detail and perspective.

[5] http://www.ordnancesurvey.co.uk/oswebsite/ontology/index.html
[6] http://seres.uni-muenster.de/ont/_dom_/gwsg070612.owl
[7] http://www.environmentontology.org/
[8] http://sweet.jpl.nasa.gov/sweet

2.2.3 Ontological Challenges in the Geospatial Domain

Despite the advantages, the application of ontology to the geographic domain presents some distinct challenges (Smith and Mark (1998)). There are numerous attributes inherent in geospatial data (i.e. data pertaining to geographic features), which set the domain apart from other domains of study in which ontology has been embraced (e.g. medicine) (Kuhn (2005) and Hart and Dolbear (2007)). These challenges, and some possible solutions, are discussed below.

While geospatial data is grounded in physical, measureable reality it is often also based on human perception and social agreement (Kuhn (2005)). Humans conceptualize their environment based on multiple criteria such as their cultural background, knowledge, motivation and particularly space and time. Thus, to create a successful and useful ontology for the geographic domain it is imperative that the ontology reflects and captures the users' conceptualizations of geographic features (Janowicz et al. (2008a)). However, the classification of geographic features is difficult given their ambiguity, vague boundaries and dependence on the contextual setting (Bennett et al. (2008)). Mereology is the theory of parts and wholes i.e. what parts constitute a whole entity and the relations of part to part within a whole. Topology refers to the geometric properties and spatial relations (e.g. separated, connected etc...) of entities that are unaffected by changes in shape or size of the features. These two principles, 'mereotopology', are fundamental in ontology development (Smith (1996)). However, they are difficult to apply to geographic features as they, and their parts, rarely possess determinant, prominent, complete and static boundaries (Smith and Mark (1998)). Despite this, geo-ontologies still rely on merotoplogy in their definitions of features (e.g. BOS Hydrology Ontology and the EnvO).

Their vague boundaries and mesoscale also make the exact location of geographic features difficult to determine accurately (Hart and Dolbear (2007)). These issues are explored in depth by Smith and Mark (Smith and Mark (2003, 1998)). Geographic features can also change their parts and topology markedly over time due to seasonal change, natural geomorphic evolution, or anthropogenic activities. For example, a river in the south of Spain may only contain water during the wet season. A forest may be harvested and so not contain trees. Nevertheless, both entities maintain their identity as a river or a forest. Such inconsistencies prompt authors such as Frank (2003a) to advocate the inclusion of a temporal component in any geospatial ontology.

In addition, the definitions of different geographic features are often separated by vague adjectives. For example, canal, stream, creek and river or lake and pond are semantically separated by vague adjectives like big, small; wide, narrow; artificial, natural (Mark (1993) presents a good example of this). These adjectives are conceptually fuzzy and highly dependent on the context in which they are used (Bennett et al. (2008)). One person's *river* may be another's *stream*. Thus, the conceptualization of geographic features is highly subject to the context in which they exist and dependent on local knowledge (Mallenby (2006)). Bennett et al. (2008) suggests the use of supervaluation semantics to incorporate vague adjectives into formal ontologies. Such 'vagueness' in defining geographic features is a useful and integral part of language and allows subtleties to be appreciated. Thus, it is desirable to embrace this vagueness allowing the user to decide feature definitions most appropriate to their situation.

Given these challenges Kuhn (2001) suggests that it is more effective to create an ontology based, to as large an extent as possible, on user needs and activities. The author introduced a method exploiting texts describing human activities and using the notion of affordances to connect activities to objects (Kuhn (2001)). A similar approach could be applied to the geospatial domain as geographic features perform different ecosystem functions and services. Ecosystem functions are the biophysical processes which take place in an ecosystem and can be characterised without a human context while ecosystem services are the outcomes of ecosystem functions which benefit humans (de Groot et al. (2002)). In the context of this work these functions and services are considered to be 'roles', similar to the anthropogenic notion of 'affordances' used by Kuhn (2001). Smith and Mark (2003) also discuss the importance of the 'role' of mountains in attempting to define them.

The 'roles' of geographic features transcend the spatial and temporal vagueness which hampers the use of traditional mereotopology in formally defining them. They are likely to represent commonalities and distinctions in different local conceptualizations. For example, rivers, wherever they are, play the role of transporting water. In Germany a river could also play the role of providing transport (as their constant flow of water makes them navigable). However, in southern Spain the lack of water in some rivers most of the year means they are not perceived to have the role of providing transport.

2.2.4 Ontology Evaluation and Similarity Measures

Various methods exist to evaluate ontologies (Brank et al. (2005)). Semantic similarity reasoning is one method to test whether the use of 'roles' to define geographic features in this context improves the ontology's ability to capture human conceptualization. Similar to spatial reference systems, which can be used to measure the distance between objects in space, semantic reference systems allow semantic distance between types and individuals, represented in description logics, to be measured (Janowicz et al. (2008a)). This distance is known as the similarity. Similarity can be employed as a traditional top–down approach to check the adequacy of the knowledge base with respect to the application domain (Baader and Kuesters (1998)).

Semantic similarity reasoning compares the meanings of concepts rather than simply their structural, or syntactic similarity (Janowicz and Wilkes (2009)). It has been suggested to facilitate successful online geographic information search and retrieval (Rodriguez and Egenhofer (2003); Janowicz et al. (2007); Janowicz and Kessler (2008)) and for use as an indicator of the quality of ontology engineering (Janowicz et al. (2008a)). Janowicz et al. (2008c) examine how similarity can be used to support information retrieval in SDIs. For these purposes, similarity measures should have the ability to handle the expressivity of the description logics used and need to reflect human similarity rankings for the same set of concepts (Janowicz and Wilkes (2009)). This issue is further discussed in section 3.7.

2.3 Multiple Conceptualizations of Geographic Features

In addition to challenges defining geographic features ontologically and evaluating these ontologies discussed above, Egenhofer and Mark (1995) note, people use multiple conceptualizations of geographic space. For example, a forest can simultaneously be a protected area, a recreational area, a plantation, an agricultural area etc... depending on one's perspective. This is apparent in INSPIRE where numerous Data Themes overlap[9]. Smith and Mark (1998) suggest that, given their meso scale and vague boundaries, geographic features are also more subject to cultural differences in conceptualization than objects in other domains (these cultural differences are discussed further in section 3.3.1). This underlines the importance of maintaining local conceptualizations. However, different conceptualizations are often conflicting. For example, the 950 definitions for forest available in the literature vary in the threshold of required tree cover from 0 – 80 percent (Lund (2009)).

Standardization, i.e. getting all parties to agree on a common definition, is notoriously difficult (Uschold (2000)) and may result in a compromise that is too broad or too narrow and does not adequately satisfy the needs of users. This difficulty in reaching a common, shared conceptualisation is compounded in the case of geographic features by their inherently vague and context dependent nature which leads to inconsistent

[9] refer to European Commission, *INSPIRE Directive*. http://inspire.jrc.ec.europa.eu/index.cfm, 2009, Accessed 25–10–2009

usage of terminology (e.g. Bennett et al. (2008); Smith and Mark (2003)). Even when a common conceptualization is created and agreed upon by a collaborative group, Brodaric and Gahegan (2007) revealed great difficulties in achieving practical, in-the-field, consensus on conceptualisations of geographic features (in their case geological regions).

2.3.1 Local versus Global Conceptualizations

It has been established that the local context of use is an extremely important factor in defining geographic features (Hart and Dolbear (2007)) and local, context-aware definitions are needed. The need for local conceptualizations to be maintained and made interoperable is supported by Uschold (2000) who outlines and explores the issues associated with creating, integrating and maintaining multiple local terminologies or ontologies. He identifies four requirements for the standardization of local ontologies. These are *local autonomy* allowing local groups to own, create and use their own terminology; *flexibility and ease of maintenance* of the local ontologies; *global access* allowing local groups to access things from other groups even if terminology is different and; *stability* such that the systems using the ontologies are not regularly disrupted by changes to them.

He presents arguments for and against the use of a global reference ontology to manage different local ontologies. Authors such as Frank (2003b) and Kuhn (2005) recognize that, to make local conceptualizations meaningful and shareable, a linked architecture of ontologies is required. Janowicz and Kessler (2008) suggest the use of a domain level ontology, based on affordances and actions, to facilitate mapping between local vocabularies. This would remove the need to agree on common conceptualizations.

2.3.2 Defining the Global Level

If a global reference ontology is to be used to allow interoperability between different local ontologies (representing local conceptualizations) a suitable scope and method for creating this global ontology needs to be found. Uschold (2000) outlines a continuum of options for creating the global ontology. These include top–down or bottom–up approaches created from the union or the intersection of sets of terms from all local ontologies. The choice of method should depend on the intended use of the ontologies.

Measures such as the computation of most specific concept (MSC) of an individual and least common subsumer (LCS) between concepts have been suggested to facilitate a bottom–up approach to knowledge base construction (Baader and Kuesters (1998) and Janowicz et al. (2008c)). These methods could also be applied to define a common global level ontology based on local ontologies. The MSC is the lowest, or least, concept of which an individual is an instance. The most specific concept of each reference individual are compared and the characteristic common to all is the least common subsumer between individuals.

2.4 Ontological Modularity

An analogous topic is that of modularity in ontologies. This allows smaller, self–contained ontologies to be merged to create more complex ones (Grau et al. (2007)). Modularity is desirable as representations are easier to understand, reason with, debug and extend (Grau et al. (2007)). Keeping ontologies modular minimizes human errors inherent in maintaining a multiple inheritance hierarchy (Horridge et al. (2004)) and allows collaborative ontology engineering and reuse (Jimenez-Ruiz et al. (2008)). Modularity also allows different (sometimes conflicting) perspectives on a domain or concept to be taken into account and integrated in a meaningful way as shown by Hois et al. (2009).

Much research has been undertaken into achieving modularity in, and mapping between, ontologies (e.g. Grau et al. (2007); Bateman et al. (2007); Jimenez-Ruiz et al. (2008); Cruz and Sunna (2008); Hois et al. (2009). However, according to Shvaiko and Euzenat (2008), who review the state of ontology matching solutions, an acceptable, integrated solution, usable by non–experts has yet to be developed.

Work by Bateman et al. (2007) suggests that, as handling spatial phenomena requires the use of numerous different theories concerning space, the ontological modules representing these theories must be formally related using ontological engineering. Hois et al. (2009) employed Econnections to link ontological modules based on different logical assumptions. This allowed them to reconcile different perspectives on phenomena in the architecture domain to take into account different ambient intelligence (AmI) requirements in design criteria.

2.5 Microtheories

The challenge of modularity and handling local conceptualizations at a global level has also been a core topic in Artificial Intelligence research for many years (McCarthy (1987); Wachsmuth (2000)). One promising approach is the use of domain specific microtheories (also called contexts). Microtheories are internally consistent sets of facts which are relevant to a particular domain (Cycorp (2002a)). Each microtheory

(MT) is designed as a coherent set of statements about vocabularies. Microtheories are usually related hierarchically such that all the assertions in more general microtheories hold for more specific ones. For the purposes of this work, a microtheory can be thought of as a modularized, local ontology.

Microtheories are employed in Artificial Intelligence (AI) to facilitate modularity and improve inference efficiency in large knowledge bases (e.g. Cycorp (2002a)). They have been applied in natural language understanding, common sense reasoning systems and model based reasoning (Guha et al. (2004)). The use of microtheories has also been proposed to achieve the Semantic Web vision (Guha et al. (2004)). However, their potential for use in reconciling local and global ontologies (Uschold (2000)), which may hold conflicting definitions for the same concept, has yet to be fully explored. The use of microtheories to organize SDIs has also yet to be examined.

The most advanced and well-documented use of microtheories is that of the Cyc Knowledge Base (KB) which aims to create the 'world's first true artificial intelligence, having both common sense and the ability to reason with it'[10]. The Cyc KB formally defines a vast quantity of human knowledge in the form of thousands of microtheories.

2.5.1 Why Use Microtheories?

The use of microtheories for knowledge representation and reasoning has numerous benefits (Cycorp (2002a)). These benefits are discussed below with respect to the geospatial domain.

2.5.1.1 Ability to handle conflicting definitions

One of the greatest benefits of microtheories, and that of most relevance to this work, is that separate microtheories can hold information about the same concept but provide incompatible facts (Cycorp (2002b)). Usually microtheories are organized in subsumption hierarchies such that facts specified in the super-microtheory must also hold in each of its sub-theories. Sibling-theories, however, may contain contradicting conceptualisations. As each microtheory is considered an object in its own right and is only applied to the world under specific conditions, two microtheories can hold conflicting facts without undermining the reasoning capacity of the entire KB (Hovy (2002); Cycorp (2002b)). This allows inconsistencies which would otherwise make reasoning impossible.

For example, from the INSPIRE perspective, different conceptualizations of the

[10]refer to www.cyc.com

same geographic feature may conflict with each other. Germany's conceptualization of 'river' may state that it contains flowing water or this may be inferred by river's subsumption by the broader classification of waterbody. However, in Spain, where rivers may be dry for most of the year, the definition of river can not be that it contains flowing water. Traditional ontologies, which are strongly bound by the rules of logic, cannot accept such conflicts. Therefore, in order to merge the definitions of rivers in Spain and Germany to create a Europe-wide conceptualization of 'rivers' (as needed for INSPIRE), one or the other or both would need to be changed to an unrealistic definition that does not reflect the nature of the features in that country. This is undesirable and undermines the success of the INSPIRE initiative.

Microtheories offer a solution to this challenge by allowing conflicting definitions of the same concept to be held within one KB (i.e. the concept of river can simultaneously be defined as containing flowing water in one microtheory but not in another).

2.5.1.2 Increased inference, update and re-use efficiency

In addition, microtheories can be used to provide modularity in ontologies. This makes reasoning and querying more efficient as only relevant microtheories are used to answer a query (Cycorp (2002b)). For example, to process a description of the flow regime of a Spanish river, no information is required about the ecological services of a shallow wetland, which could be included in the same class if an overly broad definition is used.

As Brodaric and Gahegan (2002) recognised, concepts in geoscientific domains are regularly evolving as better understanding is achieved. This results in semantic gaps between understanding and the static way concepts are represented in ontologies. The use of modular, self-contained microtheories could allow conceptualisations to be updated more quickly and simply without having to make widespread changes in the KB. This increases the likelihood of safely re-using ontologies.

2.5.2 Structuring Microtheories

So far, microtheories have only been structured by establishing hierarchical relationships between them, i.e. by generalization. Other potential ordering principles such as space, time, or cultural background have received virtually no attention in the semantic Web community. While their importance has been recognized recently, existing work reduces space and time to simple latitude-longitude pairs and time stamps. Tobler's First Law of Geography states that 'Everything is related to everything else, but near things are more related than distant things' (Tobler (1970)). Climatic, geographic and geological factors, all of which adhere to the above law, govern the character of geographic features and hence influence their categorization. Besides their role in the gradual change of the environment, space and time are the most fundamental ordering relations used in human cognition and language.

2.6 Conclusion

This chapter introduced the goals of the INSPIRE initiative which depend upon the creation of standardized data specifications to overcome semantic heterogeneity and allow interoperability between geospatial data sets across the EU. It outlined factors such as vagueness, temporal dynamics and multiple perspectives, which are inherent in our conceptualization of geographic features and mean that conceptualizations and terminologies vary from place to place. This highlighted the importance of maintaining local conceptualizations for geographic features while making them interoperable at a common global level. It discussed methods to achieve this based on existing literature and introduced microtheories as a promising solution.

Chapter 3

Methodology

This chapter describes the rationale of the work and introduces the geographic features to be treated. It continues the discussion of microtheories evaluating different possible structuring techniques and suggesting the most effective with respect to the INSPIRE scenario. It outlines the steps performed to conceptualize and formalize local definitions of geographic features as microtheories. It also details the reasoning performed on these microtheories to determine an appropriate Europe-wide[1] microtheory which could act as a 'global' definition providing interoperability between the local definitions.

3.1 Rational

The purpose of an SDI is to manage data in an effective way and facilitate efficient query of the data. This requires that data, and the real world features they refer to, are faithfully represented in the knowledge base. Thus, broad definitions which encompass the spectrum of possible conditions are not effective. For example, if you are searching for data from rivers you do not want your search to yield results for seas, lakes and estuaries as well as more relevant results.

Deciding the level of detail to be represented at the local and at the global level is a critical step in creating successful, interoperable ontologies (Uschold (2000)). Microtheory structuring principles, discussed below, are used to determine an appropriate scale for the 'local' microtheories. This work then adopts a bottom up approach to defining an appropriate[2] global level microtheory by first creating local, natural lan-

[1] In this work the term 'global' is sometimes used instead of Europe-wide to refer to the upper-level definition.

[2] Appropriate is defined here as a conceptualization that is neither too broad nor too specific.

guage definitions, using all relevant detail, based on research into the character of the features in each country. These local definitions are expanded into conceptual models mapping the relations between entities in the domain. The conceptual models are then translated into formal language (OWL) using Protégé to create local microtheories.

A good common subsumer (GCS) between local microtheories is calculated and used to determine a global-level definition which can be used to map between the local microtheories. This ensures the global definition is neither too broad nor too narrow. Similarity reasoning, using the SIM-DL reasoner (Janowicz et al. (2007)), is used to provide insight into how well the conceptualizations capture the domain and reflect human intuition.

A flow chart outlining the methodology employed in this work is shown in Figure 3.1. This methodology demonstrates how each E.U. member state can create its own microtheory for features to be used in INSPIRE based on existing natural language definitions. The upper-level, 'global' definition could be used as a guide and as a default microtheory should a country not have created its own.

3.2 Geographic Feature Scenarios

This work focuses on the definition of rivers, forests and estuaries throughout Europe. These features were chosen to best demonstrate the benefits of the microtheories approach as they are highly dependent on geographic conditions including climate, topography and biogeography. Thus, their characteristics and conceptualizations vary from place to place. Their definition is often subject to conceptual vagueness and they are prone to multiple conceptualizations.

Rivers, forests and estuaries cover large areas and produce many environmental and commercial goods and services. They are of high importance from economic, social and environmental perspectives and have many stakeholders. The ability to understand, study and effectively manage these features is of the utmost importance. Information sharing between stakeholders, facilitated by an SDI, is fundamental to achieving this aim.

The features treated here are highly complex eco-systems and commercialized anthropogenic entities. The definitions presented in this work do not claim to encompass all their elements. They show how diverse elements can be used to define local conceptualizations without undermining global interoperability. This work also improves our understanding of the complex nature of the features and elucidates some ontological principles and areas that require future work.

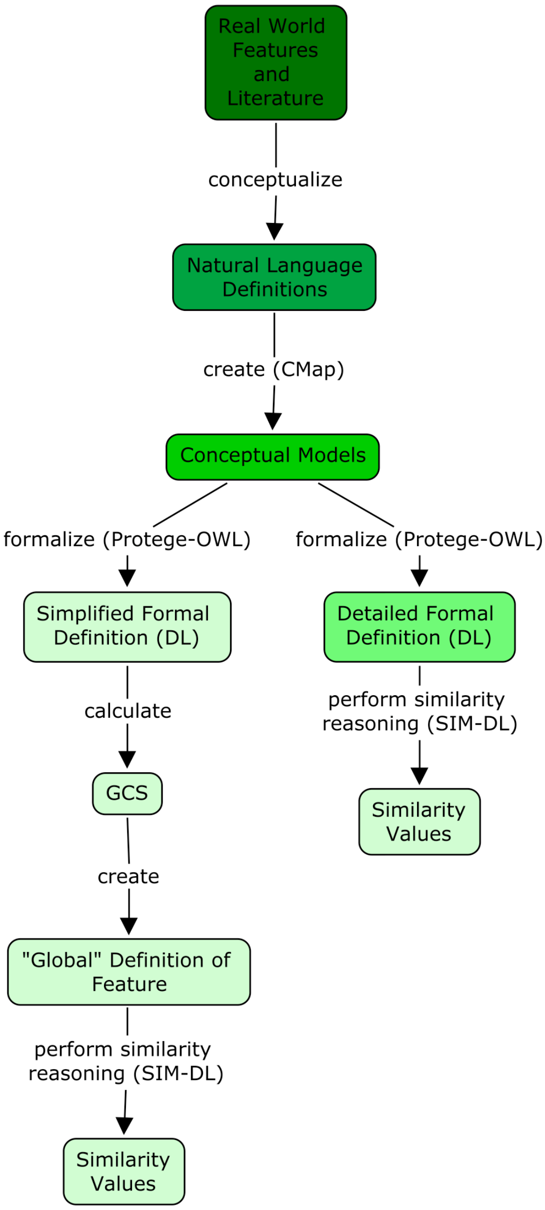


Figure 3.1: Flow chart showing the steps taken to arrive at the 'global' level microtheory for each geographic feature. Decreasing color intensity indicates decreasing expressivity of the language used.

The definitions of each feature were created to reflect the local (i.e. country specific) conceptualization. The countries used for each feature were chosen based on the contrasting conditions (and therefore, conceptualizations) of the features between them. Spain and Germany were chosen for the treatment of rivers as they represent different ends of the spectrum of contrasting river conditions across Europe. Forests in Austria, the Czech Republic, Denmark, Italy, Portugal, Spain, Switzerland and the UK were chosen in order to capture the array of different forest types across Europe and because they employed comparable parameters in their legal definitions of forest. Norway and the Netherland's were chosen as case studies for the estuary scenario as they provide stereotypical examples of two contrasting estuary types (fjord and bar-built). It is important to remember that the characteristics of features may vary across the country and the definitions represented here reflect only the most common conditions. The implications of this and ways to overcome it using microtheories are discussed in section 5.2.2.

3.3 Structuring Microtheories

The structuring of microtheories refers to the way they are related to one another and the granularity of reference of a single, 'local' microtheory. It is important to choose a structuring principle that is effective from ontological, geographic, cultural and political perspectives. In Cyc microtheories are usually related to one another hierarchically, using a generalization relationship (refer to Cycorp (2002b)). This creates an ordered set of microtheories such that more specific ones are subsumed by more general ones. The potential of structuring of microtheories using other ordering principles such as space, time or cultural background have received virtually no attention in the semantic Web community.

An exception is Janowicz (2009) who suggests further research into the use of microtheories, structured by time and space, to harmonise heterogeneous data sources based on differing ontological assumptions. In this work, individual microtheories can be thought of as local ontologies (discussed by Uschold (2000)). The following discusses the potential use of microtheories structured based on spatial (geographic, climatic and administrative) and cultural or linguistic boundaries, with respect to the INSPIRE initiative. Figures 3.2, 3.4 and 3.5 illustrate the proposed 'global' level, Europe-wide microtheories for geographic features (in blue) and the different possible structuring methods for lower level microtheories (in grey). Each feature (represented in green) would be a sub-microtheory to these lower level microtheories.

The hierarchy of microtheories in the OpenCyc KB is created using a formal generalization relationship called *genlMt* (see Cycorp (2002b) for details). To structure microtheories using other containment principles this generalization relationship could be extended such that lower microtheories are also spatially contained within the upper microtheories. However, the structuring approaches are treated at a conceptual level here. The formalization of these relations is left for future work.

3.3.1 Cultural and Linguistic Structuring

Multi-lingualism presents a substantial problem to ontology matching and further complicates the pursuit of interoperability throughout the EU (Vossen (2008)). The Knowledge Yielding Ontologies for Transition-based Organization (KYOTO) recognises the importance of allowing 'experts and non-experts to access information in their own language, without recourse to cultural background knowledge' (Vossen (2008)). Adopting cultural or linguistic boundaries as ordering principles (outlined in Figure 3.2) for microtheories could overcome barriers posed by natural language boundaries and very accurately represent conceptualizations across the European Union.

Frank (2009) discusses difficulties translating directly between vocabulary and terms which may hold different connotations from language to language. Mark (1993) demonstrated that different distinctions in physical characteristics are combined to define similar landscape features in French and English. Thus, geographic terms are not strictly able to be translated and the case for language-specific conceptualisations, and microtheories structured around language boundaries, is strong. However, there are numerous factors which would hinder the use of language boundaries as structuring principles for microtheories when defining features across the EU.

The EU has 23 official languages and more than 10 additional unofficial, but widely spoken, languages[3]. Often the same language is spoken in different countries where geographic conditions may differ considerably. For example, English is the official language of the UK and Malta which are on opposite sides of the EU and experience vastly different climatic and geographic conditions. Thus, a single 'English' microtheory would most likely not reflect the conditions in both places well. Also, many countries in the EU have more than one official language overlapping in the same place. From a practical perspective this could cause confusion and unnecessarily complicate the process.

In addition, including more than one language in a single ontology has not, to the author's knowledge, been attempted. Even using microtheories capable of bearing inconsistencies, it may not be possible given the present state of technology. Future research in this direction would be useful.

[3] www.wikipedia.org

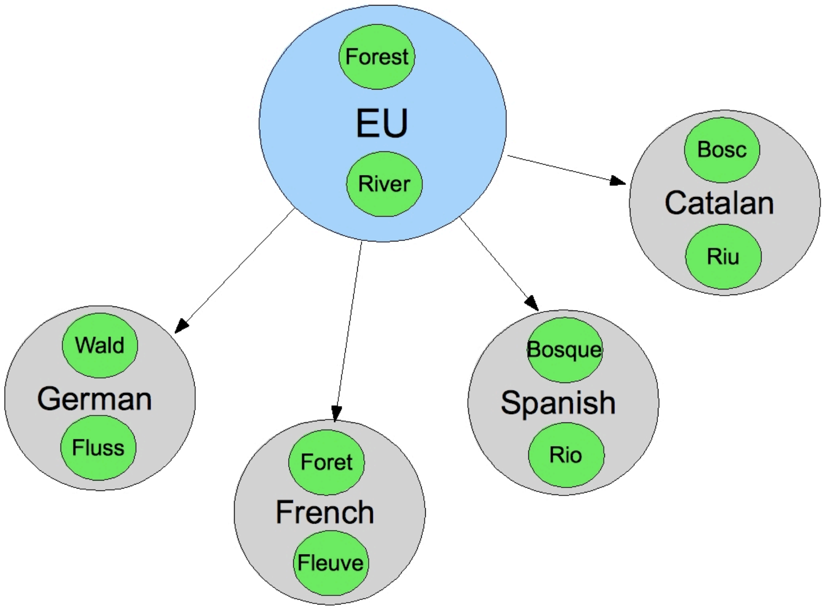


Figure 3.2: Example of language-based structuring for geographic feature microtheories.

3.3.2 Spatial Structuring

Tobler's First Law of Geography states that 'Everything is related to everything else, but near things are more related than distant things' (Tobler (1970)). Climatic, geographic and geological factors, all of which adhere to the above law, govern the character of geographic features. Thus, it is probable that they also influence how these real-world features are conceptualized. As INSPIRE aims to create a *Spatial* Data Infrastructure it makes sense that the conceptualizations of geographic entities be structured using spatially relevant principles.

It can be argued that the cultural and linguistic differences in a peoples' conceptualization of geographic features, discussed by Smith and Mark (1998), are driven primarily by factors which vary spatially such as climate and geology. For example, the Yindjibarndi Aboriginal people from arid, outback Australia, where no permanent or even seasonal watercourses exist, do not have a word for 'river' in their language (Mark and Turk (2003)). Thus, it is probable that linguistic and cultural changes in conceptualization largely follow spatial or climatic boundaries.

3.3.2.1 Climatic Boundaries

Figure 3.3 shows the climatic regions across Europe. These were used as guidelines to produce the structuring divisions suggested in Figure 3.4. Southern Europe coincides with subtropical dry summer and semiarid climatic zones including most of Spain, Portugal and Italy. The North-Western Europe microtheory would represent the humid oceanic climatic zones of France, Germany, the UK and part of Norway. Eastern Europe covers the many countries identified as humid continental zones (Figure 3.3) while the Northern Europe microtheory includes the northern parts of Norway, Sweden and Finland.

Using this structuring method some countries would belong to more than one microtheory group. For example, Spain would be included in the North-Western and Southern Europe microtheories and Germany would be represented half by the North-Eastern and half by the North-Western European microtheory. Thus, using climatic boundaries as the structuring unit for microtheories across Europe would also require numerous countries, with very different cultures and languages, to agree on a single conceptualisation. The North-Eastern Europe microtheory, for instance, would require agreement between more than 15 countries on conceptualisations of geographic features. This would be virtually no different to the standardization of feature definitions across the entire EU and would create complexities and be politically difficult to coordinate and implement in practice.

In addition, each area captured in the climatic structured microtheories is very large and likely to possess local heterogeneity not captured using this broad-scaled structuring system. Many complex factors, including topography and land cover, induce spatial climatic variations at a local and regional scale (Thomas and Herzfeld (2004)). For example, some valleys in the inner Alps, have vastly different climatic regimes due to orographic lifting and the rain shadow effect controlling precipitation in the area. These factors are likely to be reflected in the character of many geographic features in the area.

3.3.2.2 Administrative Boundaries

A structuring of microtheories based on administrative boundaries takes advantage of the spatial principles discussed above but goes some way towards over coming the challenges and shortcomings of structuring microtheories based on climatic boundaries. This structuring scenario (Figure 3.5) can be seen as a compromise between the cultural, linguistic and climatic boundary structuring methods mentioned above.

Figure 3.3: Climatic Zones of Europe[4]

.

Although administrative boundaries do not reflect climatic and geographic boundaries completely they do break the problem to a smaller size and are intuitive divides from a political perspective. Under this method each EU member state would be free to have its own microtheory, best reflecting conditions and legislation in its country. It is likely to be politically achievable and align well with present data models which are usually nationally created.

Thus, administrative structuring seems the most effective method, although it is not ideal as the territories of countries are large and diverse themselves. Also, a country may possess territory in a different location where geographic features may be very different (for example, the UK and Gibralta). To overcome these issues, autonomous or independent regions could be free to make their own sub-microtheory for features where necessary. A nation-wide microtheory could then be generalised from the internal regions. Figure 3.5 shows a small demonstration of potential division of microtheories using administrative structuring. Of course, different situations may require different microtheory splits. Administrative containment is the structuring approach used in this work.

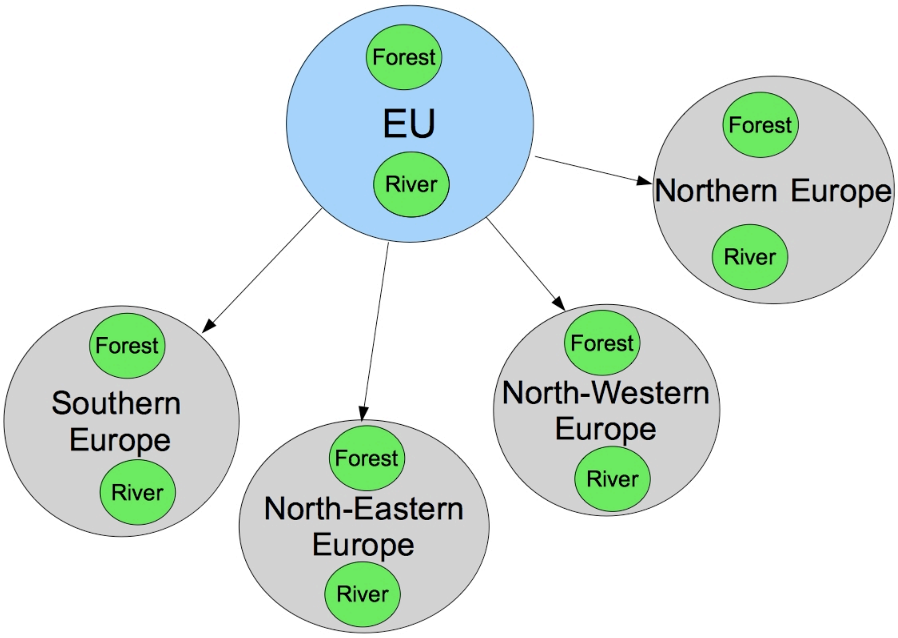


Figure 3.4: Example of climate and geography-based structuring for geographic feature microtheories.

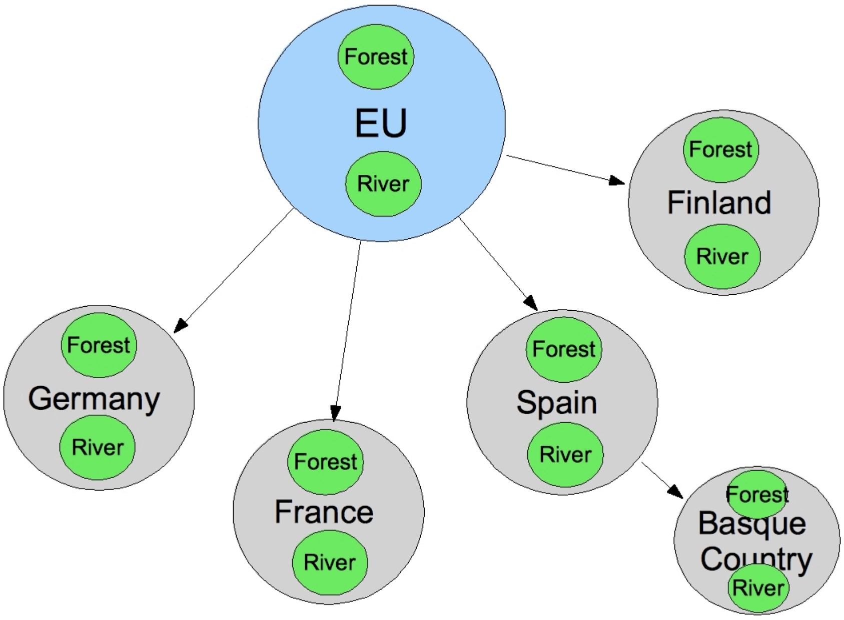


Figure 3.5: Example of administration-based structuring for geographic feature microtheories.

3.4 Natural Language Definitions

Having decided upon administrative boundaries as the most effective microtheory structuring method for the INSPIRE scenario, expressive natural language definitions for each geographic feature were derived in consultation with numerous sources. To create intuitive and practical definitions and overcome some of the challenges discussed in section 2.2.3, many characteristics were included. In this work, geographic features in different countries were defined based on their *physical parts* (the endurants of which they consist), *qualities* (their measureable attributes), their *participation in perdurants*, and their *roles*.

Geographic features perform different ecosystem functions and services. In this work these functions and services are considered to be thematic *'roles'*, similar to the anthropogenic notion of affordances used for modeling by Kuhn (2001) and others in GIScience. These properties of geographic features transcend the spatio-temporal vagueness and variability which hamper the traditional use of mereotopology in defining rivers and are likely to represent commonalities and distinctions in different local conceptualizations (as discussed in section 2.2.3).

3.5 Conceptual Modelling

The natural language feature definitions were expanded into semi-formal conceptual models mapping the relations between entities in the domain. IHMC CmapTools[5] was used to create the conceptual models.

3.5.1 Top–level Ontology Alignment

It is widely accepted that to create a coherent, systematic and complete ontology (and microtheory) it should be aligned to a foundational (top–level) ontology (Schneider (2003); Frank (2003a)). Top–level ontologies specify the meanings of important, overarching, domain–independent entities in the real world and the relationships between them (Bittner et al. (2009)). They provide a 'common neutral backbone' from which more specific domain ontologies can be built (Smith (2003)).

Currently several top–level ontologies exist (e.g. DOLCE (Masolo et al. (2003)), BFO (Spear (2006)) and SUMO (Niles and Pease (2001))). This work was loosely aligned to DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) as the top level ontology (Masolo et al. (2003)). DOLCE aims to negotiate the meaning of entities at a foundational level which will enable co–operation and consensus between humans and artificial agents (Gangemi et al. (2002)). It is recommended for use in the creation of spatial ontologies[6]. Klien (2008) states that DOLCE's cognitive bias and its top-level notations make it particularly suited to representing geographic entities.

DOLCE divides all things (particulars) into four categories *Endurant, Perdurant, Quality* and *Abstract* (Gangemi and Mika (2003)) (the taxonomy of DOLCE basic categories is shown in Figure 7.2 of the Appendix). Endurants are entities which are present, in full, at any time that they are present, while perdurants are processes which extend through time by accumulating different temporal parts. Thus, perdurants are only partially present at any time as their past and future 'parts' are not present at all times (Gangemi et al. (2002)). Physical endurants have a clear spatial location, while the spatial location of perdurants is determined by the endurants which participate in them. Qualities are basic entities that can be perceived or measured (Masolo et al. (2003)). The notion of roles, as they are discussed here, refers to 'functional roles' recognised in DOLCE (Masolo et al. (2003)).

These categories were taken into account during conceptual modeling and entities

[5]http://cmap.ihmc.us/conceptmap.html

[6]Bremen Ontology Research Group. *Spatial Ontology*, http://www.ontospace.uni-bremen.de/spatialOntology.html, 2009. Accessed 15–12–2009

are divided broadly between DOLCE's upper-level classes. However, the microtheories presented here were not directly aligned to DOLCE for the following reasons. The direct alignment of a geographic feature ontology to the DOLCE categories requires an intermediate 'core' or 'domain' ontology (Klien (2008)). A satisfactory core ontology does not yet exist for the geographic domain. Direct alignment would also mean that each microtheory is considerably larger and more complex. In addition, the ambiguity of geographic features creates some difficulty when attempting to align them with the upper-level classes defined by DOLCE (Duce (2009)).

3.5.2 Ontology Grounding

In addition to top-level alignment, it is argued that practical 'grounding' of ontologies in real-world measurements will better clarify inter-domain similarities between entities and help overcome vagueness (Mallenby (2007); Scheider et al. (2009); Kuhn (2009)). Measureable attributes often help to define and distinguish features (Mallenby (2007)) (e.g. level of salinity and depth help define the type of estuary). Thus, to establish more meaningful semantics and interoperability the measurable qualities of entities (e.g. water depth, water quality etc...) were included in the conceptual models with the intention of linking them with 'measurement ontologies' currently under creation (refer to Kuhn (2009); Scheider (2009)).

3.6 Formalization

The conceptual models were formalized into the Web Ontology Language (OWL). OWL is an expressive knowledge representation language, based on Description Logic (DL). It adopts the open world assumption (i.e. the knowledge base does not contain all there is to know and a statement is not necessarily untrue just because it is not contained in the KB) and is also free from the unique name assumption (Horridge et al. (2004)). OWL is compatible with the architecture of the World Wide Web and the Semantic Web (McGuinness and van Harmelen (2004))and is endorsed by the World Wide Web Consortium (W3C) (W3C (2007)).

Protégé is an open source, java based, ontology editor which allows added functionalities through plug-ins (Horridge et al. (2004)). It is the standard editor for DL based reasoning today (Janowicz et al. (2007)), and the Protégé OWL Plugin (Knublauch et al. (2004)) was used to create, explore and modify the river, forest and estuary definitions as OWL ontologies. Protégé version 4 was used to create and perform standard subsumption reasoning and check the consistency of the microtheories. Protege version

3.3.1 was employed for the similarity reasoning as it provides plug-ins to the SIM-DL reasoner.

Note that in this work local definitions of each feature were represented in a single ontology to allow reasoning in Protégé. In fact, they are in separate microtheories and hence are all named appropriately (i.e. *SpanishRiver, GermanRiver*). Difficulties arising from semantic heterogeneities are captured by this approach as all definitions have an administrative scope.

3.6.1 Computation of the Global Level

Once the local, member state specific, conceptualizations were formalized, it was proposed to employ the reasoner to compute the common, global level definition for each feature type. Least common subsumers (LCS) are concepts or restrictions that are common to the definition of different entities in an ontology. The LCS of concepts represents the most specific concept which subsumes all the concepts. This inference technique has been suggested to facilitate a bottom–up approach to knowledge base construction (Baader and Kuesters (1998); Baader et al. (2007); Janowicz et al. (2008c)). In this work, the LCSs were intended to determine the intersection between different local microtheories defining the same feature. The elements of commonality in each microtheory will be used as a global definition of the geographic feature which is neither too broad nor too narrow.

It is important to note that present methods for computing LCS are restricted to inexpressive description logics which do not allow for disjunction (i.e. the use of the union operator) in concept definitions (Baader et al. (2007)). In such cases the LCS would simply be the disjunction of the collection of concepts which does not reveal anything new and does not well represent a common global definition. As the definitions for geographic features presented here rely heavily on the use of disjunction to produce intuitive definitions this is unsuitable.

To overcome this Baader et al. (2007) propose the computation of a 'good' common subsumer (GCS) with respect to a background terminology. The use of vivification, whereby a series of disjointed concepts is replaced by their LCS, has been suggested to allow reasoning on knowledge bases which contain disjunction (Cohen et al. (1992)). While the vivified concept is only an approximation of the original it is useful to overcome limitations posed by the use of disjunction in the definitions and so was employed in this work. Both these methods require manual interaction and, as the microtheories produced in this study were small, the intuitive 'good common subsumer' was calculated by hand in accordance with the generic rules defined below.

First, in order to determine where to apply each rule, the normal form of concept definitions in each microtheory was defined. This ensured that the correct corresponding parts of each sibling microtheory would be compared when calculating the GCS i.e. the *roles* defined for a SpanishRiver would be compared with the *roles* defined for a GermanRiver and so on. A conjunctive normal form was used. It was structured as follows:

$$GeoFeature \equiv Roles(GeoFeature) \sqcap PhysicalParts(GeoFeature) \sqcap Qualitaties(GeoFeature)$$

$Roles(GeoFeature)$ is a disjunction of possible roles for the geographic feature, and $PhysicalPart$ and $Qualitaties$ are conjunctions of concepts related to the physical and qualitative properties of the geographic feature.

The generic rules applied to the normal form concepts of two microtheories to compute the good common subsumer (GCS) are listed below with examples:

- **Roles**. All disjunct roles of each microtheory will be vivified back to their nearest common super–role.
  $mt_1 : \exists R_1 \sqcup R_2$
  $mt_1 : \exists R_1 \sqcup R_2 \sqcup R_3$
  $GCS : \exists R$
  Where R is the nearest common super–role of R_1, R_2 and R_3.

- **PhysicalParts**. Elements belonging to the *PhysicalParts* part of the definition will be present in the GCS only if they are shared by both microtheory concepts. If they are restrictions, and the same physical part is not shared they will be vivified back to their nearest common superclass (NCS).
  $mt_1 : \exists P.F \sqcap \exists P.C$
  $mt_2 : \exists P.F \sqcap \exists P.B$
  $GCS : \exists P.F \sqcap \exists P.A$
  Where A is the nearest common superclass of C and B.

- **Qualities**. When nominal, qualitative or numeric threshold values are included as the fillers in restrictions, the GCS is created using the nearest common superclass, as long as both microtheories have the same defined hierarchy over these nominals. Otherwise, a filler is created with the union of the nominals.
  $mt_1 : \exists P.Spain$
  $mt_2 : \exists P.Germany$
  $GCS : \exists P.\{Spain, Germany\}$, or

$GCS : \exists P.MemberState$

if there exists a common hierarchy with member states in both microtheories.

3.6.2 Classifying Instances

The built–in FaCT++ (Fast Classification of Terminologies) (Horrocks (2003)) Reasoner in Protégé 4 was used to infer a class hierarchy and check the consistency of the ontologies. It was also employed to test if the restrictions were sufficient to allow the reasoner to correctly classify instances which it had been given.

3.7 Similarity Reasoning

Similarity reasoning between concepts in an ontology can be used to assess how well the definitions capture the domain and reflect human conceptualizations (Janowicz et al. (2008a)). SIM-DL[7] is an asymmetric, context-aware similarity measurement theory used for information retrieval (Janowicz et al. (2007)). It compares a DL search concept with one or more target concepts, all in canonical form, by measuring the degree of overlap between their definitions. A high degree of overlap indicates high similarity (refer to Janowicz et al. (2007) for a detailed description).

SIM-DL was ideal for use in this work as it does not require a populated ontology, can handle the expressivity of different description logics and has an intuitive user interface (Janowicz and Wilkes (2009)). SIM-DL has the added functionality of allowing the ontology creator to estimate the similarity of concepts based on their own conceptualization or the conceptualization of experts in the field (Janowicz et al. (2008a)). This estimated similarity can then be compared to the calculated similarity between the defined concepts in the ontology. This gives an indication of how well the ontology captures the conceptualization of the real world. Similarity using SIM-DL has been suggested to improve the intuitiveness of information retrieval in SDIs (Janowicz et al. (2008c))

The similarity value calculated by SIM-DL is sensitive to the context which is defined, the way the conceptualization is defined and the relationships used. Therefore, the ontologies had to be created specifically for use with the SIM-DL reasoner and could not contain instances. The restrictions used to define concepts had to be given in the same order and use the same properties etc... The overall similarity between concepts

[7]The current version of SIM-DL only supports a subset of OWL-DL and can be freely downloaded at http://sim-dl.sourceforge.net/.

is the normalized sum of the similarities for all parts (i.e. restrictions, subconcepts and superconcepts). A single similarity value is not indicative of absolute concept similarity and can only be considered relative to the other similarity values. Thus, SIM-DL delivers a normalized similarity ranking and orders the concepts from most (a similarity value of 1.0) to least (a similarity value of 0.0) similar.

Similarity reasoning was used in this work for two purposes:

1. To determine if the EU wide definitions created based on the GCS were appropriate and would allow a search to retrieve data from all member states and to see if this is also the case for the definitions provided by INSPIRE and other organizations (e.g. FAO and BOS).
2. To provide some indication of how well the local microtheories defined for the feature types, represent how those features are intuitively conceptualized.

To achieve these aims the SIM-DL similarity server was run using the *maximum* and *asymmetry* modes. When concepts are composed by disjunction, SIM-DL distinguishes between *maximum* and *average* similarity modes. The *maximum* similarity mode computes the similarity value between two concepts using the most similar concept that is part of the disjunction (Janowicz et al. (2007)). As the concept definitions in this work relied heavily on the use of the union-operator (disjunction) this was the most appropriate setting to illustrate the level of similarity between two concepts. The default *asymmetry* setting of SIM-DL means that the direction in which concepts are compared also affects the results (i.e. the similarity between a and b is not necessarily equal to the similarity between b and a).

To achieve the first aim, the EU wide definitions were compared to the local definitions and the definitions of other organizations (e.g. INSPIRE, BOS and FAO) and the similarity between them calculated. Ideally, in the context of searching, the similarity between the broadest term which is the search term (in this case the EU wide definition) and the intended terms (in this case the local definitions) should be 1.0. If this is the case a search for the broad term, in a semantics-enabled interface for Web gazetteers or Web Discovery Service in INSPIRE, would return all the intended terms and satisfy the user's requirements. A lower similarity value may or may not return all desired results depending on the similarity threshold defined (as discussed in Janowicz et al. (2008c)).

To achieve the second aim the similarity between concepts calculated by SIM-DL was compared to estimated similarity. This comparison is most relevant if the feature is treated in more than two member states and so was performed for the forest scenario.

As experts in the field were not available, the estimations of similarity between forest definitions were made by the author, based on the geographic proximity of the countries (given Tobler's 1st law of geography (Tobler (1970))) and the author's personal conceptualization of the features in each country. In the forest scenario three different ontologies were compared. The initial one was the standard ontology created and used in the calculation of GCS. The second was the standard ontology with the 'roles' removed. The standard ontology was then enriched by adding details of dominant forest types in each country to see if this improved the definitions of forest by making them more intuitive and better correlated with the estimated similarities.

In contrast to previous work using SIM-DL (e.g. Janowicz et al. (2008a,c)), this work compares different conceptualizations of the *same* feature (e.g. forests in Austria and forests in Spain), not different features (e.g. river and lake (Janowicz et al. (2008a)).

3.8 Conclusion

This chapter established administrative boundaries as the most appropriate structuring method for microtheories in the INSPIRE scenario. It introduced the geographic features to be treated and detailed the methodology to be followed in achieving the aims of this work. Rules were defined governing the calculation of the good common subsumer between local microtheories to create an appropriate global level microtheory and the principles of similarity reasoning to be performed were outlined.

Chapter 4

Application and Results

This section describes the application of the methodology, outlined in the previous chapter, to the scenarios of rivers, forests and estuaries to calculate an appropriate Europe-wide definition for each feature and test the merits of the microtheory approach. The outcomes of similarity reasoning performed on each feature are then described.

4.1 Rivers

> *The law is not so much carved in stone as it is written in water, flowing in and out with the tide.* Melvoin (1992)

It was decided to focus on the scenario of rivers as their conceptualization (Mark (1993); Pires et al. (2005)) and formalization (e.g. Santos et al. (2005); Mallenby (2007); Hart et al. (2007); Bennett et al. (2008) and Janowicz et al. (2008a)) has been well treated in the literature. In addition, they are highly sensitive to climatic and environmental factors and thus vary in character across Europe.

The traditional northern European perspective of rivers is of a continuously flowing body of water which may also be navigable (Taylor and Stokes (2005)). This view is reflected in the *INSPIRE Consolidated UML Model*'s[1], Hydrography theme. In this classification *river* is not explicitly defined. Instead, the broader class *watercourse* is defined as 'A natural or man–made flowing watercourse or stream'[2]. The European *Water Framework Directive* (WFD)[3] is a Europe-wide water legislation governing the

[1] `http://inspire-twg.jrc.ec.europa.eu/inspire-model/` Generated 24 August 2009 (v3, Revision 873)

[2] `http://inspire-registry.jrc.ec.europa.eu/registers/FCD/items/105`

[3] Directive 2000/60/EC, `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32000L0060:EN:NOT`.

quality, monitoring and reporting of all inland and coastal waters. The WFD definition for river is also included in the INSPIRE FCD as it is regarded as a candidate spatial object for Annex III theme 'Area Management...'. The WFD defines a river as '[a] body of inland water flowing for the most part on the surface of the land but which may flow underground for part of its course'[4]. The British Ordnance Survey (BOS) Hydrology Ontology[5] adopts a more detailed definition but still requires flowing water while the Environment Ontology (EnvO)[6] only contains the restriction that rivers contain water.

These definitions seem broad, however, their requirement of flowing water, or at least water, may be too specific to encompass some rivers in the Mediterranean climes of southern Europe - especially taking the effects of global warming into account. For example, some rivers in southern Spain are highly ephemeral and may contain water (let alone flowing water) only for short periods coinciding with floods or strong rainfall events. In these regions, the conceptualization of river includes channels or depressions through which water flows, even if they are dry and do not show distinct banks (Lamaro et al. (2007)).

4.1.1 Natural Language Definitions

Rivers in Spain and Germany are defined below in natural language. The terminology used is in accordance with that found in the WFD or documents from the relevant national governing body. Fundamental differences between the two definitions are obvious.

Spanish River A river in Spain is defined as a channel, with a bed and more or less defined banks, which transects a river basin at a low point in the topography. It drains water which falls as precipitation on the river basin. It has a flow regime which refers to the average presence or absence of water within the channel throughout a year. It may participate in flood events and provides the ecological service of protecting against these events. Spanish rivers can provide terrestrial or aquatic habitat and terrestrial or aquatic recreational areas and play the role of supplying water and transportation.

German River A river in Germany is defined as a channel, with a river bed and river banks, which contains flowing water, transects a river basin and has another waterbody as its destination. It represents the above ground expression of the groundwater table

[4] http://inspire-registry.jrc.ec.europa.eu/registers/FCD/items/421
[5] http://www.ordnancesurvey.co.uk/oswebsite/ontology/index.html
[6] http://www.environmentontology.org/

and also drains water, from precipitation or snow melt, in the river basin. It may participate in flood events and provides the ecological service of protecting against these events. German Rivers provide aquatic habitat and aquatic recreational areas and play the role of supplying water and transportation.

4.1.2 Conceptual Models

Figures 4.1 and 4.2 show the conceptual models created for rivers in Spain and Germany with differences highlighted in red. The temporal presence of flowing water in Spanish rivers is expressed by making the *Contains.Water* relationship optional (shown red in Figure 4.1). Less importance is placed on the presence of river banks by making the *hasPart.RiverBanks* relationship also optional. German rivers are defined as having a waterbody as their destination but this is not required of Spanish rivers. German and Spanish rivers were defined as having precipitation and ground water as sources of water with Germany having snow melt as an additional source.

Natural disasters like flood events (and drought events in Spain) are of the utmost management importance and thus were included in the conceptual models. Erosion is also a process of particular environmental and management importance and so was included in the conceptual models.

The definitions deliberately avoid reference to rivers being artificial or natural as these terms are vague and cause confusion. For example, natural rivers can have artificial components (e.g. bank stabilization measures) or an artificial flow regime (e.g. due to the presence of a dam etc...). While these distinctions may help distinguish between some features (e.g. canal and river...), they should not be of primary importance to the definition of river.

4.1.3 Formalization

It was necessary to keep the formal definitions as simple as possible for reasoning. Thus, not all the relations depicted in Figures 4.1 and 4.2 were included in the formal ontology. Figure 4.3 shows how Spanish and German rivers were defined.

Current reasoners only work on concepts which are 'defined', i.e. those whose restrictions are necessary and sufficient or 'equivalent' in Protégé terminology. Restrictions on class membership were made using the existential qualifier, meaning 'at least one' or 'some', to link the property and the filler. For example, *GermanRiver hasPart.RiverBed* describes the individual or set of individuals that have at least one

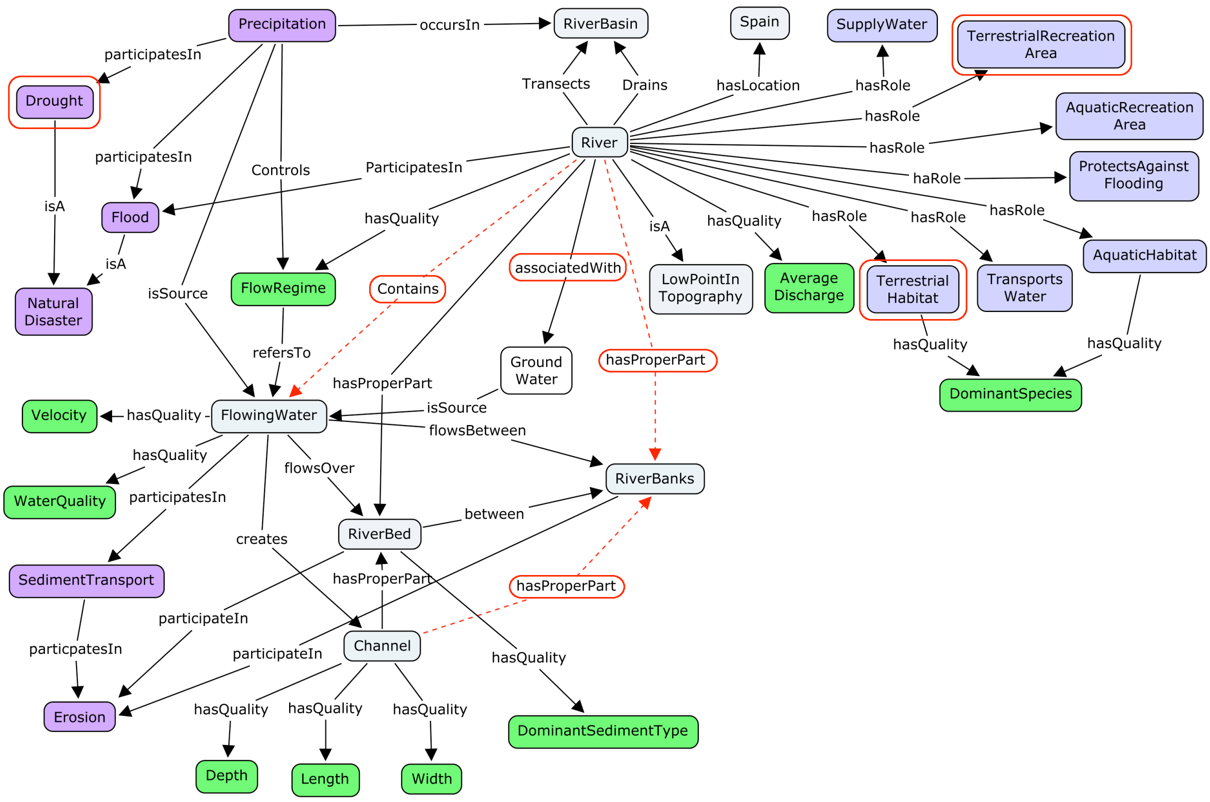


Figure 4.1: Conceptual model showing the relations between entities defining a Spanish River. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue) and quality (green). Red dotted lines indicate optional relations. Elements of difference from the German river definition are circled in red.

part that is an individual from the class *RiverBed*.

The *hasRole* relation, used in the conceptual models to provide more intuitive and anthropogenically meaningful definitions, is not suitable for use with the reasoner. Instead the roles were added as named classes to the concept of rivers in each country. The *union* (OR) operator was used to relate these named classes, meaning that each river may or may not fall into that category (i.e. perform that role) and the river can be a member of more than one category. For example, the disjunction of roles means that a *GermanRiver* can be a *RecreationalArea* and/or an *AquaticHabitat* and so on.

Restrictions that are obligatory for membership in a class were described using the *intersect* (AND) operator. The semantics of the optional relations used for *Contains.Water* and *hasPart.RiverBanks* in the Spanish river definition could not be formalized. Hence, these restrictions, present in the conceptual model, were excluded from the Spanish river definition in OWL. Given the open world assumption of OWL, by

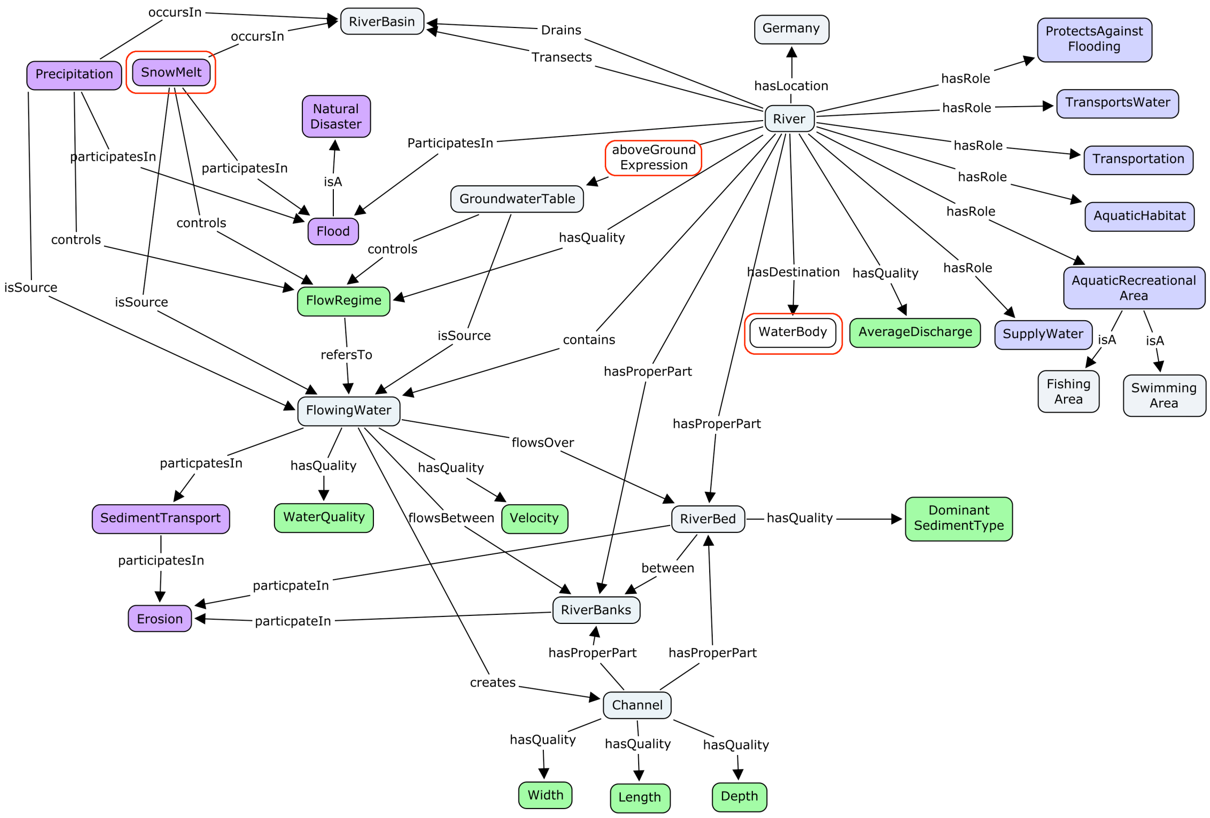


Figure 4.2: Conceptual model showing the relations between entities defining a German River. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue) and quality (green). Elements of difference from the Spanish river definition are circled in red.

leaving out that restriction Spanish rivers can contain flowing water but do not have to.

Entities were not declared *disjoint* (i.e. if an individual is a member of a class it cannot also be a member of a class that is declared disjoint from that class) because, given their broad scale, a single geographic feature can have multiple conceptualizations, multiple links and may fall within one or more member states. The *domain* and *range* of properties were not specified as doing so caused unintended classification results.

4.1.4 Good Common Subsumer

A good common subsumer for Spanish and German rivers was calculated, according to the rules presented in section 3.6.1, and used to define a global-level, Europe-wide definition which could be suitable for INSPIRE. All restrictions common to both definitions were included in the global-level definition, *RiverEurope* (refer to Figure 4.3).

As the Spanish definition lacks the *Contains.FlowingWater* and *hasPart.RiverBanks* restrictions, these were excluded from the European river definition. Vivification was used when the union operator was encountered. Thus, the EU definition included the nearest common superclass (NCS) of all the disjunct roles of each river definition. This was a class called *Roles*. The NCS of the sources (precipitation, groundwater and snowmelt) was *Source* and *NaturalDisaster* was the NCS of *Flood'* and *Drought*. The common filler between the two definitions for the *hasLocation* property is *MemberState*. Thus, a European river was defined as:

$RiverEurope \equiv Roles \sqcap \exists hasPart.RiverBed \sqcap \exists transects.RiverBasin$
$\sqcap \exists hasSource.Source \sqcap \exists hasLocation.MemberState$
$\sqcap \exists hasQuality.FlowRegime \sqcap \exists ParticipatesIn.NaturalDisaster$

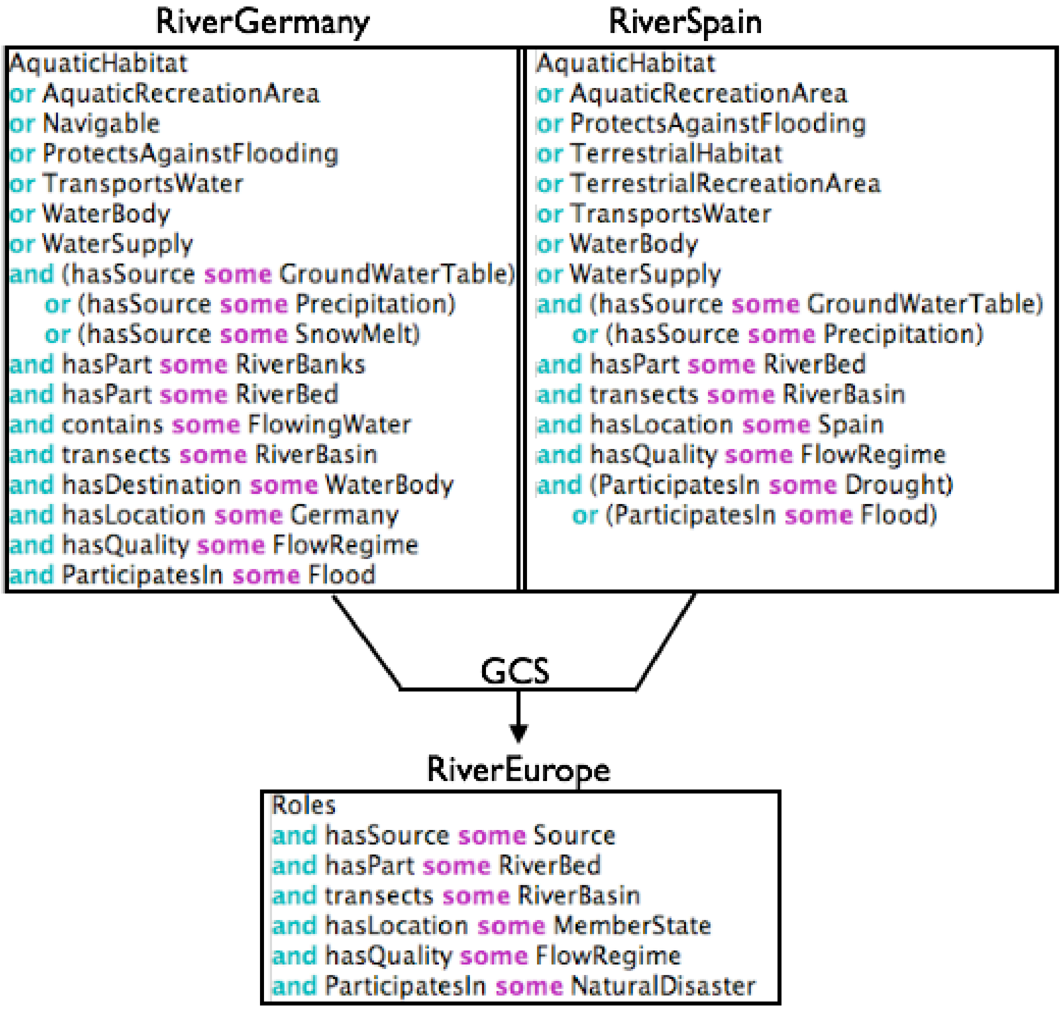


Figure 4.3: Restrictions used to define Spanish and German Rivers and the good common subsumer of these concepts used to define European Rivers.

Classification of inferred classes using FaCT++ showed that the GCS calculation was successful as Spanish and German rivers were subsumed by European river (Figure 4.4). This reasoning also proved that the INSPIRE definition, while it captured the definition of German rivers, was too narrow and did not capture Spanish rivers. The same was true of the BOS Hydrography Ontology and EnvO definitions of river. The EnvO definition was the broadest and was inferred to subsume INSPIRE, BOS and German definitions. The requirement of water in all those definitions was the restriction that did not fit with the character of most Spanish rivers.

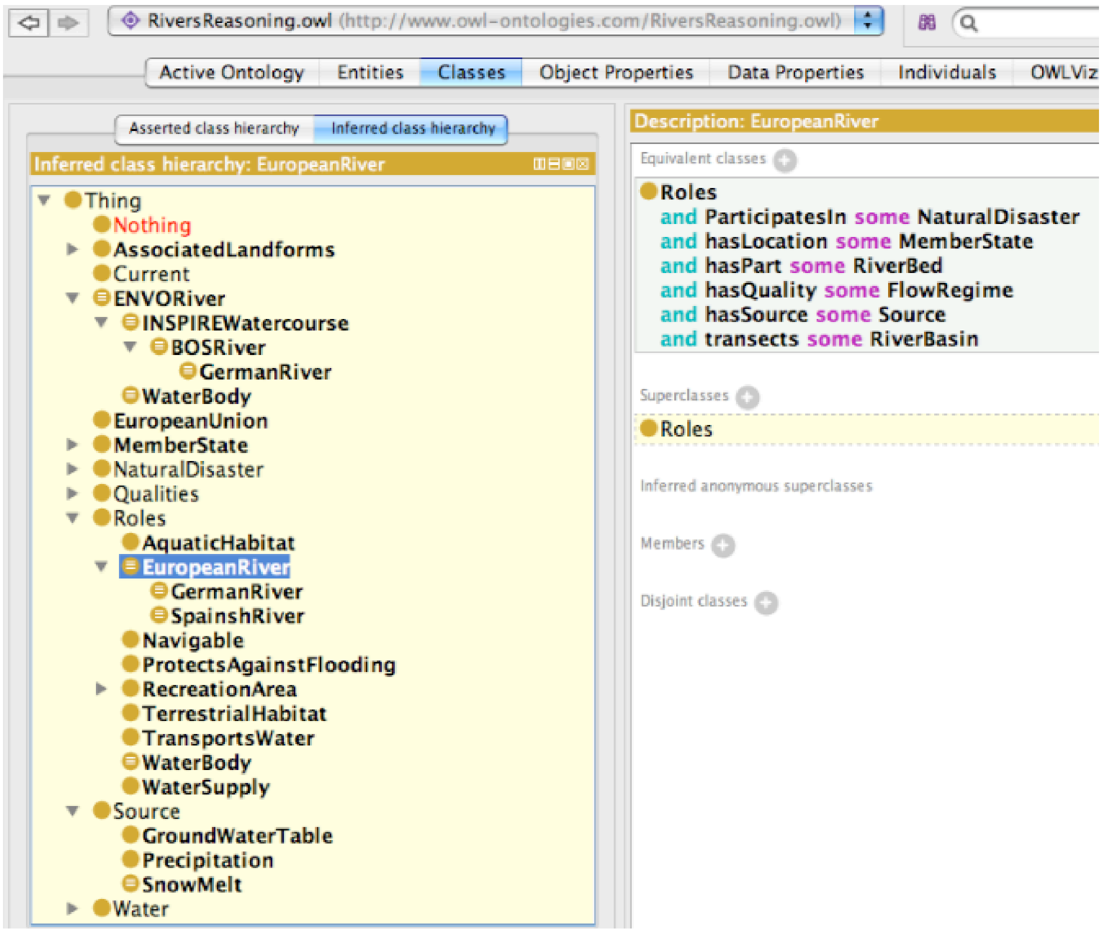


Figure 4.4: Screen shot showing the inferred class hierarchy and the restrictions used to define European Rivers. Please note that the ontology was able to infer that German and Spanish Rivers were both European Rivers based on the definition used. However, INSPIRE Watercourse only subsumes German Rivers as Spanish Rivers cannot satisfy the definition.

4.2 Forests

Forests are extremely diverse and notoriously difficult to define. The UN Food and Agriculture Organisation (FAO) recognized the need to harmonize forest related definitions to facilitate communication and negotiation between international conventions, processes and instruments as well as to reduce the reporting requirements and therefore costs for countries[7]. Bennett (2001) shows that enormous complexity is involved in the adequate ontological representation of forests given the possibility of multiple conceptualizations and the vagueness of their parts and boundaries.

[7] FAO. Process to Harmonize Forest–Related Definitions, http://www.fao.org/forestry/cpf/definitions/en/, Accessed: 16–10–09.

There are more than 950 definitions of forest currently in the literature, all of which differ depending on the perspective of the creator (Lund (2009)). This creates problems when trying to integrate data, monitor, manage and report on resources. The importance of improving access to forestry information has increased in recent years in the light of international commitments towards sustainable forest management, climate change and biodiversity (Schuck et al. (2005)). A vast amount of effort has been, and is being, invested towards the harmonisation of national forest inventories and reporting in Europe alone (e.g. Ministerial Conference on the Protection of Forests in Europe[8], European Forest Fire Information System[9], European Forest Data Centre[10]).

4.2.1 Natural Language Definitions

The elements most commonly used to define what constitutes a forest in legal definitions are minimum area, width, canopy cover and tree height requirements (Rennolls (2005); FOEN (2007); Neeff et al. (2006)). Distinctions between forests as land use, land cover or administrative areas are also found in definitions (Lund (1998)). Comber et al. (2008) suggested that biological, socioeconomic and spatial/structural terms are of importance in distinguishing between types of forest which are conceptualized as land use or land cover. This distinction is commonly used (and mis-used) in the literature defining forests (Comber and Fisher (2005)). This work avoided distinguishing between land use, land cover by attempting to define 'forests' as entities incorporating all elements important to their conceptualization, biological, socioeconomic and structural.

Table 4.1 shows the threshold values used in the legal definition of forest in eight E.U. member states[11] and the FAO.

In addition to these thresholds, forests were defined as performing the following roles (Nasi et al. (2002)); *Carbon Sink, Erosion Prevention, Avalanche Prevention, Increase Soil Infiltration, Recreational Area, Terrestrial Habitat, Commercial Forest* (which play the role of providing timber OR non-timber forestry products) and *Natural Forest*. Forests in Austria, Czech Republic, Denmark, Italy and Switzerland have the additional role of *Avalanche Prevention* while those in Italy, Portugal and Spain may

[8] http://www.mcpfe.org/

[9] http://effis.jrc.ec.europa.eu/

[10] http://efdac.jrc.ec.europa.eu/

[11] Please note that other member states (e.g. Finland) adopt entirely different definitions for forest based on land's capability to produce a minimum volume of stem wood per hectare (Finland Forest Association, 2009. www.forest.fi, Accessed: 12–10–09). This work treats only definitions which follow the more traditional format. However, there is no reason that other definition techniques could not be treated in the same way.

also play the role of *Agricultural Areas* (as they may simultaneously be used for grazing and growing crops like olives or cork) and participate in *Forest Fires*.

Table 4.1: Threshold values used to define forests in selected Euorpean Union countries. Blanks indicate no threshold values are stipulated.

Country (source)	**Min. Width (m)**	**Min. Canopy Cover (%)**	**Min. Area (ha)**	**Min. Mature Tree Height (m)**	**Notes**
Austria (Schuck *et al.* 2002)	10	30	0.05	-	
Czech Republic (Schuck *et al.* 2002)	20	-	0.01	-	Includes land after clear cutting ready for reforestation and forest roads, water surfaces, land above the timber line and other land serving for forest management.
Denmark (Schuck *et al.* 2002)	20	-	0.5	6	
Italy (Schuck *et al.* 2002)	20	20	0.2	-	Includes areas temporarily bare due to cutting or exceptional occurrence.
Portugal (Schuck *et al.* 2002)	15	10	0.2	1.5	Includes areas recently harvested or burnt.
Spain (CMA, Spain 2009)	20	5	0.2	-	There are 3 different categories which may occur over cultivated or pasture land. This is the most liberal of the definitions.
Switzerland (FOEN, Switzerland 2007)	25	20	0.065	3	Afforested, regenerated, burnt, cut or damaged areas are exempt from height and canopy cover restrictions but must have the potential to achieve these levels.
UK (Schuck *et al.* 2002)	50	20	2	-	Includes areas temporally without tress due to forest operations.
FAO (Neeff *et al.* 2006)	20	10	0.5	5	Includes areas normally forming part of the forest area which are temporarily unstocked as a result of human intervention such as harvesting or natural causes but which are expected to revert to forest.

4.2.2 Conceptual Models

The conceptual model for Swiss forest is shown in Figure 4.5. The threshold values used in legal definitions (Table 4.1) were all included as were the roles performed by forests. The conceptual models for forests in all the countries followed the same format and can be found in the Appendix.

Each country's forest was classified as a *NaturalForest* or *CommercialForest*. This split was introduced to reconcile the intuitive assumption that forests have trees, with the clause in most country's legal definitions that a forest need not have trees if it has been cleared for forestry management purposes. Thus, the conceptual model shows *NaturalForests* as having *properPart.Trees* while this restriction is optional for *CommercialForests*. Forest management activities (e.g. harvesting and planting) were included as perdurants for *CommercialForests*. These factors are highly important in reporting and monitoring forests. Their further definition, specifically for each member state, will help to overcome the issues of data harmonization and reporting discussed above. Qualities of trees such as age and species were also included and will help to standardize reporting.

To overcome the fact that a variety of tree species may be present, forests were recognized as having a *DominantSpecies* of tree and it is to these that threshold values are applied. *DominantSpecies* were defined to *havePart.Canopy* which is measured to determine the *CanopyCover* (CC) quality.

The European Environment Agency (EEA) has created a classification of 'forest types' throughout Europe to improve reporting on sustainable forest management (Barbati et al. (2006)). Their classification identifies key-factors which distinguish forest type, including terrain, hydrology, bio-geographic region, geographic distribution, dominant tree type and level of human influence. *ForestType* was identified as a quality of forests and these key factors were included. These factors could be expanded upon to enrich the definitions of forest.

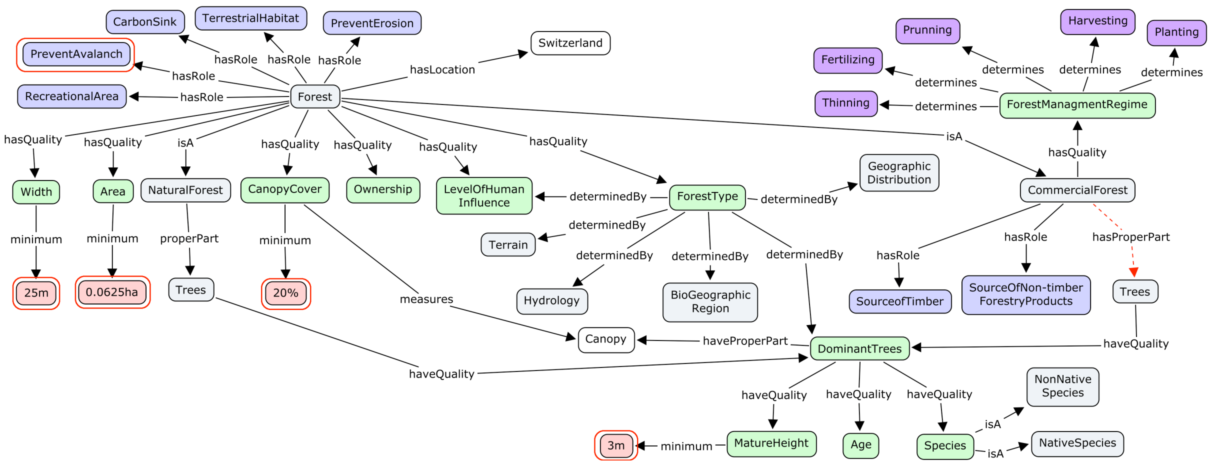


Figure 4.5: Conceptual model showing the relations between entities defining a forest in Switzerland. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue), quality (green). The threshold values for qualities used to define the feature are pink. The red dotted line indicates optional relations and elements of difference from the other forest definitions (refer to Appendix Figures 7.3 to 7.9) are circled in red.

4.2.3 Formalization

The same principles described in section 4.1.3, for rivers, were adhered to in formalizing the definitions of forests. Reification, making a data model for a previously abstract concept, was used wherever possible to make concepts more concrete and thus improve the capability of the reasoner. Subclasses of the *hasQuality* object property for each threshold quality *hasWidth, hasArea, hasMinCC* and *hasMinTreeHeight* were created. To formalize the threshold values used to define forests legally, value partitions were employed (Horridge et al. (2004)). Subsumption was used to make implicit the relative value of each class. For example, the *minimumArea2Hectares* class is subsumed by *minimumArea0.5Hectares* which is subsumed by *minimumArea0.01Hectares*. This allows the reasoner to infer the sizes with respect to each other (Figure 4.6). As was the case for rivers (refer to section 4.1.3), the optional *hasProperPart* relation for trees in the *CommercialForest* class could not be formalized and so was left out of the formal definition.

4.2.4 Good Common Subsumer

A good common subsumer for all the forest definitions was calculated, according to the rules presented in section 3.6.1, and used to define a global-level, Europe-wide definition which could be suitable for INSPIRE. All restrictions common to both definitions were included in the global-level definition. As not all countries forest definitions specify thresholds for minimum canopy cover and minimum mature tree height these restrictions could not be included in the European definition. All definitions employ a minimum width and minimum area threshold. The minimum values among the definitions for width and area, 10 meters and 0.01 hectares respectively (refer to Table 3.1), acted as NCS and were used in the European definition. Vivification was used when the union operator was encountered. Thus, the EU definition included the nearest common superclass (NCS) of all the disjunct roles of each forest definition, *Roles*. The common filler between the definitions for the *hasLocation* property is *MemberState*.

Thus, European forest was defined as:

$$ForestEurope \equiv Roles \sqcap \exists hasArea.MinimumArea0.01Hectare$$
$$\sqcap \exists hasWidth.MinimumWidth10Meters \sqcap \exists hasLocation.MemberState.$$

The definition created for a European forest is shown in Figure 4.6. Classificiation of inferred classes using FaCT++ showed that the calculation was successful as all the member state forest definitions were subsumed by the European forest class (Figure 4.6). This reasoning also showed that none of the country's definitions were subsumed by the FOA definition suggesting that its requirements are too stringent.

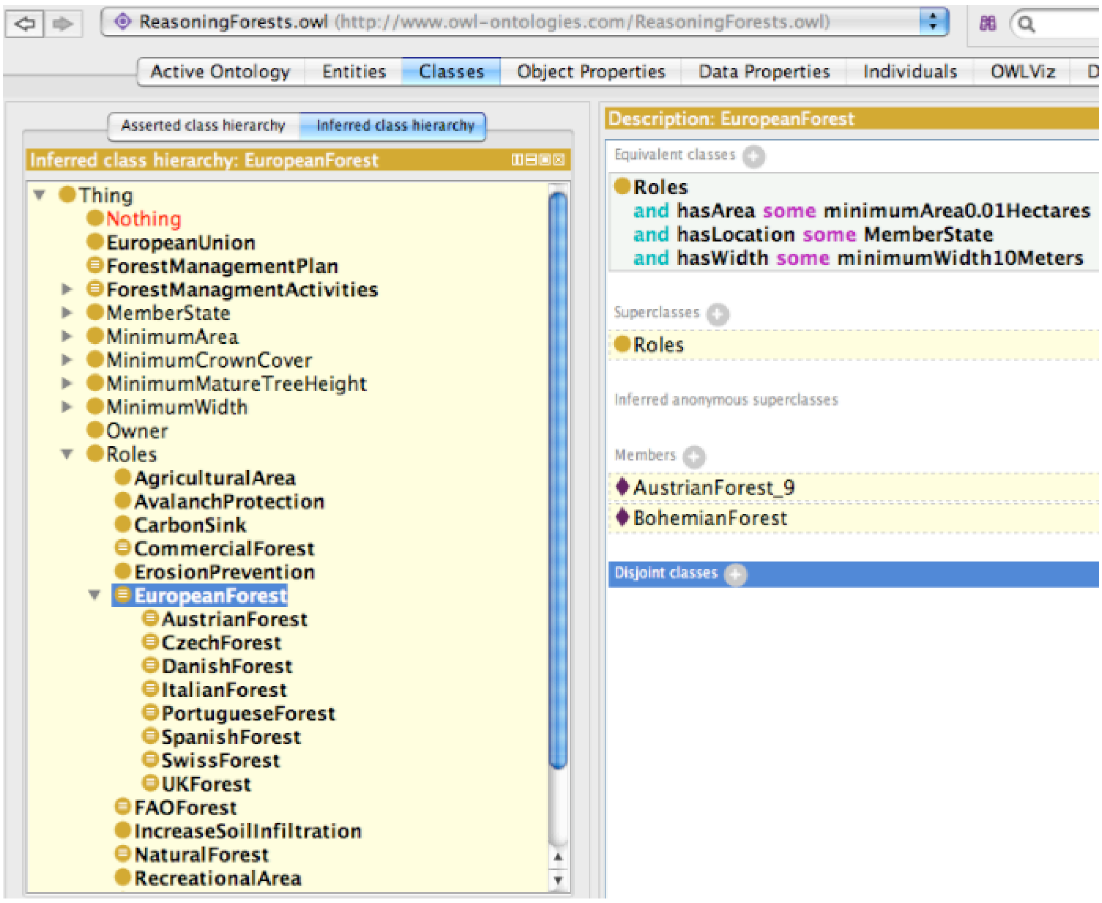


Figure 4.6: Screen shot showing the inferred class hierarchy and the restrictions used to define European forests.

4.3 Estuaries

Estuaries are partially enclosed bodies of water along coastlines where fresh water and salt water meet and mix (Shumchenia (2001)). They are highly dynamic features whose parts and properties constantly vary. Vague adjectives are used to distinguished between different types of estuaries. For example, deep, shallow; high salinity, low salinity; strong flow regime; little tidal mixing and so forth. Unlike the forestry domain, empirical values are not used to define these thresholds in the literature.

4.3.1 Natural Language Definitions

Typical Estuaries in the Netherlands and Norway are defined in natural language (Shumchenia (2001); NOAA (2008)). Vague adjectives used in the definition are shown in italics. INSPIRE does not yet have a definition for estuary and they are only mentioned by default under the protected areas data theme. In the WFD estuaries fall into the *Transitional Water* class and are defined as 'bodies of surface water in the vicinity of river mouths which are partly saline in character as a result of their proximity to coastal waters but which are substantially influenced by freshwater flows'[12].

Netherlands Bar-Built Estuaries Most estuaries in the Netherlands are *shallow* estuaries with ocean and river inputs. Where the two meet, at the mouth, sand bars are built up along the coast by waves and currents. These sand bars limit the exchange of water between the river and ocean and thus there is *low* tidal mixing and *moderate* salinity. The river usually has a *low* flow regime and there is a *moderate* to *macro* tidal regime. Estuaries in the Netherlands provide intertidal, brackish and avian habitat, food, transportation and recreation areas, protect against flooding and storm surge and filter pollutants and sediments from up stream.

Norway Fjord Type Estuaries Most estuaries in Norway are *deep* estuaries with ocean and river inputs. Where the two meet, at the mouth, there is a *shallow* barrier. This barrier limits the exchange of water between the river and ocean and thus there is *low* tidal mixing. The river usually has a *high* flow regime while the tidal range of the ocean is *small*, thus salinity is *very low*. They provide freshwater, brackish and avian habitat, food, transportation and recreation areas and protect against flooding and storm surge.

4.3.2 Conceptual Models

Figures 4.7 and 4.8 show the conceptual models created for estuaries. The important elements of these models are those that distinguish the features, depth, tidal range, tidal mixing, flow regime, salinity and formation process. Different roles are attributed to the estuaries. As very little mixing occurs in Norwegian estuaries they were not recognized as having the role of filtering pollutants and sediments.

The BOS Hydrography Ontology defines estuaries as *isPartOf.River* which *hasDestination.Sea*. This implies that water flows from the river to the sea. While this is true, water from the ocean also enters the estuary and in estuaries with low flow rivers may in fact be the dominant source of water. In this work the *hasInput* property was employed as the relation between estuaries and the ocean and river. This relation is thought to better reflect the dynamics of the relationship. The qualities of *TidalRange*

[12] http://inspire-registry.jrc.ec.europa.eu/registers/FCD/items/424

and river *FlowRegime* then determine the extent of river and ocean influence and the extent of tidal mixing and amount of water salinity.

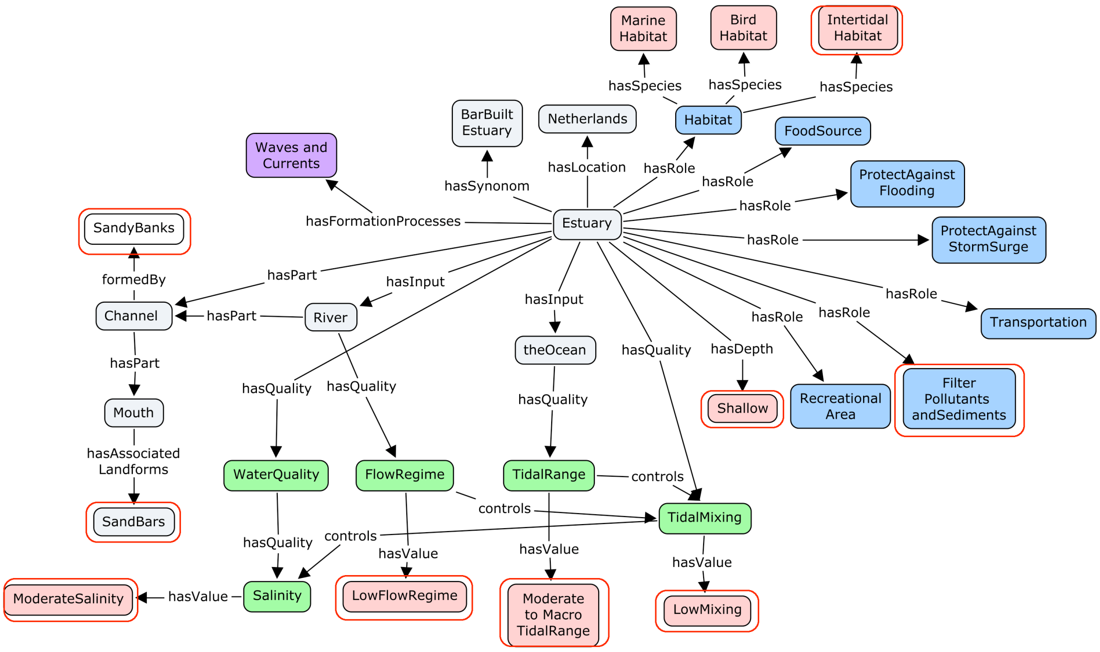


Figure 4.7: Conceptual model showing the relations between entities defining an estuary in the Netherlands. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue) and quality (green). The threshold values for qualities used to define the feature are pink.

4.3.3 Formalization

The same principles described in section 4.1.3, for rivers, were adhered to in formalizing the definitions of estuaries. The different types of estuaries are distinguished using quality thresholds (e.g. tidal regime, tidal mixing etc...). However, unlike the forestry domain where thresholds are given numeric values (e.g. 30 percent crown cover) in the estuary domain they are described using vague adjectives.

This work employes a non-restrictive approach using value partitions where subsumption was used to infer the relative value of each class. For example, *Depth* has subclasses of *Deep* which is subsumed by *ModeratelyDeep* which is subsumed by *shallow*. *HighSalinity* is subsumed by *MediumSalinity* and *LowSalinity* and so forth. This method allows nuances and subtleties between the conceptualization of these adjectives in each country to be maintained. It also allows for temporal fluctuation in values which is common in these features (e.g. an estuary that is usually very saline can become

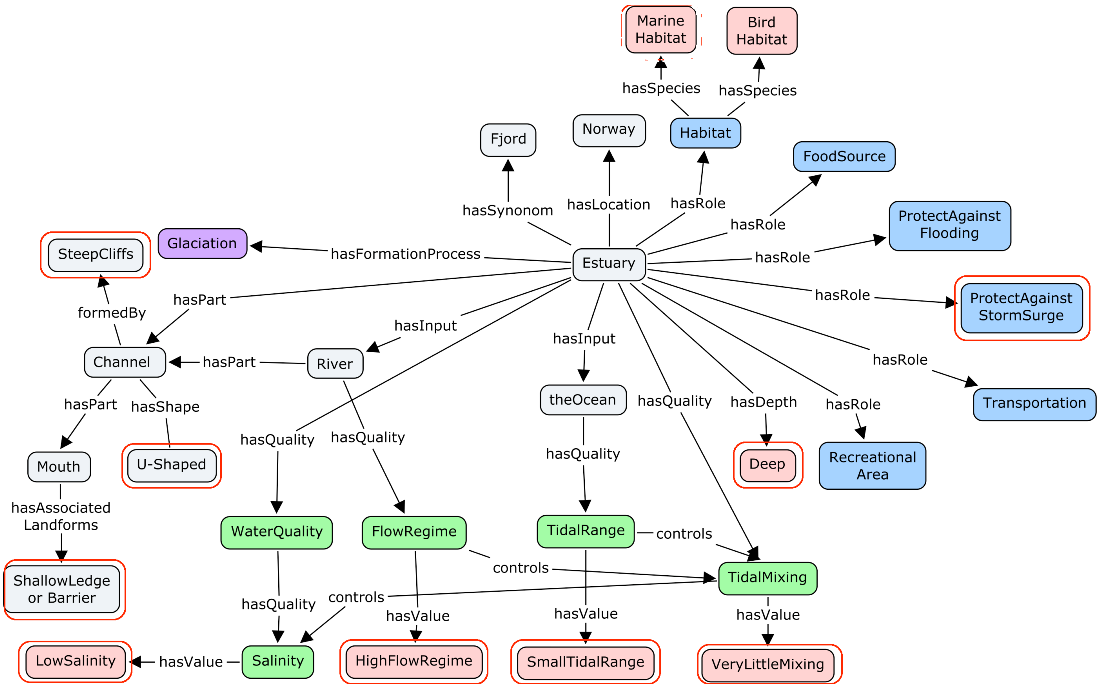


Figure 4.8: Conceptual model showing the relations between entities defining an estuary in Norway. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue) and quality (green). The threshold values for qualities used to define the feature are pink.

almost fresh during flood events).

Reification was employed, making subclasses of the *hasQuality* object property for each threshold (e.g. *hasTidalRange* and *hasSalinity*). These properties were used to restrict the different estuaries with the threshold fillers discussed above (Refer to Figure 4.9). The BOS definition of estuary was also formalized as:
$BOSEstuary \equiv hasDestination.Ocean \sqcap hasPart.River \sqcap isSubjectTo.Tide.$

4.3.4 Good Common Subsumer

A good common subsumer for the local estuary definitions was calculated, according to the rules presented in section 3.6.1, and used to define a global-level, Europe-wide definition which could be suitable for INSPIRE. All restrictions common to both definitions were included in the global-level definition (refer to Figure 4.9). Vivification was used when the union operator was encountered. Thus, the EU definition included the nearest common superclass (NCS) of all the disjunct roles of each estuary definition, *Roles*. The common filler between the two definitions for the *hasLocation* property is

Europe (not Member State as Norway is not a member of the EU). Each of the thresholds was defined to the NCS that quality as the common filler of the definitions (e.g. $hasTidalRange.TidalRange$).

Thus, European estuary was defined as:

$$EstuaryEurope \equiv Roles \sqcap \exists hasInput.(Ocean \sqcap \exists hasTidalRange.TidalRange) \exists hasInput.River \sqcap \exists hasFlowRegime.FlowRegime \sqcap \exists hasPart.(Channel \sqcap \exists formedBy.Banks \sqcap \exists hasAssociatedLandforms.AssociatedLandforms \sqcap \exists hasPart.Mouth)) \sqcap \exists hasLocation.Europe \sqcap \exists hasSalinity.Salinity \sqcap \exists hasTidalMixing.TidalMixing \sqcap \exists hasEstuaryType.EstuaryType \sqcap \exists hasDepth.Depth \sqcap \exists hasFormationProcess.FormationProcess.$$

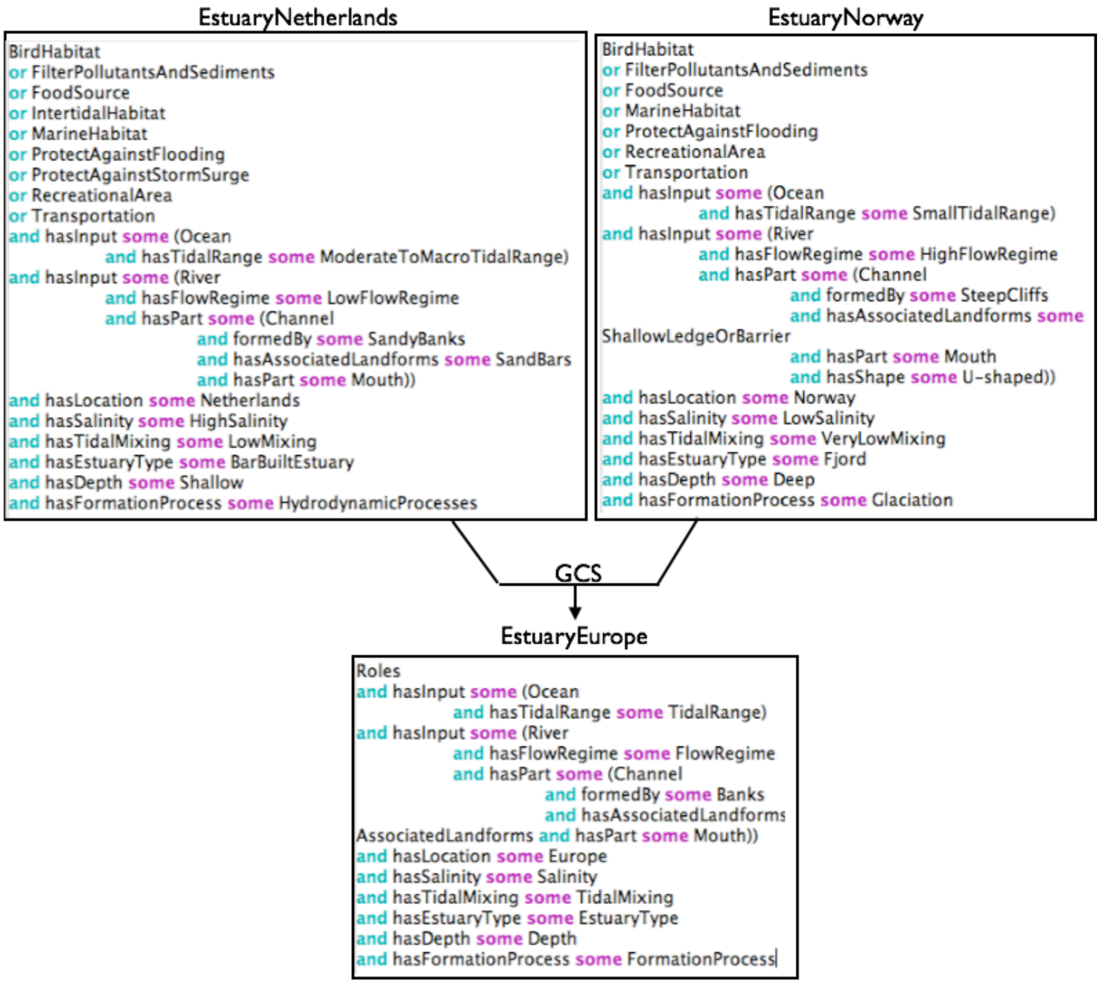


Figure 4.9: Restrictions used to define Norwegian and Dutch estuaries and the good common subsumer of these concepts used to define European Estuaries.

Classification of inferred classes using FaCT++ showed that the calculation of GCS was successful as Dutch and Norwegian estuaries were subsumed by the *EuropeanEstuary* class (Figure 4.10). The reasoner also showed that the *BOSEstuary* was classified as a subclass of *Ocean* due to OWL's open world assumption and the fact that both were defined as being *subjectTo.Tide*. The BOS definition did not subsume either Norway's or the Netherland's estuary definitions.

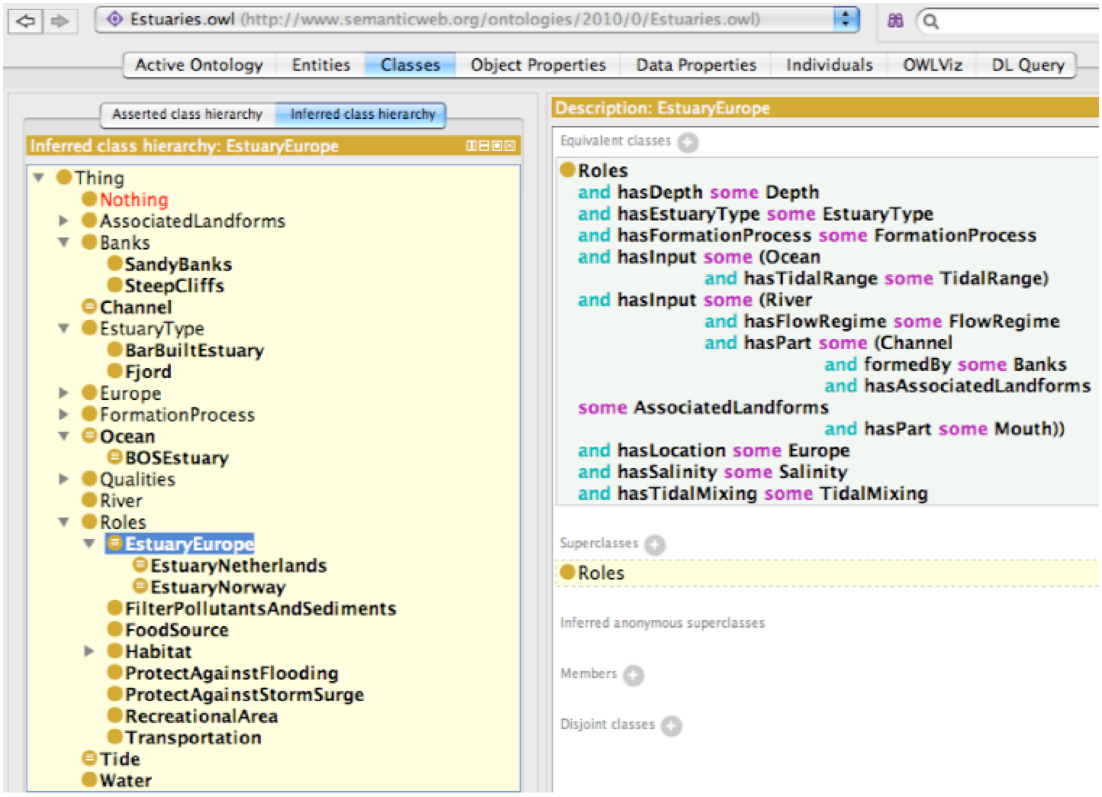


Figure 4.10: Screen shot showing the inferred class hierarchy and the restrictions used to define European estuaries.

4.4 Similarity Reasoning

This section describes the results of similarity reasoning performed on the microtheories using the SIM-DL plugin for Protégé. It also compares the estimated and calculated similarities for each feature type scenario.

4.4.1 Rivers

The concept of European rivers was compared with that of German and Spanish rivers and the INSPIRE watercourse class. As expected, from the perspective of the European river definition, the similarity to the Spanish and German definitions was 1.0. This means that users of a semantics-enabled interface for Web gazetteers or Web discovery service in INSPIRE, searching for *EuropeanRivers* will be satisfied retrieving both Spanish and German rivers. This is not surprising as the EU wide definition is the super concept of both. Comparison to the INSPIRE's watercourse definition resulted in a very low (0.01) similarity to the EU definition. This is to be expected as SIM-DL measures the conceptual overlap between two concepts and the INSPIRE definition does not contain several statements made for the EU wide definition (e.g. *Roles* and *RiverBed*). Note, however, that creating the EU definition based on the GCS of more member states could conceivably broaden the EU river definition.

When SIM-DL reasoning is run in a different direction i.e. comparing the *INSPIREWatercourse* concept to the other river definitions. The INSPIRE definition was found to have a similarity of 1.0 to the German definition but only 0.5 to the European and Spanish definitions. Thus, while results for German rivers would certainly be retrieved by a search using this definition, Spanish and other rivers may not. These results demonstrate that INSPIRE's definition is inadequate to represent rivers throughout Europe as its requirement for flowing water places too much restriction on membership.

4.4.2 Forests

The concept of *EuropeanForest*, created from the GCS, was compared with the forest concepts of the 8 member states and the FAO. As expected, the results showed the European forest definition had a similarity value of 1.0 to the definitions of all member states. This means that users of a semantics-enabled interface for Web gazetteers or Web Discovery Service in INSPIRE, searching for *EuropeanForest* will be satisfied retrieving forest data from all of the member states included. This is not surprising as the EU wide definition is the super concept of all of them.

When the FAO forest concept was compared with the rest, none of the definitions had a similarity of 1.0. The highest similarity was with Swiss and Danish forests (0.84 and 0.83 respectively). These three definitions shared all the same restrictions although different threshold values were adopted (refer to Table 4.1). The least similar were Czech forests (0.32) and the EU forest definition (0.19). As they lacked restrictions for minimum canopy cover and tree height. These results are in keeping with the definitions for each category shown in Table 4.1 and show that the FAO definition does not serve as a good global definition as its requirements are too stringent.

The concept of forest in Austria was compared with the other member states and the European definition created from the GCS. The results showed the Austrian forest definition was most similar (0.9) to Swiss and UK forests (refer to Table 4.2). Portuguese, Spanish and Danish forests had similarity values of 0.87, 0.84 and 0.81 (respectively) when compared to Austrian forests. The definition of Czech forest had a similarity of only 0.71. The similarities estimated by the author are shown in Table 4.3. These are poorly correlated (-0.058) (Figure 4.12) and differ by a total of 131 (refer to Table 4.3) with the similarities calculated by SIM-DL suggesting that the ontology does not capture the human conceptualization of the domain well. This is most likely due to the use of legal definitions rather than factors which are more likely to intuitively define the conceptualization of forests.

To provide some indication of whether the use of 'roles' (e.g. *CarbonSink*, *AvalanchProtection* etc...)in the definitions helped them to better approximate the human conceptualization of forests the similarity reasoning was run again, this time without including the roles in the definition. As shown in Table 4.3, the correlation between estimated and calculated similarity was the same (-0.058) but the total difference was lower (123) than in in the definitions which included the 'roles'. This suggests that including roles did not improve the onology's ability to capture the human conceptualization of the domain.

In an attempt to improve the ontology's ability to capture the human conceptualization of forests in each of the member states the definitions were enriched by including information about the dominant forest type in the area. The frequency of EEA Forest Types (Barbati et al. (2006)) was examined and the most frequently occurring type in each country was defined formally in OWL (Refer to Table 4.2 and Figure 7.10 in the Appendix which shows the frequencies and forest types used). Each forest type was defined based on biogeographic region, dominant species, hydrology, level of human influence and terrain (refer to Barbati et al. (2006)). The dominant forest type in each country was made equivalent to the definition of *Forest* in each country (refer to Figure 4.11).

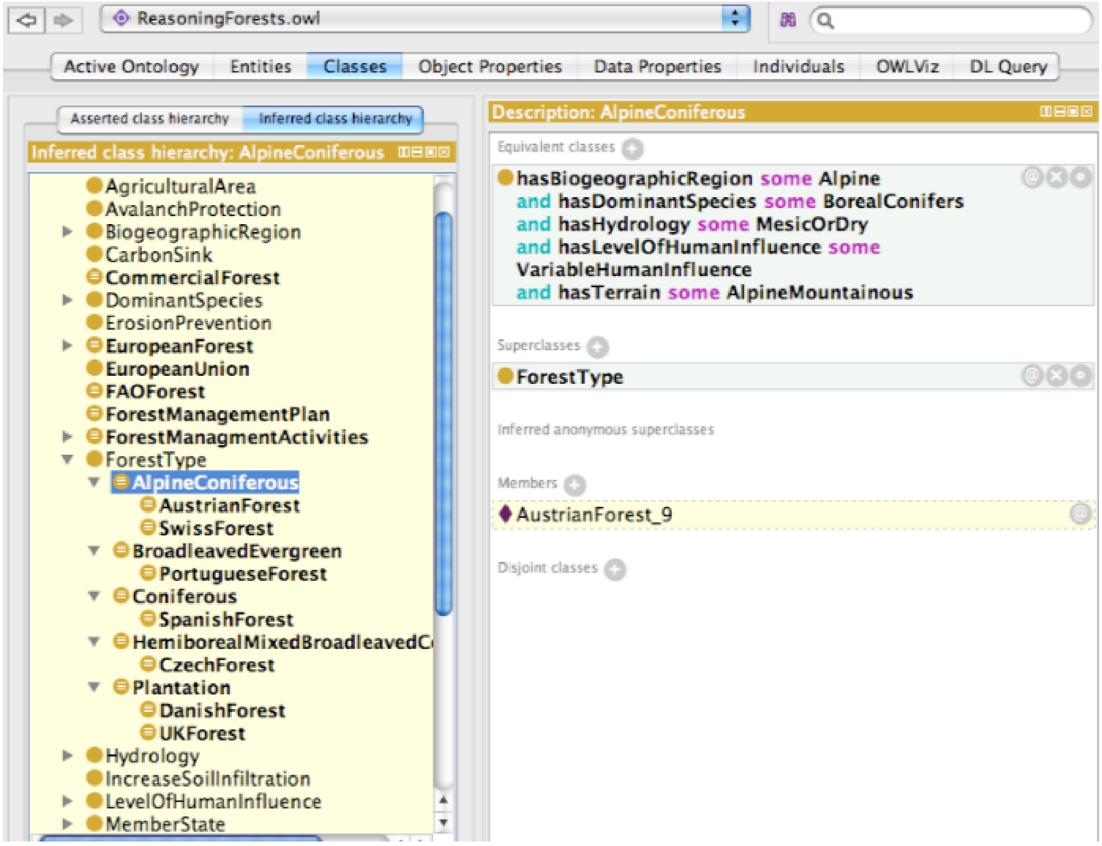


Figure 4.11: Screen shot showing the inferred class hierarchy and the restrictions used to define the Alpine Coniferous Forest Type.

Similarity reasoning comparing the estimated forest similarities and the similarity values calculated using the enriched definitions showed an improved, though still not strong, correlation (0.086) and considerably lower total difference (81) (refer to Table 4.3). As shown in Figure 4.12 the enriched definitions improved the correlation between estimated and calculated similarity values for Danish and UK forests and slightly improved those of Portugal and Spain. Nevertheless, the lowest correlations between estimated and calculated similarities were for Spanish and Portuguese forests. In both cases, the estimated similarity was considerably lower than the calculated similarity. This suggests that, either the estimated conceptualization is inaccurate, or the formal definition does not well represent the condition of forests in these countries. It may be important to note that the dominant forest types in Portugal and Spain accounted for less than 50 percent of forests in those countries (Refer to Table 4.2). It is possible that the forest types used are not dominant enough to accurately characterize the perception of forests in these countries. This reveals a shortcoming in the methodology and suggests that in countries with varying forest types a more fined grained structuring method may be required for microtheories.

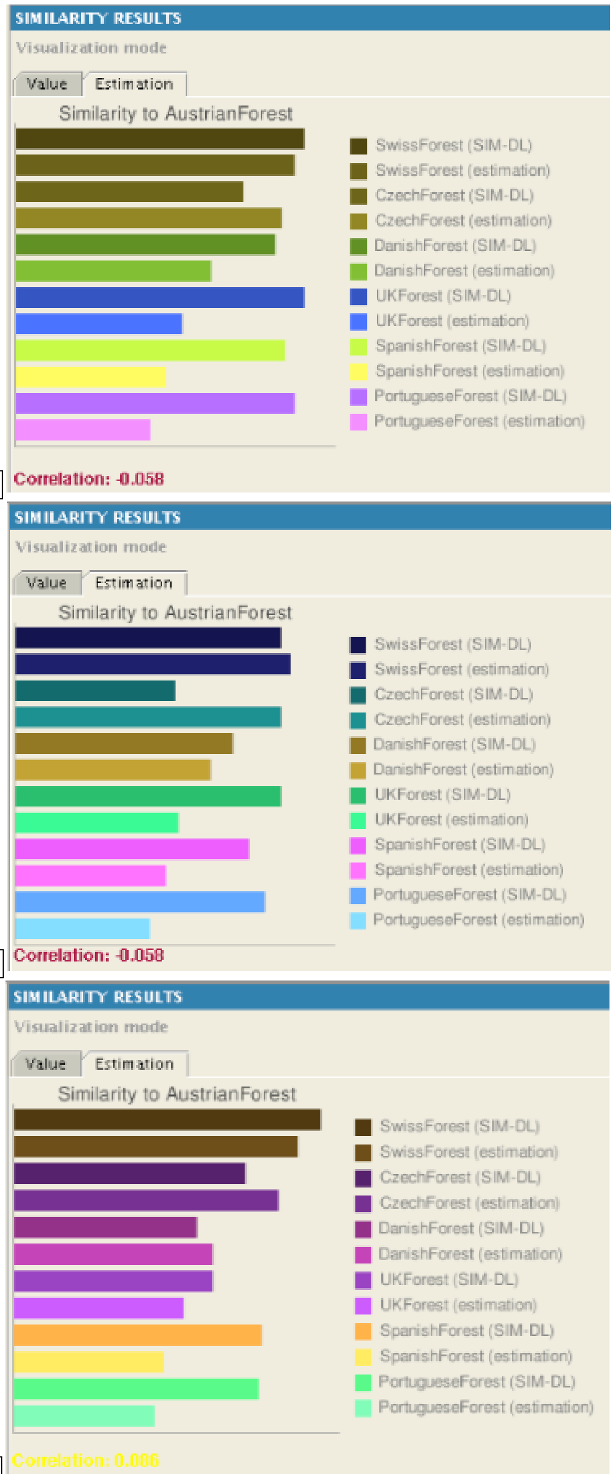
Table 4.2: EEA Forest type categories used in this work.

EEA Label	Forest Type Category (shortened version)	Country (frequency of this forest type %)
2.	Hemiboreal forest and nemoral coniferous and mixed broadleaved-coniferous forest (Hemiboreal Mixed Broadleaved Coniferous)	Czech Republic (69)
3.	Alpine coniferous forest	Austria (65), Switzerland (50)
9.	Broadleaved evergreen forest	Portugal (48)
10.	Coniferous forests of the Mediterranean, Anatolian and Macaronesian regions (Coniferous)	Spain (43)
14.	Plantations and self sown exotic forests (Plantation)	Denmark (60), UK (61)

Source: Barbati *et al.* (2006)

Table 4.3: Table comparing estimated and calculated similarities between Austrian forests and other forests using different definitions ('normal', 'without roles' and 'enriched'). In each case the difference between and correlation between estimated and calculated similarity is presented.

Search Forest Concept	**Target Forest Concept**	**Estimated Similarity (%)**	**Calculated Similarity: Normal Definition (%)**	**Difference**	**Calculated Similarity: Without Roles (%)**	**Difference**	**Calculated Similarity : Enriched Definition (%)**	**Difference**
Austrian	Swiss	95	90	5	83	12	94	1
Austrian	Portuguese	50	87	37	78	28	75	25
Austrian	Czech	90	71	19	50	40	71	19
Austrian	Danish	70	81	11	68	2	56	14
Austrian	Spanish	55	84	29	73	18	76	21
Austrian	UK	60	90	30	83	23	61	1
			Correlation -0.058	**Total 131**	**Correlation -0.058**	**Total 123**	**Correlation 0.086**	**Total 81**



[a] [b] [c]

Figure 4.12: Comparison of the estimated and calculated similarity between different forest microtheories using a) the standard definitions, b) the definitions with 'roles' removed and c) the enriched definitions.

4.4.3 Estuaries

The concept of *EuropeanEstuary*, created from the GCS, was compared with the estuary concepts of the Netherlands and Norway. As expected, the results showed the European estuary definition had a similarity value of 1.0 to both the definitions. This means that users of a semantics-enabled interface for Web gazetteers or Web discovery service in INSPIRE, searching for *EuropeanEstuary* will be satisfied retrieving data from estuaries in both Norway and the Netherlands. This is not suprising as the EU wide definition is the super concept of both. The EU wide definition of estuary was found to have a similarity value of just 0.03 to that of the BOS Hydrology Ontology. This is to be expected as SIM-DL measures the conceptual overlap between two concepts and the BOS definition does not contain several statements made for the EU wide definition (e.g. *Roles*, *AssociatedLandforms* and many of the qualities of estuaries). Note, however, that creating the EU definition based on the GCS of more member states could conceivably broaden the EU estuary definition.

The definition of estuaries in the Netherlands was compared with that of estuaries in Norway and the created European estuary definition. The results showed the Dutch estuary definition had a similarity of 0.73 to the Norwegian and only 0.59 to the European definition created. The similarity with the BOS estuary was only 0.03. However, this can be attributed largely to the fact that it is lacking some restrictions and is structured differently (refer to section 5.3).

These results did not correlate well (correlation: 0.2) with the estimated levels of similarity (Figure 4.13). The calculated similarity between Dutch and Norwegian estuaries (0.73) was considerably higher than the estimated similarity (approximately 0.5) while the opposite was true for the BOS definition in which estimated similarity was substantially higher than calculated similarity. This may reveal shortcomings in the ability of these ontologies to capture human conceptualizations, potentially due to the different importance humans give to different factors (refer to section 5.3). However, as only 2 features were compared, the results do not necessarily indicate how well the microtheories capture the features in each country.

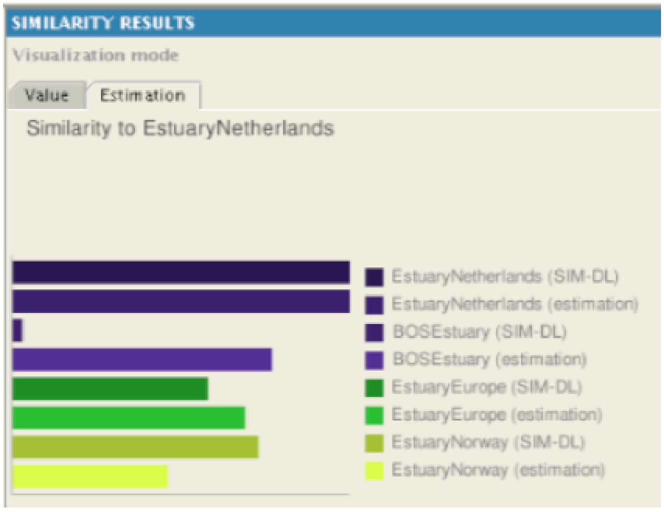


Figure 4.13: Screen shot showing the estimated and calculated similarity between different estuary microtheories.

4.5 Conclusion

This chapter has presented the results of this study and constitutes the main contribution of this work. It has detailed the creation of natural language definitions and detailed conceptual models of rivers, forests and estuaries in different European Countries. It described how these concepts were formalized in OWL and reasoned upon to calculate the good common subsumer. It was shown through inference and similarity reasoning that the GCS provided an appropriate global-level definition to allow interoperability between local definitions. Challenges and shortcomings in the process were also revealed and will be discussed in the following chapter.

Chapter 5

Discussion

To achieve semantic interoperability, local conceptualizations of geographic features must be made explicit and unambiguous. This work suggested the use of microtheories to make local conceptualizations formal and interoperable at a global level. The application of the microtheories technique to three geographic features confirmed the potential of the method and revealed important challenges. This section discusses the findings of this study within the context of related literature. It outlines the advantages of the microtheory approach, discusses some challenges to its implementation and suggests possible ways to overcome these challenges. It also recognizes valuable areas for future research and discusses the practical implications of this work to the INSPIRE initiative.

5.1 Creation of Microtheories to Define Geographic Features

Microtheories can be thought of as small, modular ontologies. Thus, to discuss their potential for use in reconciling local and global definitions from an SDI perspective, it is necessary to discuss the ontological challenges encountered in creating formal microtheories for geographic features.

5.1.1 Restrictions of Formal Language

The restrictions of formal languages, such as Description Logics, make the ontological definition of geographic features difficult (Smith and Mark (1998)). Traditional ontological tools including subsumption and mereotopology are not sufficient to adequately and intuitively define geographic features. The dynamic, interlinked, temporally variable processes require many complex relations which are not possible to formalize in present DL languages. This work revealed that some co-relations and mutual dependencies between form and function, which are inherent to the geographic domain and were present in conceptual models of geographic features, confuse ontologies resulting in unwanted classifications. For example, that flowing water is contained in a river channel and also creates the channel through erosion of the bed and banks was difficult to formalize in a consistent manner.

Also, entities used to define geographic features are often difficult to assign categorically to the DOLCE upper-level classes (Duce (2009)). For example, the most general of distinctions, between endurant and perdurant, is often difficult to make for entities like 'precipitation' which can be seen as a process or an endurant. The most common relation between endurants and perdurants defined in the DOLCE is that of participation (Gangemi et al. (2002); Masolo et al. (2003)). As the relations between endurants and pedurants in the geographic domain are complex, different types of participation should be identified and defined.

An additional barrier to the applied use of ontologies is that the formalization of natural language definitions requires knowledge of ontological engineering which cannot be expected of domain experts. Ideally, natural language could be automatically or semiautomatically translated into formal languages. Achieving this requires the semantics of natural language adjectives to be defined. Hart et al. (2007) have conducted useful work in this direction and have developed 'RABIT' a structured natural language to allow domain experts to formalize their knowledge into ontologies.

5.1.2 Vagueness

Another challenge to ontology in the geographic domain is the formalization of the vague adjectives and vague terms often used to define geographic features. Two different types of vagueness were encountered in this work and are known to hamper the creation of formal definitions for geographic features (Bennett (2001)). These are sorites vagueness and conceptual vagueness.

5.1.3 Overcoming Vagueness

Different methods can be used to overcome each of the different types of vagueness. Conceptual vagueness can be overcome by 'grounding' ontologies. Conceptual vagueness occurs when no single, adequate definition for a term exists (Bennett (2001)). In this work conceptual vagueness was encountered in the concept of *crown cover* which was used in legal definitions of forests. The definitions did not specify which methodology was to be used to determine the percentage of crown cover in an area. Thus, the concept is vague. To remove conceptual vagueness the term, and how it is measured, must be precisely defined. This could be done by grounding the property of interest in reproducible, observation procedures as suggested by Scheider et al. (2009). For example, a strict methodology for the measurement of crown cover could be defined.

Sorites vagueness is the indeterminacy associated with the thresholds at which different properties, e.g. 'deep' and 'shallow', are expressed (Bennett (2001)). Sorites vagueness is present in the definition of many geographic features (Mark (1993)). It is also probable that the conceptualized meaning of these vague terms could differ from country to country depending on many of the same factors which govern the conceptualization of geographic features themselves. In this work sorites vagueness was best demonstrated in the estuary scenario which required the use of many vague adjectives to distinguish Nowegian estuaries from Dutch estuaries (refer to section 4.3). This work employed subsuming value partitions to implicitly define the vague adjectives (refer to section 4.3.3). However, perhaps the semantics of these adjectives could be formalized using supervaluation semantics to set firm thresholds defining a vague term as proposed by Bennett (refer to Bennett (2001); Bennett et al. (2008) for details).

To do this, any conceptual vagueness relating to the property of interest and how it is measured, must be removed by grounding the ontology as discussed above. Once this is achieved, automated knowledge acquisition and data mining techniques (e.g. Musen et al. (2000); Sanz et al. (2008)) could be employed to inductively reason on actual data and deductively determine appropriate values to define thresholds for vague adjectives (Simoudis et al. (1996)).

For example, in the estuary scenario, the adjectives 'deep' and 'shallow' are used to distinguish between estuaries in the Netherlands and Norway. No value ranges for 'deep' or 'shallow' exist in the literature and it is probable that their conceptualization differs depending on cultural background, knowledge, motivation, and especially local conditions. For example, in the Netherlands, an estuary with a mean water depth of 10 meters[1] may be considered 'deep' whereas in Norway (where estuaries may be hundreds of meters deep) this would be considered 'shallow'.

By performing automated text mining of real data pertaining to Norway and Netherlands, the range of values considered 'deep' or 'shallow' in each country could be obtained and used to define these adjectives more meaningfully. A similar approach could be used for salinity and other parameters. Of course there are numerous challenges

[1]Tidal variations in estuaries make the measurement of definitive depth values difficult. For this reason mean water depth is used in this example

to implementing this approach, the availability of sufficient data being one. The, often marked, temporal fluctuation of values for parameters such as depth and salinity driven by factors such as tide and river flow, could also hamper this approach. However, further discussion falls beyond the scope of this work.

Another way to overcome vagueness is to prescribe an arbitrary threshold value. This was the method used in the forests scenarios. Legal definitions were used to provide values for thresholds defining the feature (e.g. minimum crown cover of 20 percent)[2]. There is rarely any documentation of how these threshold values are determined though they reflect the characteristic of forests in each country to a certain extent. For example, Mediterranean forests (i.e. those in Spain and Portugal) have no restrictions for tree height in their definition as tree species are usually small. They also have lower canopy cover requirements as forests in these areas are typically quite sparse.

However, the results of similarity reasoning on forest definitions (Table 4.3 and Figure 4.12 showed poor correlation between the author's estimated similarity of forest in different countries and the calculated similarity based on legal definitions. This suggests that the legal definitions, and the microtheories created based upon them, do not accurately reflect the conditions of forests in those countries.

5.2 Creation of Intuitive Microtheories

It is recognized in the literature that intuitive definitions for geographic features are desirable (Janowicz et al. (2008a)). These should be in accordance with a naive subject's conceptualization of the world (Smith and Mark (2001)) and thus need to be locally specific. The SIM-DL similarity reasoner provides a way to infer the degree to which definitions are intuitive and how well they capture a domain by comparing calculated and estimated similarities (Janowicz et al. (2008a)).

5.2.1 Inclusion of Roles

To make definitions more intuitive and to overcome the problems encountered in formalizing the definition of geographic features discussed above, this work endeavored to use 'roles', based on the ecological functions and services performed by the features. Similarity reasoning was used to approximate whether this use of roles was effective in the forest scenario. The results showed that the inclusion of roles in definitions

[2] Please note that, as mentioned above, definitions did not specify which methodology is to be used to determine the percentage crown cover in an area. Thus, conceptual vagueness creeps into what appear to be firm boundaries.

did not improve the correlation of calculated similarity and estimated similarity (Table 4.3) which suggests that roles do not improve a definition's ability to capture the human conceptualization. However, for reasons discussed below, these results should be considered with caution and not taken to be conclusive.

5.2.2 Inclusion of Feature 'Types'

This work also tested whether more intuitive and meaningful definitions for forests could be obtained with the addition of 'forest type' factors identified by the EEA (Barbati et al. (2006)). This did improve the correlation between estimated and calculated similarities (refer to Table 4.3 and Figure 4.12) though correlation was still poor for Portuguese and Spanish forests. This highlights the fact that forest types are by no means uniform throughout a country and do not follow administrative boundaries.

To overcome this, it is suggested that each of the forest types defined by the EEA (Barbati et al. (2006)) could be formalized as a separate microtheory. Then member states could choose which forest types are present in their country and calculate the GCS between these definitions to determine the country wide definition for forest. The global, Europe wide definition could be created from the GCS of all 14 EEA forest type microtheories. A similar methodology could usefully be applied to the estuary scenario as well. The literature classifies estuaries into different types based on geological features and water circulation (Shumchenia (2001)). Each type could be defined as a microtheory and the GCS methodology presented here used to create an appropriate global level.

5.3 Similarity Reasoning

Similarity reasoning was used in this work to determine the overlap between different conceptualizations (local and global) of the same feature. This provided insight into whether searching for the broad, global level definitions created would satisfactorily return all local definitions of the feature. However, as Janowicz et al. (2008b) point out, perdurants and processes, which are often fundamental to human conceptualizations, are difficult to formalize and so are not taken into account in similarity reasoning.

The BOS definition of estuary had a calculated similarity value of just 0.03 when compared to the European definition (refer to section 4.4.3). This reveals the sensitivity of similarity reasoners to the structure and terminology used in definitions. Current logic languages are unable to capture the true semantics of classes and object properties and thus depend upon structural similarity and similarity of definitions in canonical form. For example, the European definition created here stated that estuaries have *River* and *Ocean* as *inputs*, while the BOS Hydrography Ontology defines estuaries as *partOf.River* which *hasDestination.Sea*. Both are semantically similar (though not the same). However, as different relations and different class names are used in the two definitions they are deemed by SIM-DL to be completely dissimilar. This makes it difficult to compare ontologies which define the same feature but use different class names and object properties.

The similarity reasoning used to infer how well the microtheories captured the human conceptualization of features was useful but should be treated with some caution. Firstly, the estimated values were based upon the author's conceptualizations of forests throughout Europe and were not guided by a forestry expert. For this reason they may be biased and inaccurate and may not provide an ideal approximation against which to judge how well the definitions capture the domain.

Secondly, it is hypothesized here that humans may employ a cognitive weighting of factors when conceptualizing features. For instance, many factors used to define two features can be the same and only one factor different. This would result in high similarity calculated by the reasoner. However, if that one feature of difference is of the most importance to human subjects their estimated similarity will be much lower than the reasoner's calculation. This is supported in the estuaries scenario where estimated similarity between Dutch and Norwegian estuaries was 0.5 but calculated similarity was 0.73 (section4.13). This 'cognitive weighting' hypothesis requires empirical testing through social surveying. If it is confirmed future research should be undertaken into how to 'weight' the elements of a definition to better reflect human intuition and conceptualizations.

5.4 Interoperability between Local Microtheories

Formal microtheories can represent local conceptualizations in an unambiguous way, fundamental to achieving semantic interoperability. The work presented here showed (often great) contrast between the definitions of the same geographic feature in different countries. To make these definitions interoperable an appropriate global level definition is required (Uschold (2000)). This work proposed and tested a bottom-up approach based on reasoning between local level definitions for the same feature, Since present DLs do not support the calculation of Least Common Subsumers in complex and disjunct ontologies (Baader et al. (2007)) a different methodology was formulated to compute a good common subsumer (GCS).

5.4.1 Good Common Subsumers

This work showed how a good common subsumer, calculated based on firm rules and using vivification to remove disjunctions, could create an appropriate global level definition for geographic features. Calculating the GCS between different microtheories did not reveal any information not present in the local microtheories. However, it did allow interoperability between local definitions and identified fundamental similarities between definitions which are important in distinguishing the feature from other features. The GCS was also broad enough to allow differences in important elements of the definitions to be maintained. For example, in the river scenario, the concepts of river bed and flow regime could be kept in the upper level definition as they are common to rivers everywhere and help distinguish them from other features like lakes or estuaries. However, river banks and flowing water were not common to all rivers and so were excluded so as not to overly restrict membership to the class.

5.5 Implications for INSPIRE

This work demonstrated that the INSPIRE definition for rivers was too narrow and would not capture the conceptualization of most rivers in the south of Spain. However, in other respects the INSPIRE definition could be made more specific. For example, the river definition could include it's relation to river basins which are listed in the INSPIRE Feature Concept Dictionary[3].

Other features treated here have yet to be defined for INSPIRE but there is a danger that they too may be too broad or too specific to accurately represent features throughout Europe. The use of microtheories in a manner similar to that demonstrated here could be very effectively applied to INSPIRE to ensure appropriate global definitions are used.

It is intended that consistent terminology for INSPIRE will be managed in a multilingual glossary. However, this work supports previous assertions in the literature (e.g. Mark (1993)) that direct translation of geographic feature terms from one language to another is undesirable. In order to be successful, features in different languages, with their contextual nuances, should be able to be made interoperable. The microtheory approach suggested here facilitates this.

[3] http://inspire-registry.jrc.ec.europa.eu/registers/FCD/items/409 as of 19–01–10

Administrative boundaries offer a geographically suitable and politically pragmatic structuring solution for microtheories. However, some scale problems are inherent in this method. To overcome these problems, different types of each geographic feature (if types already exist in the literature) could be defined formally and each country or suitable region could chose which best suits them or use the global-level definition. However, countries should still be free to create their own definitions and have them incorporated at the global-level.

As Uschold (2000) points out, when local and global ontologies are used, questions arise as to how updating and maintenance will be performed. In the INSPIRE case, when an country wishes to add a definition or update its existing definition for a feature GCS computation will have to be performed again at the European level, to ensure the global definition is suitable. If managed well, this need not be a major issue.

5.6 Further Applications of Microtheories

As well as the potential to be extremely useful in the creation of successful Spatial Data Infrastructures, the microtheory approach presented here could be applied to aid the interpretation of Volunteered Geographic Information (VGI) or data encountered in broad informal systems like the Semantic Web. Gangemi and Mika (2003) proposed the use of contextual evidence to aid interpretation when a complete theory is lacking in such systems. The bottom-up approach, based on the common subsumers between concepts presented here, could be used to create the upper-level context required.

This work showed the ability of microtheories to provide modularity to ontologies. Further investigation and exploitation of this property would be a valuable line of future research. The use of microtheories to define even more basic concepts than geographic features could be investigated. For example, each element of a feature definition could be its own 'micro' microtheory. A group of these topical microtheories could be put together to define more complex geographic features like rivers or forests. This would require relationships other than generalization and spatial containment to be used between microtheories, but could potentially allow more complex interlinkedness and interrelationships in the geographic domain to be made explicit without compromising the consistency of ontologies.

5.7 Conclusion

This section has discussed the results of the study and the challenges encountered in creating formal yet intuitive definitions of geographic features. It also discussed similarity reasoning and the computation of a GCS between local microtheories. It suggested the possible implications of the research to the INSPIRE initiative and suggested further applications for microtheories.

Chapter 6

Conclusion

This work developed the importance of employing local definitions for geographic features and demonstrated how microtheories, structured by administrative boundaries, can be used to overcome semantic heterogeneity caused by multiple local conceptualizations of geographic features.

The work examined novel and intuitive structuring principles for microtheories including linguistic, cultural, climatic and administrative boundaries. It determined that administrative boundaries were most effective as they are spatially and temporally relevant as well as politically pragmatic. However, results suggested that as countries are large and diverse themselves nation-wide microtheories may be too broad to define geographic features and capture human conceptualizations of them. Two possible solutions to this challenge were identified. Autonomous or independent regions could be free to make their own sub-microtheory for features where necessary and a nation-wide microtheory could be generalised from the internal regions. Or different 'types' of each feature could be defined and countries decide which type or types are present in their territory and calculate the GCS between these definitions to determine the nation-wide definition. The global, Europe wide definition could be created from the GCS of all 'types' present in member states.

This work also discussed challenges in defining geographic features ontologically including the restrictions imposed by formal logic languages and difficulties associated with vagueness in the geographic domain and how they could be overcome. It tested the use of ecological functions and services as 'roles' to formally define features in an intuitive manner. The use of roles was not conclusively found to improve the ontology's capture of human conceptualizations. However, the technique should not be discounted and merits further enquiry.

Rivers, forests and estuaries in different European countries were defined in natural language, as conceptual models and then as formal microtheories. This revealed that numerous different, and often conflicting, elements were used to define the same feature in different countries, reaffirming the importance of maintaining local conceptualizations. The simplifications necessary to formally define features also highlighted gaps in the ability of existing logic languages to deal with the expressive conceptualizations necessary to effectively define vague, dynamic and inextricably linked processes occurring in the geospatial domain and incorporate intuative conceptualizations.

Appropriate feature definitions at a European-level are fundamental to achieving interoperability. This work proposed and tested a bottom-up approach based on reasoning between local definitions of features instead of standardizing common feature types manually – which may exclude local conceptualizations. Given the inability of present reasoners to work with descriptive and disjunct ontologies, this work showed how a good common subsumer, calculated based on firm rules and using vivification to remove disjunctions, could create an appropriate global level definition for geographic features.

Results showed that the present INSPIRE and Water Framework Directive definitions for rivers were too specific in some respects and would not capture the conceptualization of most rivers in the south of Spain. Other features treated here have yet to be defined for INSPIRE but their definitions by other organizations such as the BOS and FAO were found to be too specific to represent conditions throughout Europe. It is recommended that the methodology presented here be replicated by EU member states to create their own microtheories for features, based on existing natural language definitions or based on pre-defined 'feature type' microtheories. The global-level definition could then be calculated semi-automatically using the GCS method presented here.

The structuring of microtheories by administrative boundaries presented here gives initial insights into the role of space and time for ontology modularization (Janowicz (2009)). However, the approach still requires an improved and rigid, formal underpinning. Future work should also be directed towards improving the ability of existing semantic Web representation languages and reasoners to handle the expressive conceptualizations necessary to effectively define geographic features. The potential use of topical microtheories to define smaller components of geographic feature definitions, such as river bed or forest canopy, and the combination of these components in the definition of geographic features also warrants further investigation.

Bibliography

F. Baader and R. Kuesters. Computing the least common subsumer and the most specific concept in the presence of cyclic aln-concept descriptions. In O. Herzog and A. Gunter, editors, *Proceedings of the 22nd Annual German Conference on Artificial Intelligence (KI-98)*, volume 1504, pages 129–140, 1998.

Franz Baader, Baris Sertkaya, and Anni-Yasmin Turhan. Computing the least common subsumer w.r.t. a background terminology. *Journal of Applied Logic*, 5(3):392 – 420, 2007. ISSN 1570–8683. Selected papers from the 9th European Conference on Logics in Artificial Intelligence JELIA 04.

Anna Barbati, Piermaria Corona, and Marco Marchetti. European forest types - categories and types for sustainable forest management reporting and policy. Technical Report 9, European Environment Agency, 2006.

John Bateman, Stefano Borgo, Klaus Luettich, Claudio Masolo, and Till Mossakowski. Ontological modularity and spatial diversity. *Spatial Cognition and Computation*, 7 (1):97–128, May 2007.

B. Bennett, D. Mallenby, and A. Third. An ontology for grounding vague geographic terms. In Carola, editor, *Formal Ontology in Information Systems – Proceedings of the Fifth International Conference (FOIS 2008)*, volume 183 of *Frontiers in Artificial Intelligence and Applications*, pages 280–294, 2008.

Brandon Bennett. What is a forest on the vagueness of certain geographic concepts. *Topoi*, 20(2):189–201, 2001.

Yaser Bishr. Overcoming the semantic and other barriers to gis interoperability. *International Journal of Geographical Information Science*, 12(4):299–314, 1998.

Thomas Bittner, Maureen Donnelly, and Barry Smith. A spatio-temporal ontology for geographic information integration. *International Journal of Geographical Information Science*, 23(6):765–798, June 2009.

J. Brank, M. Grobelnik, and D. Mladenic. A survey of ontology evaluation techniques. In *Conference on Data Mining and Data Warehouses (SiKDD 2005)*, 2005.

Boyan Brodaric and Mark Gahegan. Distinguishing instances and evidence of geographical concepts for geospatial database design. *Lecture Notes in Computer Science*, 2478:22–37, January 2002.

Boyan Brodaric and Mark Gahegan. Experiments to examine the situated nature of geoscientific concepts. *Spatial Cognition and Computation*, 7(1):61 –95, May 2007.

W.W. Cohen, A. Borgida, and H. Hirsh. Computing least common subsumers in description logics. In *In Proceedings of the 10th National Conference on Artificial Intelligence*, pages 754–760. MIT Press, 1992.

A.J. Comber and P.F. Fisher. What is land cover? *Environment and Planning B: Planning and Design*, 32:199–209, 2005.

A.J. Comber, R.A. Wadsworth, and P.F. Fisher. Using semantics to clarify the conceptual confusion between land cover and land use: The example of 'forest'. *Journal of Land Use Science*, 3(2-3):185–198, September 2008.

Max Craglia. Introduction to the international journal of spatial data infrastructures research. *International Journal of Spatial Data Infrastructures Research*, 1:1–13, 2006.

Isabel F. Cruz and William Sunna. Structural alignment methods with application to geospatial ontologies. *Transactions in Geographic Information Science. Special Issue on Semantic Similarity Measurement and Geospatial Applications*, 12(6):683 – 711, 2008.

Cycorp. Cyc 101 tutorial – microtheories. Technical report, Cycorp Inc., 2002a.

Cycorp. Contexts in cyc. Technical report, Cycorp Inc., 2002b. URL http://www.cyc.com/cycdoc/course/contexts-basic-module.html.

Rudolf S. de Groot, Matthew A. Wilson, and Roelof M. J. Boumans. A typology for the classification, description and valuation of ecosystem functions, goods and services. *Ecological Economics*, 41(3):393–408, June 2002.

Dong-Po Deng. Building ontologies of place name for spatial information retrieval. In *GIS Development*. GIS Development, 2007.

"Metadata" Drafting Teams "Data Specifications", "Network Services". Inspire technical architecture - overview. Technical report, INSPIRE Drafting Teams, 2007.

Stephanie Duce. Towards an ontology for reef islands. In Krzysztof Janowicz, Martin Raubal, and Sergei Levashkin, editors, *GeoSpatial Semantics*, volume LNCS 5892, pages 175–187. Springer-Verlag Berlin Heidelberg, 2009.

M.J. Egenhofer and D. M. Mark. Naive geography. In Andrew U. Frank and Werner Kuhn, editors, *Spatial Information Theory: A Theoretical Basis for GIS*, volume LNCS 988, pages 1–15. Springer-Verlag, Berlin, 1995.

FOEN. Switzerland's initial report under article 7, paragraph 4 of the kyoto protocol. Technical report, Federal Office for the Environment (FOEN), Switzerland, 2007.

Andrew Frank. A linguistically justified proposal for a spatiotemporal ontology. In Werner Kuhn, M. Worboys, and S. Timpf, editors, *Spatial Information Theory: Foundations of Geographic Information Science*, number 2205 in LNCS. Springer, 2003a. Position Paper in COSIT03 Ontology Workshop.

Andrew U. Frank. *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, chapter Chapter 2. Ontology for Spatio-temporal Databases, pages 9–77. Number 2520 in LNCS. Springer, 2003b.

Andrew U. Frank. Multi-cultural aspects of spatial knowledge. In Krzysztof Janowicz, Martin Raubal, and Sergei Levashkin, editors, *GeoSpatial Semantics*, volume LNCS 5892, pages 1–8. Springer-Verlag Berlin Heidelberg, 2009.

A. Gangemi, N. Guarino, C. Masolo, A. Oltrami, and L. Schneider. Sweetening ontologies with dolce. *Lecture Notes in Computer Science*, 2473, 2002.

Aldo Gangemi and Peter Mika. Understanding the semantic web through descriptions and situations. In Robert Meersmand, Zahir Tari, and Douglas C. Schmidt, editors, *On the Move to Meaningful Internet Systems 2003: CoopIS/DOA/ODBASE*, pages 689–706, 2003.

Carlos Granell, Michael Gould, Miguel Angel Manso, and Miguel Angel Bernabe. Spatial data infrastructures. In Hassan A. Karimi, editor, *Handbook of Research on Geoinformatics*, chapter Chapter V, pages 36 – 41. Information Science Reference, November 2008.

B.C. Grau, Y. Kazakov, and U. Sattler. A logical framework for modularity of ontologies. In *20th International Joint Conference on Artificial Intelligence*, pages 183–196, 2007.

T.R. Gruber. A translation approach to portable ontology specifications. Technical Report KSL 92.71, Knowledge Systems Laboratory, Stanford University, 1993.

N Guarino. Formal ontology in information systems. In N Guarino, editor, *Proceedings of FOIS-98, Trento, Italy*, June 1998.

R. Guha, R. McCool, and R. Fikes. Contexts for the semantic web. In *International Semantic Web Conference (ISWC 2004)*, number 3298 in Lecture Notes in Computer Science, pages 32–46. Springer, 2004.

Glen Hart and Catherine Dolbear. What's so special about spatial? In Arno Scharl and Klaus Tochtermann, editors, *The Geospatial Web*, chapter 4, pages 39–44. Springer, 2007.

Glen Hart, Catherine Dolbear, and John Goodwin. Lege feliciter: Using structured english to represent a topographic hydrology ontology. In *Proceedings of the OWL Experiences and Directions workshop (OWLED 2007)*, 2007.

Joana Hois, Mehul Bhatt, and Oliver Kutz. Modular ontologies for architectural design. In R. Ferrario and A. Oltramari, editors, *Formal Ontologies Meet Industry*, pages 66–78. IOS Press, 2009.

Matthew Horridge, Holger Knublauch, Alan Rector, Robert Stevens, and Chris Wroe. A practical guide to building owl ontologies using the protege-owl plugin and co-ode tools. Technical Report Edition 1.0, The University of Manchester, August 2004.

Ian Horrocks. The fact system. Technical report, University of Manchester, 2003. URL http://www.cs.manchester.ac.uk/~horrocks/FaCT/.

Eduard Hovy. Comparing sets of semantic relations in ontologies. In Rebecca Green, Carol A. Bean, and Sung Hyon Myaeng, editors, *The Semantics of Relationships: An Interdisciplinary Perspective*, chapter Chapter 6. Kluwer Academic Publishers, 2002.

Krzysztof Janowicz. The role of place for the spatial referencing of heritage data. In *The Cultural Heritage of Historic European Cities and Public Participatory GIS Workshop*. The University of York, UK, 17 –18 September 2009.

Krzysztof Janowicz and Carsten Kessler. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 10(22):1129–1157, 2008.

Krzysztof. Janowicz and M. Wilkes. SIM-DL_A: A novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity. In *The 6th Annual European Semantic Web Conference.*, volume Lecture Notes in Computer Science 5554, pages 353–367. Springer, 2009.

Krzysztof Janowicz, Carsten Kessler, M. Schwarz, M. Wilkes, I. Panov, M. Espeter, and B. Baeumer. Algorithm, implementation and application of the sim-dl similarity server. In F. T. Fonseca, A. Rodriguez, and Sergei Levashkin, editors, *Second International Conference on GeoSpatial Semantics (GeoS 2007)*, number 4853 in LNCS, pages 128–145. Springer, 2007.

Krzysztof Janowicz, Patrick Maue, Marc Wilkes, Sven Schade, Franca Scherer, Matthias Braun, Soren Dupke, and Werner Kuhn. Similarity as a quality indicator in ontology engineering. In C. Eschenbach and M. Grueninger, editors, *5th International Conference on Formal Ontology in Information Systems (FOIS 2008)*, pages 92–105. IOS Press, 2008a.

Krzysztof Janowicz, Martin Raubal, Angela Schwering, and Werner Kuhn. Semantic similarity measurement and geospatial applications. *Transactions in Geographic Information Science*, pages 1–10, 2008b.

Krzysztof Janowicz, M. Wilkes, and M. Lutz. Similarity-based information retrieval and its role within spatial data infrastructures. In T. J. Cova, H. J. Miller, K. Beard, A. U. Frank, and M. F. Goodchild, editors, *5th International Conference on Geographic Information Science*, volume LNCS 5266, pages 151–167. Springer, 2008c.

Krzysztof Janowicz, Sven Schade, Arne Broering, Carsten Kessler, Patrick Maue, and Christoph Stasch. Semantic enablement for spatial data infrastructures. *Transactions in Geographic Information Science*, TBA:TBA, forthcoming.

Ernesto Jimenez-Ruiz, Bernardo Cuenca Grau, Ulrike Sattler, Thomas Schneider, and Rafael Berlanga. Safe and economic re-use of ontologies: A logic based methodology and tool support. In *The Semantic Web: Research and Applications*, volume 5021 of *Lecture Notes in Computer Science*. Springer Berlin, May 2008.

Eva Klien. *Semantic Annotation of Geographic Information*. PhD thesis, Institute for GeoInformatics - University of Muenster, 2008.

H. Knublauch, R. Fergerson, N. Noy, and M. Musen. The protege owl plugin: An open development environment for semantic web applications. In *Third International Semantic Web Conference*, pages 229–243, 2004.

Werner Kuhn. Ontologies in support of activities in geographical space. *International Journal of Geographical Information Science*, 15(7):613–631, October 2001.

Werner Kuhn. Geo–ontologies for semantic interoperability. Presentation at the Workshop on Geo–Ontology, Ilkley, UK, 2002.

Werner Kuhn. Geospatial semantics: Why, of what and how? *Journal of Data Semantics*, III:1–24, 2005.

Werner Kuhn. A functional ontology of observation and measurement. In Krzysztof Janowicz, Martin Raubal, and Sergei Levashkin, editors, *GeoSpatial Semantics*, volume LNCS5892, pages 26–43. Springer-Verlag Berlin Heidelberg, 2009.

Yang Kun, Wang Jun, and Peng Shuang-yun. The reseach and practice of geo-ontology construction. In *Proceedings of the International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion*, 2005.

Elizabeth Lamaro, Robert Stokes, and Mark Patrick Taylor. Riverbanks and the law: The arbitrary nature of river boundaries in new south wales, australia. *The Environmentalist*, 27(1):131–142, 2007.

H. Gyde Lund. Definitions of forest, deforestation, afforestation, and reforestation, 1998. URL http://home.comcast.net/~gyde/DEFpaper.htm.

H. Gyde Lund. Definitions of forest, deforestation, afforestation, and reforestation, 2009. URL http://home.comcast.net/~gyde/DEFpaper.htm.

D. Mallenby. Handling vagueness in ontologies of geographic information. In *Doctoral Consortium of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR2006)*, Lake District, United Kingdom, 2006.

David Mallenby. Grounding a geographic ontology on geographic data. In *Commonsense 2007 - 8th International Symposium on Logical Formalizations of Commonsense Reasoning*, 2007.

David M. Mark. Toward a theoretical framework for geographic entity types. In I. Campari and Andrew U. Frank, editors, *Spatial Information Theory A Theoretical Basis for GIS*, volume 716, pages 270–283. Springer, Heidelberg, 1993.

David M. Mark. The cognition of landscape features: Ontology, cross-language variation, and feature extraction. Guest Lecture at Vespucci Summer School, 2009.

David M. Mark and A. G. Turk. Landscape categories in yindjibarndi: Ontology, environment and language. *Lecture Notes in Computer Science*, –(2825):31 – 49, 2003.

C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltrami. Ontology library deliverable d18. Technical report, ISTC-CNR, 2003.

J. McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 12: 1030–1035, 1987.

Deborah L. McGuinness and Frank van Harmelen. Owl web ontology language overview. w3c recommendation. Technical report, World Wide Web Consortium, 2004. URL http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Jeff Melvoin. Northern exposure, crime and punishment, 1992. URL http://www.quotationspage.com/quotes/Jeff_Melvoin/.

Mark A. Musen, Ray W. Fergerson, William E. Grosso, Natalya F. Noy, Monica Crubzy, and John H. Gennari. Component–based support for building knowledge-acquisition systems. In *In Proceedings of the Conference on Intelligent Information Processing of the International Federation for Information Processing World Computer Congress*, pages 18–22, 2000.

R. Nasi, S. Wunder, and A. J.J. Campos. Forest ecosystem services: Can they pay our way out of deforestation? Technical report, CIFOR for the Global Environmental Facility (GEF), 2002.

Douglas D. Nebert, editor. *Developing Spatial Data Infrastructures: The SDI Cookbook*. Global Spatial Data Infrastructure, 2.0 edition, 2004.

Till Neeff, Heiner von Luepke, and Dieter Schoene. Choosing a forest definition for the clean development mechanism. Forests and Climate Change Working Paper 4, Food and Agriculture Organisation, 2006.

I. Niles and A. Pease. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, October 2001.

NOAA. Classifying estuaries – by geology, 2008. URL http://oceanservice.noaa.gov/education/kits/estuaries/estuaries04_geology.html.

Paulo Pires, Marco Painho, and Werner Kuhn. Measuring semantic differences between conceptualisations: The portugues water bodies case - does education matter? In *On the Move to Meaningful Internet Systems 2005: OTM Workshops*, volume Lecture Notes in Computer Science 3762, pages 1020–1026. Springer Berlin, 2005.

Keith Rennolls. A partial ontology for forest inventory and mensuration. In *2nd International Workshop on Forest and Environmental Information and Decision Support Systems*, pages 679–683. IEEE Computer Society Press, 2005.

M. Andrea Rodriguez and Max. J. Egenhofer. Determining semantic similarity among entitiy classes from different ontologies. *Knowledge and Data Engineering*, 15(2): 442–456, March/April 2003.

Paulo Santos, Brandon Bennett, and Georgios Sakellariou. Supervaluation semantics for inland water feature ontology. In *Proceedings of the 19 International Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 564–569. Edinburgh, 2005.

I. Sanz, M. Mesiti, G. Guerrini, and R. Berlanga. Fragment-based approximate retrieval in highly heterogeneous xml collections. *Data and Knowledge Engineering*, 64(1): 266–293, 2008.

S. Scheider, K. Janowicz, and W. Kuhn. Grounding geographic categories in the meaningful environment. In K. S. Hornsby, C. Claramunt, and G. Ligozat, editors, *Spatial Information Theory, 9th International Conference, COSIT 2009.*, volume LNCS 5756, pages 69–87. Springer, 2009.

Simon Scheider. The case for grounding databases. In Krzysztof Janowicz, Martin Raubal, and Sergei Levashkin, editors, *GeoSpatial Semantics*, volume LNCS 5892, pages 44–69. Springer-Verlag Berlin Heidelberg, 2009.

L. Schneider. Designing foundational ontologies - the object-centered high-level reference ontology ochre as a case study. In *Proc. of Conceptual Modeling*, Chicago, Illinois, USA, 2003. Springer.

A. Schuck, G. Andrienko, S. Folving, M. Kohl, S. Miina, R. Paivinen, T. Richards, and H. Voss. The european forest information system - an internet based interface between information providers and the user community. *Computers and Electronics in Agriculture 47*, 47:185–206, 2005.

Emily Shumchenia. Estuarine science, 2001. URL http://omp.gso.uri.edu/ompweb/doee/science/intro.htm.

Pavel Shvaiko and Jerome Euzenat. Ten challenges for ontology matching. Technical Report DISI-08-042, Dipartimento di Ingegneria e Scienza dell'Informazione, University of Trento, August 2008.

Evangelos Simoudis, Brian Livezey, and Randy Kerber. *Integrating Inductive and Deductive Reasoning for Data-Mining*, pages 353–373. American Association for Artificial Intelligence, 1996.

Barry Smith. Mereotopology: A theory of parts and boundaries. *Data and Knowledge Engineering*, 20:287–303, 1996.

Barry Smith. Ontology: An introduction. In L. Floridi, editor, *Proceedings of the Blackwell Guide to the Philosophy of Computing and Information*, pages 155–166. Oxfort: Blackwell, 2003.

Barry Smith and David M. Mark. Ontology and geographic kinds. In *International Symposium on Spatial Data Handling (SDH'98)*, pages 308 – 320. Vancouver, Canada, July 1998.

Barry Smith and David M. Mark. Geographical categories: An ontological investigation. *International Journal of Geographical Information Science*, 15(7):591–612, 2001.

Barry Smith and David M. Mark. Do mountains exist? towards an ontology of landforms. *Environment and Planning B: Planning and Design*, 20(2):411–427, 2003.

J. Sowa. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Pacific Grove, CA, 2000.

Andrew D. Spear. *Ontology for the Twenty First Century: An Introduction with Recommendations*. Institute for Formal Ontology and Medical Information Science, Saarbruecken, Germany, 2006.

M. Taylor and R. Stokes. Up the creek: What is wrong with the definition of a river in new south wales? *Environment and Planning Law Journal*, 22(3):193–211, 2005.

Axel Thomas and Ute C. Herzfeld. Regeotop: New climatic data fields for east asia based on localized relief information and geostatistical methods. *International Journal of Climatology*, 24(10):1283–1306, August 2004.

Waldo Tobler. A computer model simulating urban growth in the detroit region. *Economic Geography*, 46(2):234–240, 1970.

Michael Uschold. Creating, integrating and maintaining local and global ontologies. In W Horn, editor, *Proceedings of 14th European Conference on Artificial Intelligence (ECAI'00), Berlin*, Amsterdam, 2000. IOS Press.

Piek Vossen. Welcome to the kyoto project, 2008. URL www.kyotoproject.eu.

W3C. Web ontology language (owl), October 2007. URL http://www.w3.org/2004/OWL/.

I. Wachsmuth. The concept of intelligence in ai. In *Prerational Intelligence – Adaptive Behavior and Intelligent Systems without Symbols and Logic*, volume 1. Springer, 2000.

Chapter 7

Appendix

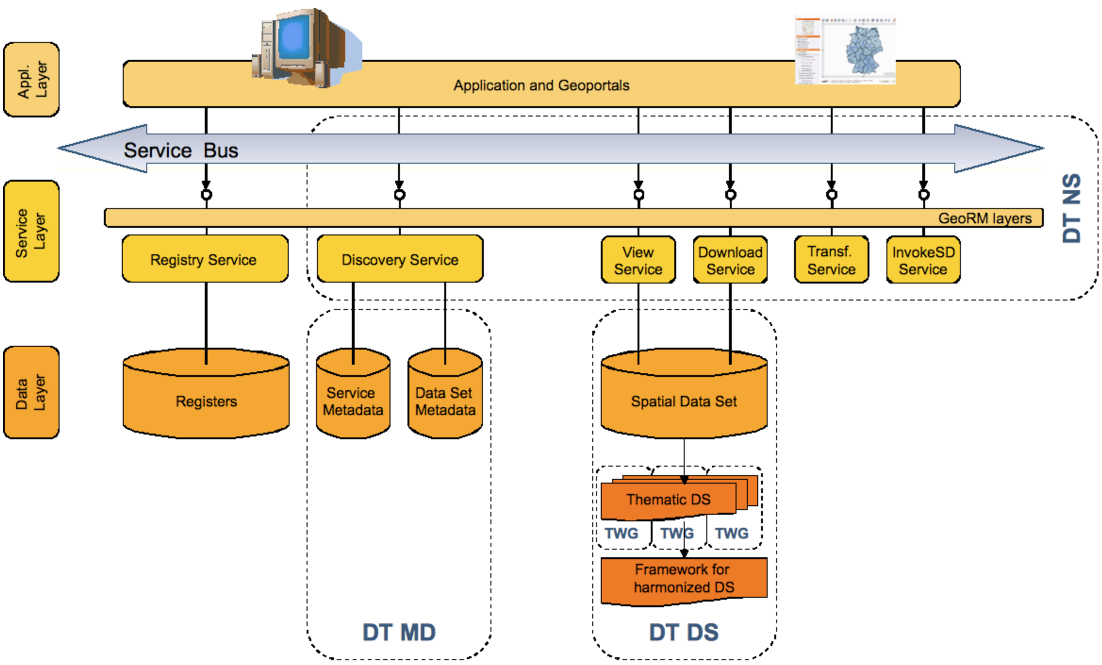


Figure 7.1: INSPIRE Technical Architecture Overview. Source: Drafting Teams "Data Specifications" (2007)

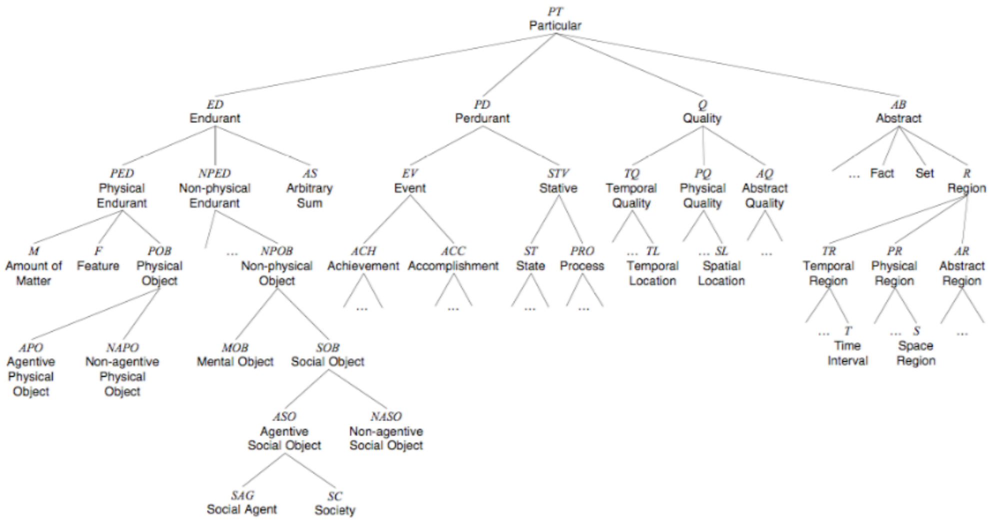


Figure 7.2: The DOLCE top-level classes. Source: Masolo et al. (2003)

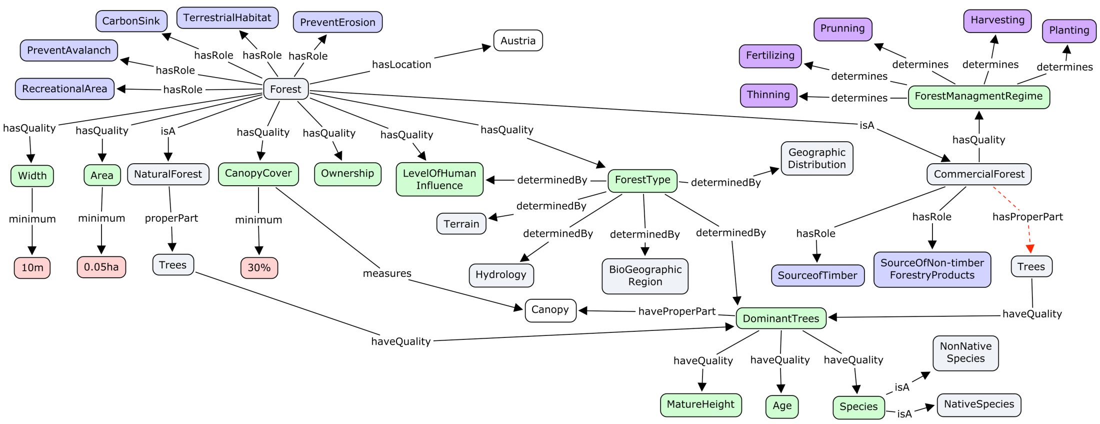


Figure 7.3: Conceptual model showing the relations between entities defining a forest in Austria. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue), quality (green). The threshold values for qualities used to define the feature are pink. The red dotted line indicates optional relations.

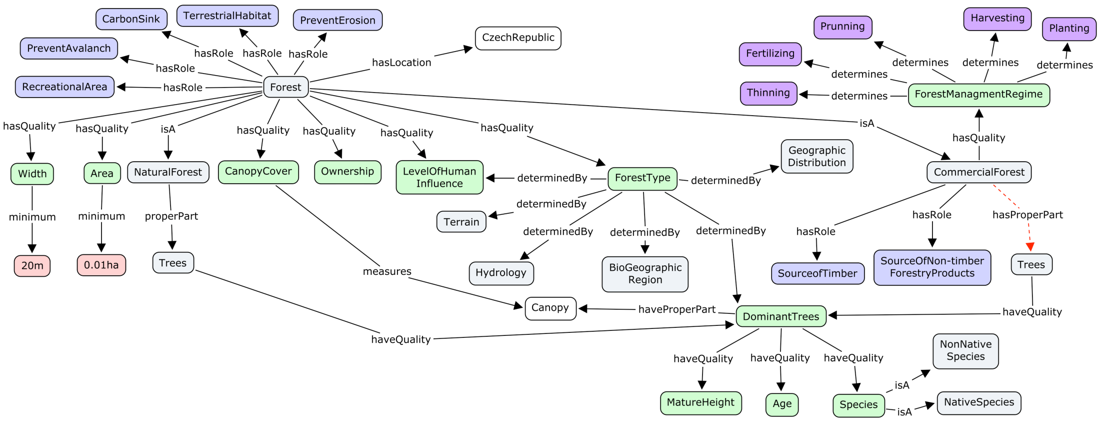


Figure 7.4: Conceptual model showing the relations between entities defining a forest in the Czech Republic. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue), quality (green). The threshold values for qualities used to define the feature are pink. The red dotted line indicates optional relations..

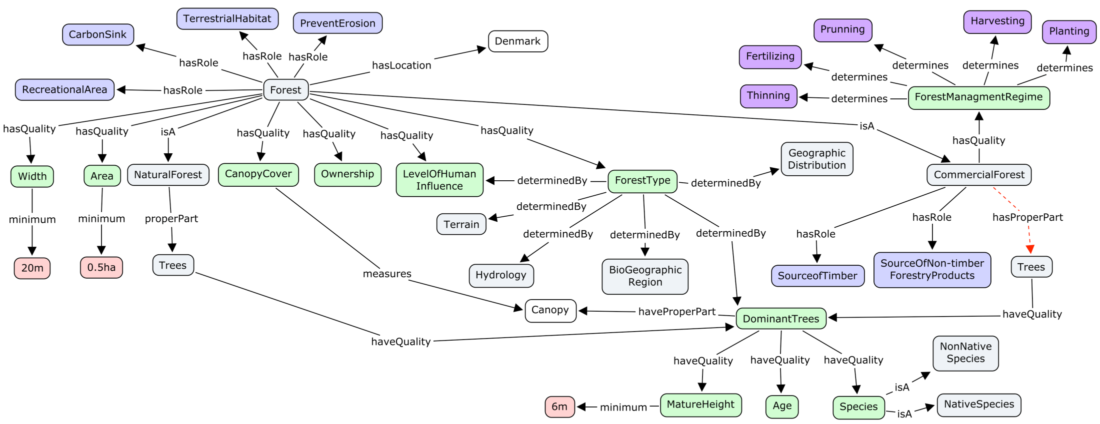


Figure 7.5: Conceptual model showing the relations between entities defining a forest in Denmark. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue), quality (green). The threshold values for qualities used to define the feature are pink. The red dotted line indicates optional relations..

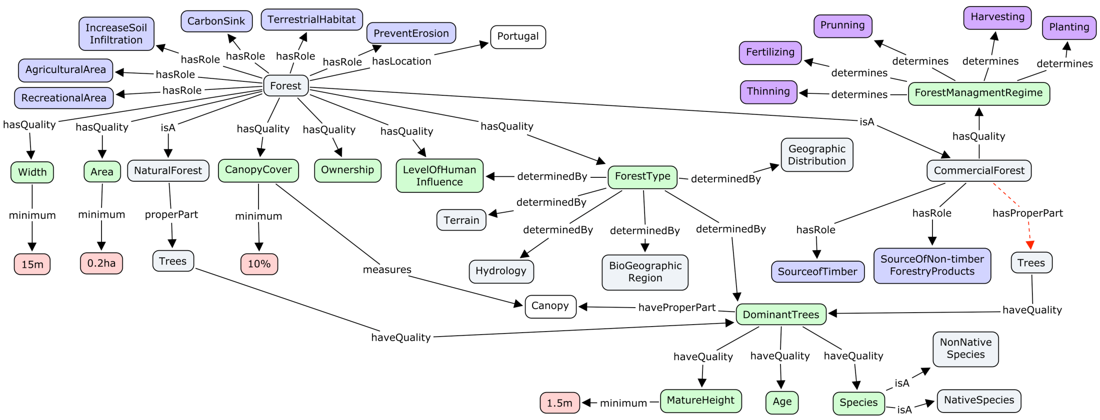


Figure 7.6: Conceptual model showing the relations between entities defining a forest in Portugal. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue), quality (green). The threshold values for qualities used to define the feature are pink. The red dotted line indicates optional relations..

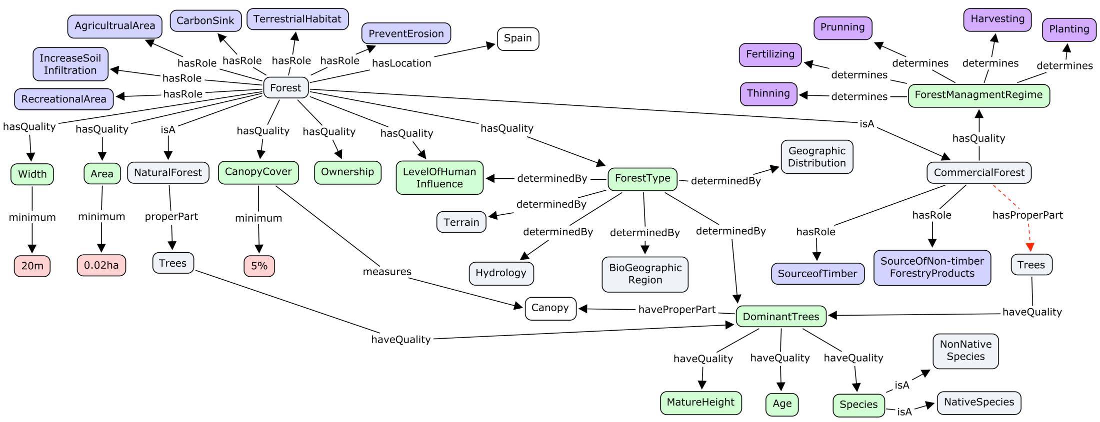


Figure 7.7: Conceptual model showing the relations between entities defining a forest in Spain. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue), quality (green). The threshold values for qualities used to define the feature are pink. The red dotted line indicates optional relations..

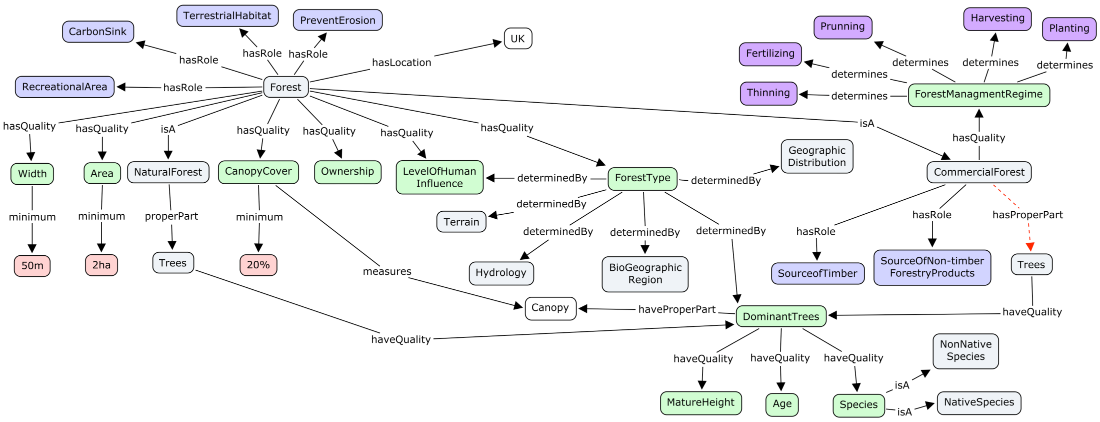


Figure 7.8: Conceptual model showing the relations between entities defining a forest in the UK. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue), quality (green). The threshold values for qualities used to define the feature are pink. The red dotted line indicates optional relations..

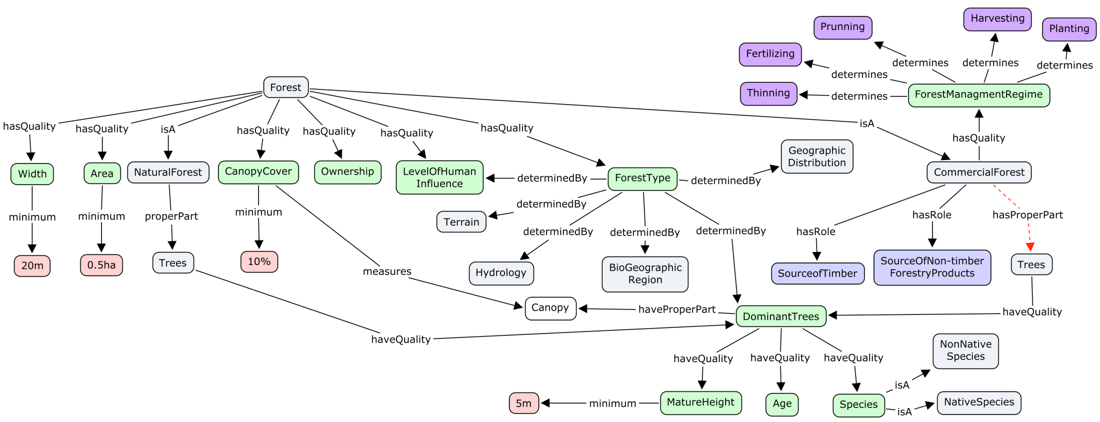


Figure 7.9: Conceptual model showing the relations between entities defining a forest by the FAO of the UN. The entities are divided roughly into the DOLCE top-level classes: physical endurant (white), perdurant (purple), role (blue), quality (green). The threshold values for qualities used to define the feature are pink. The red dotted line indicates optional relations..

	Category (% of the ICP level I plots)														
Country (no of ICP level I plots)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**Total Country (%)**
Andorra (3)	0	0	100	0	0	0	0	0	0	0	0	0	0	0	100
Austria (136)	0	24	65	0	3	1	6	0	0	0	0	1	0	0	100
Azores (6)	0	0	0	0	0	0	0	0	67	33	0	0	0	0	100
Belarus (405)	8	62	0	0	1	0	0	0	0	0	13	1	14	0	100
Belgium (10)	0	0	0	10	10	0	0	0	0	0	0	0	0	80	100
Bulgaria (103)	0	8	26	0	5	12	7	17	0	17	0	0	0	8	100
Canaries (13)	0	0	0	0	0	0	0	0	38	62	0	0	0	0	100
Croatia (84)	0	1	6	2	15	20	11	14	2	6	0	19	0	2	100
Cyprus (15)	0	0	0	0	0	0	0	0	0	100	0	0	0	0	100
Czech Republic (140)	0	69	1	1	9	4	4	0	0	0	0	1	3	9	100
Denmark (20)	0	0	0	0	10	30	0	0	0	0	0	0	0	60	100
Estonia (92)	7	77	0	0	1	0	0	0	0	0	12	0	3	0	100
Finland (595)	88	3	0	0	0	0	0	0	0	0	4	0	6	0	100
France (511)	0	4	9	6	24	7	5	14	4	10	0	1	2	15	100
Germany (451)	0	51	4	1	8	12	6	0	0	0	0	1	3	14	100
Greece (91)	0	0	0	0	9	2	10	19	16	43	0	0	0	1	100
Hungary (73)	0	5	0	0	21	7	0	19	0	0	0	5	7	36	100
Ireland (19)	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100
Italy (255)	0	0	23	1	2	0	16	40	4	4	0	0	3	6	100
Latvia (95)	19	59	0	0	0	0	0	0	0	0	0	0	22	0	100
Lithuania (63)	5	76	0	0	2	0	0	0	0	0	0	0	17	0	100
Luxembourg (4)	0	25	0	25	25	25	0	0	0	0	0	0	0	0	100
Moldova (10)	0	0	0	0	80	0	0	10	0	0	0	0	0	10	100
Neherlands (11)	0	0	0	9	27	0	0	0	0	0	0	0	0	64	100
Norway (442)	68	4	0	0	0	0	0	0	0	0	0	0	27	0	100
Poland (433)	0	75	5	1	7	2	2	0	0	0	0	1	5	1	100
Portugal (133)	0	0	0	0	1	0	0	4	48	29	0	0	1	18	100
Romania (226)	0	1	16	0	21	22	21	10	0	0	0	0	2	6	100
Russia (134)	20	75	0	0	0	0	0	0	0	0	3	0	1	0	100
Serbia (130)	0	2	1	0	11	23	11	35	0	1	0	5	1	12	100
Slovak Republic (108)	0	5	39	0	16	26	10	1	0	0	0	0	0	4	100
Slovenia (42)	0	12	19	0	2	21	29	5	0	0	0	2	2	7	100
Spain (607)	0	0	3	2	2	0	2	9	26	43	0	0	0	12	100
Sweden (775)	50	39	0	1	1	1	0	0	0	0	0	0	6	1	100
Switzerland (48)	0	15	50	0	8	6	13	6	0	0	0	0	0	2	100
United Kingdom (85)	0	4	0	4	16	14	0	0	0	0	0	1	0	61	100
Total countries (6 368)	**20**	**25**	**6**	**1**	**6**	**5**	**4**	**6**	**4**	**7**	**1**	**1**	**6**	**7**	100

Figure 7.10: Rough estimate of the relative frequency of the categories of the European forest types for some European countries. Forest types used for the countries treated in this work are circled in red. Source: Barbati et al. (2006)