Department of Spatial Sciences

ONTOLOGY DRIVEN GEOGRAPHIC INFORMATION RETRIEVAL

Nicholas G Addy

"This thesis is presented as part of the requirements for the award of the Degree of Master of Philosophy by the Department of Spatial Sciences at Curtin University of Technology"

August 2009

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: …….

Date: ……...

ABSTRACT

The theory of modern information retrieval processes must be improved to meet parallel growth and efficiency in its dependent hardware architectures. The growth in data sources facilitated by hardware improvements must be conversant with parallel growth at the user end of the information retrieval paradigm, encompassing both an increasing demand for data services and a widening user base. Contemporary sources refer to such growth as three dimensional, in reference to the expected and parallel growth in the key areas of hardware processing power, demand from current users of information services and an increase in demand via an extended user base consisting of institutions and organizations who are not characteristically defined by their use of geographic information. This extended user base is expected to grow due to the demand to utilise and incorporate geographic information as part of competitive business processes, to fill the need for advertising and spatial marketing demographics. The vision of the semantic web as such is the challenge of managing integration between diverse and increasing data sources and diverse and increasing end users of information. Whilst data standardisation is one means of achieving this vision at the source end of the information flow, it is not a solution in a free market of ideas. Information in its elemental form should be accessible regardless of the domain of its creation. In an environment where the users and sources are continually growing in scope and depth, the management of data via precise and relevant information retrieval requires techniques which can integrate information seamlessly between machines and users regardless of the domain of application or data storage methods. This research is the study of a theory of geographic information structure which can be applied to all aspects of information systems development, governing at a conceptual level the representation of information to meet the requirements of inter machine operability as well as inter user operability. This research entails a thorough study of the use of ontology from theoretical definition to modern use in information systems development and retrieval, in the geographic domain. This is a study examining how the use of words to describe geographic features are elements which can form a geographic ontology and evaluates WordNet, an English language ontology in the form of a lexical database as a structure for improving geographic information recall on Gazetteers. The results of this research conclude that WordNet can be utilised to as a methodology for improving search results in geographic

information retrieval processes as a source for additional query terms, but only on a narrow user domain.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank the people who have been pivotal in the development of this thesis: my supervisor Dr. Bert Veenendaal for his guidance and wisdom, and Mr David Purnell of Whelans Survey and Mapping, for his flexibility in affording me the time to pursue my academic goals. Lastly, I wish to express my gratitude to my family for their unlimited support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Introduction and Motivation

The Internet is perhaps the most profound communication medium of the twenty first century. Whilst broadcast television has brought world wide and cultural events to the masses, its effect on the scientific community is negligible when compared to the potential to exchange ideas that the Internet's development has enabled.

As network communication infrastructure becomes a priority for more countries and the Internet becomes more pervasive, the spatial nature of data is more apparent and valuable. Given the truth in Moores Law (Moore 1965), which states that processing power doubles every twelve months, spatial data are becoming available at a consequently significant rate; the growing volume of digital information makes the precise extraction of relevant items difficult. As the volume of collected data increases the demand for improved access to online information sources also increases. In this data rich environment information dissemination rather than hardware constraints are the bottleneck in information processing.

The timely and efficient access to relevant data facilitates knowledge acquisition, scientific endeavour, and research advances. Paradoxically, as data availability increases, access to relevant information becomes more difficult (Buttenfield 1997 cited in Fabrikant and Buttenfield 2001, 1). Traditionally users of spatial information have adopted a proprietary or homogeneous approach, i.e. using a single product and using that systems proprietary file format as a de facto standard (Doyle and Daly 2007). This is often "a knee jerk reaction by many organizations trying to meet the requirement to share data and technology by standardizing to a single vendor platform, and a single content model" (Reed 2004, 5).

There are, according to Doyle and Daly (2007), hundreds of software houses developing their own or third party technologies. In addition, end users are also developing their own applications via macro or scripting languages and through the customisation of Geographic Information Systems (GIS). Subsequently there are a

great number of systems and file formats in circulation. Different geo-processing systems produce different types of data, and different vendors produce data in different formats. Proprietary software developed by vendors is not open source and therefore not interoperable between other systems. Different data producers, even when using the same vendor systems do not use the same terminology for spatial features (McKee 2003).

The same can be said about how we as individuals relate to spatial information. An extreme and obvious comparison is that not all people speak the same language. A less obvious one can be found in the technical jargon used by various professionals in different fields who access geospatial data. People not only speak a variety of languages, but within the same language use different words to refer to different concepts and entities.

For the foreseeable future there is no doubt that the availability of information will continue to improve. Subsequently, the users of spatial information will also continue to increase and the problem therefore becomes one of how to best utilize these data. Whilst the same information may be stored a many number of different ways, the information is also perceived differently depending on who has collated the information and to what purposes the information is to be used. As the data gathering process is often the most expensive aspect in any information system or scientific endeavour it makes economic sense that the full utility can be gained from it.

Standards are necessary if there is to be communication of location (and time), route, types of service, etc. across diverse technology platforms, application domains, classes of products, carrier networks and national regions. GIS models of local counties or towns would be expected to use data models tailored to meet individual needs and feature names unique to a community or culture, but “if each were to make its data available as though it is in the reference data model then combining the data of several local or geographically separated communities at a later date would be easier” (Buehler 2003, 3).

In recognizing the economic benefits, to be gained from data interoperability from local government to state and national levels, the Victorian government of Australia initiated an enquiry detailing the specification and requirements for a strategy to meet the spatial information needs of its jurisdiction. The main benefit of this strategy was the recognition of the savings in the collection of data which can be reduced or offset by acknowledging the value of the data across a variety of communities, and in addition, the benefits from the development of standards and a common logical data model.

This strategy included the proposal of six consolidated elements for a state government's GIS which were listed as:

- a set of core digital spatial data (including maps) of appropriate quality that is accessible for use by the rest of the community within that jurisdiction.
- a set of standards for exchange of the spatial data.
- a set of protocols to allow timely and efficient access to the high volume of data within the core spatial databases.
- a set of hardware and software (including application software) to support the management of the data and databases to meet the needs of the providers and users of data.
- a pool of technical, planning and management expertise available to support a cost-effective GIS needed to meet the information needs of the jurisdiction.
- a vision within the jurisdiction about using appropriate GIS to meet its strategic needs and the appreciation by top management of the need to develop and maintain that GIS.

A key component in the strategic vision of Victorian government is the AS 2482 standard for exchange for topographic data, which has been available on magnetic tape format since 1978 (Chan and Williamson 1995). The purpose of this standard is to specify a file structure and formats for the interchange of digital mapping data and associated information. This includes a code structure to be used for identifying individual types of cultural, hydrographic, relief and vegetation features associated with mapped entities. The so called feature codes defined in this standard ensure that topographic features such as hydrants and sewer manholes are recorded with their

geographic coordinates by surveyors in a recognized and accepted representation. This agreed representation can then be used and added to by cartographers, civil engineers or other professionals and professional communities, who are then able to utilize, recognize and understand this information.

The Spatial Data Transfer Standard (SDTS 2008) has also arisen because of similar motivations as a robust way of transferring data with minimal loss of information between dissimilar computer systems. This standard facilitates for interoperability issues of both the technical and semantic forms with its primary goal being in the preservation of meaning when transferring data between systems.

The Open Geospatial Consortium (OGC) is an organization developed to pursue the resolution of such issues and has made some headway into the problem of data standardization, however research suggests that the depiction of database content (or database 'semantics') can be based on a spatial or even a geographic metaphor, also called information spaces (Fabrikant 2001; Holt 2001), which can have an integrating effect both at a machine representation level as well as at a conceptual level to improve search precision and recall. To date, systematic approaches to apply spatial metaphors to information archives lack solid theoretical foundations.

The goal of this research is to investigate the current methodologies to enable the integration of information both at a machine or representation format, as well as between different and disparate groups of people who have alternate views of reality represented by the words they use to describe the objects and things which form that reality. One such methodology is that of ontology as a tool to integrate both data representations and user groups for geographic retrieval and management operations. Research is undertaken into the theoretical foundations of geographic ontology, and WordNet, an existing non spatial ontology is evaluated as a tool to improve the retrieval of geographic information using a subset of the Australian gazetteer dataset. The suitability of ontology, as compared with alternate corpus based methods is evaluated via the performance of this application, to improve data retrieval, data management, and data integration for the geographic domain.

1.2 Research Rationale

While standardization is necessary, it is not sufficient. “Standards are secondary to the requirement for gaining an understanding of business processes and how geospatial data and services by extension, and standards can best be used” (Reed 2004, 1).

Standardisation is useful as a starting point for preventing the proliferation of non standardised data; however we cannot turn back the clock, accounting for legacy datasets whose creation was not predicated by a need for spatial or representational integration from or developed to any accepted standard. These data have innate value even though their initial design does not encapsulate many of the specifications of their possible use. Additionally, since widespread heterogeneity arises naturally from a free market of ideas and products there is no way for standards to banish heterogeneity by decree (Elmagarmid and Pu 1990).

Bishr (1998, 4) defines semantic heterogeneity as “the ambiguous nature of terminology used to describe facts or concepts of the world we are interested in.” Standardising information as it is created can only achieve so much. The way in which information is perceived from the point of view of the goals of the user must be considered in order for the goal of precise and relevant information recall to be realised, and in the process disambiguate terminology.

Frequent users of GIS may in time become familiar with the underlying representation methods of a particular system such that the query terms they develop are more relevant to a retrieval task. Infrequent and first time users of such systems would experience a significant learning curve in assimilating an unfamiliar local system. In the distributed environment of the semantic web, the solution to the problem of interoperability in information retrieval cannot take on a specific localised view or make any assumptions with regards to user experience other than to treat every possible user as a first time user.

A valuable benefit of GIS amongst analysts known as service chaining or service composition relies on the integration of several data sources which in combination provide knowledge about certain phenomenon, or to achieve a specific task. For example, a landslide assessment model may require information from soil composition, rainfall, slope, etc. in combination to quantify the potential of landslide for a particular area. Similarly, the 'dial before you dig' service provides information about the location of underground services such as oil, gas, water, sewer pipelines prior to land development. The information, however, may not always come from the same providers and therefore different terms may be used to describe phenomena of a similar nature.

To allow service chaining it requires that information communities are able to maintain their particular professional or regional semantics (Gould and Hecht 2001), yet the uninitiated user should be able to recognise familiar features regardless of the terminology used to identify and subsequently integrate them. For example, cartographic software packages often include symbol libraries grouped under various headings according to the potential user's professional discipline, for example forestry, transport etc. or the nature of the symbols they contain, civil, water, emergency, tourism. Finding which symbol best represents a particular feature, for example, a bridge, hospital, toilet, etc. can be potentially tedious, given the large number of categories and choices. If symbols have both a graphical representation as well as a name they can be linked conceptually to names and categories of both the user's choosing and their existing terminology.

The unique problem in information retrieval and data management as it regards geographic information is primarily concerned with identity. This research seeks to analyze methods to enable the integration of beliefs about geographic phenomena and the relationships between these phenomena and the people that describe them, which manifest itself in language by the terms people use to describe features. This is undertaken by gaining an understanding of the methods of creating a categorical model to meet the perceptions of a group to describe the geographic domain and using this model to manage the information retrieval of geographic data. The nature of geography and geographic information retrieval and management is complex. This

research will illustrate that a specification of geographic knowledge and terms has value in the role of establishing formal ‘guidelines’ that serve as an integrating tool that has the ability to resolve ambiguity between the physical, logical, representational and implementation levels of a retrieval or data management architecture.

1.3 Objectives

The primary aim of this research centres upon the goal of text based information retrieval within the architecture of the semantic web. This is achieved by understanding the factors which make knowledge more accessible and available, and in doing so to allow patterns and discoveries to be made from spatially correlated data. This is done in the context of the theoretical foundations which are the realm of standards organizations in order to foster scientific growth and interoperability of data between groups that utilise spatial data.

Geographic data are characterised by a spatial location. However, this alone is not a sufficient method by which to subtly discern between the various forms and types of information within a spatial context, for the processes involved in information retrieval of a retrieval platform as large in its scope of ideas as the semantic web. A key problem presented by the scope and size of the semantic web is information dissemination, and how to apply a method in which to discern between data which are similar in nature in an organized system and understandable to someone with particular search results in mind.

The objectives of this research are:

1. To understand the key processes which form the information retrieval paradigm and to identify the restrictions or bottlenecks which encumber the efficient and precise retrieval of information within this model,
2. To research knowledge management and the way entities and their relationships in the real world can be modelled for the purposes of integration and information management,

3. Research ontologies in the geographic domain to define and analyse methods by which geographic entities can be distinguished in terms of their identity and related to similar entities so that an universal geographic representational knowledge structure can be realised,
4. Apply a geographic ontology to an information retrieval scenario to evaluate the effectiveness of ontology driven retrieval, and evaluate its authenticity in terms of results obtained in Objective 3.

The following sections outline the specific hurdles confronting the issues of managing the complexity of data in information retrieval specifically as it relates to data with a spatial context, which will be elaborated on further within this research.

1.4 Research Methods

For information retrieval the factors which hinder the integration of knowledge must be fully understood hence, a review is needed as to the processes which occur between how the user expresses a request for information, and to how that need is met (Objective 1). An ontology is "an explicit specification of a conceptualization" (Gruber 1993, 1). The value of *ontology* is that it connects an entity with its context (relationships with other entities). A literature review of ontology as a knowledge structure is undertaken (Objective 2) A geographic ontology as such is a method to represent the objects, concepts and other entities that are assumed to exist in the area of interest and the relationships that hold among them. The term, 'geo-ontology' is defined by Smart, Abdelmoty and Jones (2004, 2) as "a relational language with the desired abilities to represent real world geographic concepts, the data properties for each concept, the relationships between concepts and specialisation/generalisation concepts of hierarchies."

The growth in the number of users and the variety of user communities means that usability of systems comes first. The development of a formal model is required to capture the basic reasoning behind human conception of geographic phenomenon and therefore bind this with what physically exists and how it is logically perceived, with how it is represented and implemented at a machine level (Objective 3). This

approach represents a departure from traditional software design in which usability determined by the implementation of the data representation.

The creation of a domain ontology is a task that is often carried out during the requirements specification during the design phase of a software application (Gruber 1993). In this scenario, what and how things are represented in the ontology are design decisions, based on the processes, inputs and outputs of the systems, which are in many respects subjective to the designer, as well as to the specific application. The user group of the application may play some part in the design during consultation between the developers involving the representation of tasks and processes as well as providing the terminology, which is familiar to the group. The analysis of human computer interaction as such requires consolidation and reconciliation between various user groups between geographic language and its terms, therefore how people relate to the landscape must be understood in order to establish if there are common or universal concepts which all people share.

The specification of geographic ontology must consider the perceptions of geographic features by the people who use and communicate about these features. These perceptions and the terms used by groups of people are often different between cultures and groups of people in different spatial regions, distinguished for the most part by the significance a feature has to a community. These may be based on the use a feature has to a community such as physical sustenance or transport to having a spiritual significance.

A study of how users and user groups of spatial information generally perceive geographic space incorporated into the retrieval process may result in a more efficient algorithm which can mesh together in a meaningful way the features in geographic space. The research will explore the use of WordNet (Miller 1990) a lexical database of the English language as a geographic ontology, providing terms denoting such geographic concepts and their relationships. The application of WordNet as an ontology for the geographic domain will be demonstrated, using gazetteer datasets to which the retrieval method will be applied. This application will

be evaluated as a ‘proof of concept’ that ontology can improve the management of large datasets via improved recall and precision (Objective 4).

1.5 Tools Used

The application of the theories of ontology driven retrieval and management in the geographic domain are applied to gazetteer datasets to analyse the effects of ontology in the improvement of retrieval precision and relevance for large datasets, that is, to provide proof of concept. The architecture used to perform this study is an open source model of freely available and interoperating components known as WAMP. WAMP is an acronym for Windows Operating System, Apache Web Server, MySQL database and PHP (Hypertext Pre Processor) a common and widely used software bundle for internet applications development. PHP is a powerful ‘C’ like server side web language, which can facilitate database interaction between other client side web technologies such as HTML, XML and JavaScript. Used in conjunction with these technologies, PHP can facilitate web-based functionality akin to the seamless operation of desktop applications, negating the characteristic ‘click and submit’ functionality of the early form based systems. Google Maps will provide the platform from which to display the cartographic representation of the results of a search query, from which PHP provides the link between the database and the map display.

The WordNet (Miller 1990) lexical database inspired by current psycholinguistic theories of human lexical memory is available to the public domain from Princeton University via Android Technologies (Android Technologies 2007) in MySQL format. It is an English language database consisting of 115,424 words and their meanings. It is also a structure of relationships between words called synsets based on psycholinguistic cognitive theories. The current version integrates approximately 100,000 synsets, of which those pertaining to geographic terminology will be used to form a geographic ontology. This structure will be used to provide geographic terms for query expansion in information retrieval scenarios using two gazetter datasets, a small subset of geographic features of the Perth locality from the Australian gazetteer, which is freely available from Geoscience Australia (The Australian

Government 2008), and the geographic features from the British Columbia gazetteer (British Columbia Geographical Names 2008).

1.6 Thesis Outline

The first chapter of the thesis briefly outlines the background and context of the problem, goals of the research, and the methodology that will be used to achieve them as well as introducing the main concepts which will be elaborated on in subsequent chapters. Chapter 2 (Figure 1) elaborates on the background, and motivations discussed in Chapter 1 from the point of view of the users and uses of geographic data, and the various mediums in which geographic data are created and subsequently accessed. These communities, which create data in various formats, encumber knowledge integration and sharing, which has long term consequences for data utilisation.

Chapter 3 outlines the processes involved in information retrieval, from the formulation of a query to the presentation of matching results. Chapter 3 involves a description of the layered architecture comprising the interface, query expansion and results ranking, with a brief description of the various methods which currently dominate the field in each phase of retrieval. This chapter reviews the main hurdles in each of the main phases in information retrieval and the methods that current research and applications have produced to overcome them. The importance of the user's perception of spatial data is emphasized as the primary framework to which all software design should emanate.

Chapter 4 introduces ontology, with its origins in philosophy to its modern use in computer science, whether it refers to a description of beliefs about the world or a specific domain or part of it. Ontology at an abstract level is the basis of the formation of knowledge structures which can express the beliefs of a domain of interest.

Chapter 5 continues in the vain of Chapter 4 by elaborating on the motivations behind the way humans conceptualise their environment. In addition the way geography and geographic language can be cognized and organised into knowledge structures is also reviewed.

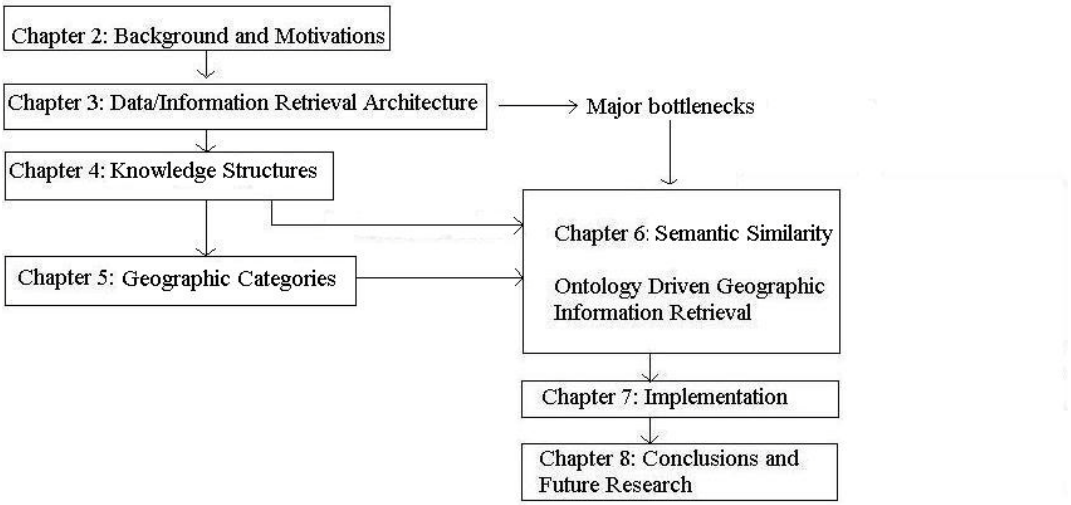


Figure 1.1: Thesis Structure

The information retrieval paradigm (Chapter 3), ontology (Chapter 4), and geographic categorization (Chapter 5) culminate (Figure 1.1) via case studies of the ways in which lexical similarity can be defined within a taxonomic structure. The WordNet (Miller 1990) database is analyzed as a lexical geographic ontology, which can provide the terms and relationships that can be exploited for query expansion in information retrieval. A literature review is presented of the ways in which similarity between concepts within this structure can be determined and what this means for users for geographic query (Chapter 6).

The result (Chapter 7) is an implementation and anaylsis of WordNet as a geographic ontology to improve the retrieval bottlenecks (Chapter 3) on two gazetteer feature datasets intended to replicate an information search and retrieval scenario for the management of geographic data of a larger scale. The results of this implementation and its implications are discussed in Chapter 8 and further research directions recommended.

2 BACKGROUND

2.1 Introduction

Geographic information can be described as information about objects or phenomena that are associated with a location relative to the surface of the earth; as such geospatial data affects almost all aspects of society. Geospatial information has applications in land and natural resources management, environmental and urban planning and development. In the health sector geospatial information is important in planning the delivery of health services and disease prediction and control, as well as monitoring pollution and climate change. Yet, other public services also utilise geographic information such as water, gas, electricity companies, police departments for crime prevention and analysis, public transport planning and provision etc.

Location based services such as those that can be found in cars or within mobile phones which utilize Global Positioning System (GPS) receivers, not only incorporate current road maps, but also, services such as petrol stations, restaurants, accommodation and shopping centres as standard features. The multidisciplinary nature of geographic information use is best illustrated by example. Consider a local government body during the preliminary stages of a new development. This process would require utilisation of geospatial information regarding, electricity and gas, water, environmental impact, such as vegetation and wildlife assessments, and cadastral information. Therefore the value of geospatial data can be seen in its importance across a wide range of organisations and disciplines. Furthermore, spatial data is always changing, cities are becoming larger, and new developments are always being made, therefore geospatial data must be consistently and regularly updated. This illustrates the need for interoperability, so that the data can maintain its value to those interested in using the data even if it is for purposes other than its initial intention.

2.2 Distributed Geographic Information Systems

Public services are expected to interoperate and make better uses of information technology within, across and beyond constituent organizations (Doyle and Daly 2007). The Internet is often cited as a heterogeneous environment due to the diverse range of hardware, software and data, which construct its form (Doyle and Daly 2007). There exists a variety of programming languages within this environment to meet a multitude of tasks. Organizations such as the World Wide Web Consortium (W3C) attempt to govern the implementation of web based languages such as HTML and XML for the purpose of improved communication between applications by maintaining standard specifications and documents. Standards serve the purpose to 'lead the development of World Wide Web (WWW) to its full potential by developing protocols and guidelines that ensure long term growth' (W3C 2008).

The Open Geospatial Consortium (OGC 2008) is a voluntary organisation committed to data interoperability for location based services. In a similar role to the W3C it is responsible for the ratification of spatial data and interfaces. It is a conglomeration between over 235 organizations, business, government and academic institutions committed to the goals of spatial data interoperability via consensus.

The OGC standards have facilitated technological advances that have allowed GIS users to take advantage of new distributed mechanisms such as GML (Geographic Mark-up Language). These standards manifest themselves in software applications development as technical documents that detail interfaces and encodings to support interoperability challenges for software developers working independently (OGC 2008). This allows components or modules of an application to interact seamlessly, also known as 'plug and play' without direct communication between software developers.

There is more data on offer than ever before, and the increase in quantity and accessibility of such data places more emphasis on communication between systems, if the potential of this data is to be realised (Doyle and Daly 2007). The technical standards specifications which form the main products of the standards organizations such as the W3C and the OGC are indicative of a bottom up approach to interoperability issues which naturally affect the design of search protocols and

algorithms. These methods are a result of concerted efforts towards realising the goals of improved communications and interoperability. To recap, therefore, the long term growth of, and value of data is able to be maintained by data once it has been developed to a known standard circumventing the need for expensive recapture or duplication.

The Open GIS specification for the conformation for the 'spatial web' encompasses spatial search engines for the discovery of online geodata sources and geoprocessing services. OGC members have cooperatively developed the GeoMobility Server (GMS) (OGC 2008) which is a set of specifications for open interfaces and schemas that support Location Based Services (LBS) as part of a combined effort between members to improve data sharing and integration of information.

“Historically, that is prior to the emergence of the WWW, digital spatial data have been isolated in the so-called ‘islands of automation’ making interoperability difficult” (McKee, 2003, 1). McKee (2003), states that information sharing between systems is made difficult by two particular problems. Firstly, interoperability issues between spatial information between different systems, and secondly, the difficulties in integrating spatial information in non spatial information systems.

Spatial data can comprise and exist in a great variety of forms. For example cartographic representations typically consist of points, lines and polygons as the basic data types. Point data can include the representation of surface levels, or at larger scales, populated places like cities. Features such as roads and contour lines naturally lend themselves to representation as line objects, whilst polygons may, for example, be used to represent buildings. Spatial data can also take the form of satellite or aerial images, in which the features within these images can be automatically identified by sophisticated software applications.

Spatial data also has a temporal dimension; satellite images are being constantly taken and streamed to ever increasing archives. As storage capacity becomes cheaper and more efficient, the custodians of such data are able to physically maintain valuable repositories of data with a greater temporal dimension.

Spatial data are particularly important in the news media industry; in which text based news reports often need to be associated with a spatial location. These reports require that precise location can be established based on textual or even verbal representation of location. Regardless of the format in which the information is represented, from the perspective of an information framework, there are commonalities and relationships between these data types which are characteristically spatial in nature. The term 'GeoFusion' is used by McKee (2001) to refer to a way of describing explicit relationships that exist between geospatial features that are somehow the same but are described differently and perhaps in different media. McKee (2001) makes the distinction between geographic and geospatial data as not all geospatial data is graphic.

2.3 Geocoding

The process of geocoding is "the conversion of an address, or other features involving a spatial component, such as images or textual place names into geographic coordinates" (Bakshi, Knoblock and Thakkar 2004, 1). Traditionally, this has involved using a street vector data source to obtain the address range and coordinates of the street segment on which the given address is located. The location of the address is then derived by estimating the position of the address on the street by the address range of the street, a method that assumes that all even numbered addresses are on one side of a street and odd numbers on the other, and that given addresses are evenly distributed (Bakshi, Knoblock and Thakkar 2004). Modern approaches to geocoding have employed additional data sources such as the lot size in estimating address position or taking the centroid of a postcode.

Adding a spatial component to data records, data sources and events is a high priority which adds both context and value to data allowing pattern matching and data discovery based on spatial location. However, the process is encumbered by a number of factors which primarily stem from the creators of data (government organizations, businesses or academic institutions) who most often create localized data terms produced subject to the specific needs of small regionalised communities. This makes subsequent aggregation and interoperability difficult. "Standards are one

method that facilitates the ability to leverage IT investment in unforeseen ways" (Reed 2004, 1).

"The development of universal standards for geospatial data transmission would exponentially increase the use of the information worldwide for numerous functions including national security, environmental management and crime mapping" Thomas Kalil, Special assistant to the president for economic policy for the National Economic Council at the (Clinton) White House; cited in Reed, 2004, 2).

Interoperability is described by Reed (2004, 3) as "the active engagement in the ongoing process of ensuring that systems, procedures and culture of an organization are managed in such a way as to maximize opportunities for exchange and reuse of information, whether internally or externally."

Reed (2004) suggests that to ensure the future growth of geospatial data accessibility organizations require:

1. An overarching commitment to interoperability and a focus on geospatial standards,
2. A commitment to collaboration,
3. A commitment to define a geospatial and interoperability framework that meets the business process requirements of the organization,
4. A commitment to the collection and maintenance of geospatial data,
5. A commitment to the retraining and education of staff and management.

McKee (2003) addresses the problem of interoperability further by highlighting two key hurdles which impeded interoperability between spatial information systems. The first being the technical interoperability issues, due to different geo-processing systems producing different types of data, and also including different vendors producing data in different formats. Because of copyright issues the software created by a vendor is not available to competing companies and therefore not interoperable between other systems. Second, there are semantic interoperability issues to consider. Different data producers such as companies and local government organisations even

when using the same vendor systems do not use the same terminology for spatial features.

Spatially referenced data can provide a relationship between technical and semantic interoperability issues, but spatial location alone is not a reliable attribute for a qualifier of the identity of a geographic feature. Landscapes change over time, and what was once a hill or a lake could just as easily be a shopping centre in the space of one or two years.

It is the latter issue which makes the process of adding spatial context to typically non spatial data through the process of geocoding difficult. Towards the resolution of semantic ambiguity, Mckee (2003) advocates that the adoption of standardized meta data creation to facilitate information management and interoperability, with greater potential for knowledge discovery and pattern matching.

2.4 Geospatial Web Services and the Semantic Web

Information records can be generated via a number of means specific to any field or industry and often this information has a spatial context; be it news reports, government census, mining and exploration, forestry or botanical repositories or any discipline or venture whose data involve a spatial location as an attribute. The semantic web operates by using textual word matching to reconcile a query with the data source, as such meta data or name tagging of non textual data are the means to link information in different formats.

The advantage of web based GIS is that it allows delivery of geospatial information to a large number of users in geographical dispersed networks without the costs and constraints of software installation and software updates for each end user. These costs are borne with the cost of the computer hardware. A key driver, which has developed as a result of the advances in data creation and communications infrastructure is the efficiency with which these data can be accessed, in a meaningful and uncomplicated manner so that the application specific user is able to utilize the information.

A fundamental pre-processing step for many geographic information retrieval systems is toponym recognition and resolution, which is the task of “mapping the name of a location to the spatial location that the name refers to” (Leidner 2006, 3). This process of *geocoding* also referred to as ‘spatial grounding’ returns as a result the hardcoded position of a text-based named entity, in a known cartographic reference systems such as latitude or longitude. (Liedner 2006, 3)

2.5 Interoperability

Firstly, “similes and metaphors are utilized in human language and may emulate human thought in a machine way” (Holt 2001, 1). A solid spatial design strategy for constructing usable information spaces will help improve communication between information seekers and database providers for efficient knowledge exchange. Ontologies have been acknowledged to be the core methodology for capturing and sharing semantics of geospatial information (Klien and Probst 2005).

At a fundamental level ‘semantics’ in language relates to the meanings of words and symbols. A clear and agreed upon definition of words used to describe geographic features, objects and scenarios is a key milestone in furthering a methodology which can resolve the standardization and interoperability issues which are essential to information retrieval. Achieving the goal of semantic interoperability to fully utilize the breadth and depth of the vast semantic web requires the development of an intelligent framework to support effective communication and the expression of human language concepts in machine form.

A sound representational framework of space grounded on ontological and semantic principles can be transferred directly to the explicit geographic domain as a basis to reduce current limitations of how geographic space is represented within geographic information systems and serve as a basis to enable textual matching of disparate data sources. A geographic ontology can serve as a solid framework for data generalisation through semantic abstraction which overcomes the limitations imposed by developing a standardisation of terms. It is thus a promising avenue for research in dealing with the problem of information overload, synonymy and polysemy if data archives are to keep growing exponentially.

Ontologies, specifically domain-specific ontologies are at the heart of most approaches to semantic interoperability (Klien and Probst 2005). Domain ontologies help to manage semantics of terms in application schemas and they may enable semantic matchmaking. This is crucial for realising semantic interoperability between different information communities (Klien and Probst 2005).

Klien and Probst (2005) use the example of the feature ‘water level measurement’ to illustrate. Study of a specific domain would reveal the important attributes of this feature which are important to the application, such as location, height, date and time. This in turn would reveal how this data could possibly be perceived by other domain applications and how these attributes would be useful in clarifying the relationships between levels of detail and attributes between domains. This would enable added value to be gained by exploiting the relationships in data between domains for pattern finding, for example, adding composition of the water type to determine how water levels are differently affected by composition of the water between fresh water and salt water bodies.

A thorough study of geospatial domain ontologies must be made prior to the definition for a formal description of geospatial web services, serving as a common ground to which the members of different communities can commit. Standards are a good starting point in this regard, so that applications are developed with a bigger picture in mind. Application ontologies further describing the available information sources can then be described using the concepts provided by an agreed upon geospatial upper level ontology encompassing a range of possible terms which can integrate the various domain ontologies. As a consequence, a user's query can then be syntactically, machine comparable to all existing domain application ontology concepts designed from the ground up via a unifying upper level ontology of less specific and better known terms. These terms have more relevance to first time and infrequent users who are not totally familiar with the intricacies of each domain.

By subsumption reasoning, the sort of terminological reasoning performed by a GIS user can automatically infer if application concepts are equivalent to, or are thus sub-

concepts to the query concept. For example an area labelled as a ‘yard’ can be defined as equivalent to an area of similar size labelled ‘cartilage’. These in turn can be classified as sub concepts to the larger spatial area class of ‘field’ which encapsulates both yard and cartilage.

As shown in Klien and Probst (2005) the integration of matchmaking capability into spatial data infrastructures overcomes some of the semantic heterogeneity problems in service discovery leading to increased recall and precision. This capability is a result of a clear definition of geographic feature such as lakes and mountains and what attribute data one is likely to expect from each.

A ‘layered ontology architecture’ separates implementation and physical universes (Klien and Probst 2005). In geospatial applications a town is often modelled as a point feature; however there is no ‘ontological relation’ between the real world and the representational structure of a point. If towns are modelled in an application by representing them as points, then this relation between town and its geometrical representation will be part of the application ontology. Klien and Probst (2005) suggest that through iterative development and analysis of application ontologies, a degree of agreement can be arrived, that is comparable to a standard. Higher level domain ontologies such as a geographic ontology can then serve as a source for building application ontologies and therefore must be highly stable in their conformance to industry accepted norms. The so-called *implementation universe* is largely governed by the specification of the OGC and is based on agreement by its constituent members. Frequent changes to these domain ontologies, say Klien and Probst (2005, 4) would “discourage service providers from referencing their application ontologies to them.”

At the implementation level the problem is that between domains similar features are often defined in different ways making matchmaking difficult. Therefore an essential definition of a concept can compensate for this problem by allowing degrees of membership of a feature to a concept. This is also known as ontological scale or granularity. For example, the clear definition of the concept of the height attribute in geographic features allows the terms hill and mountain to exist on the same

conceptual path, via definition, but these two features are separated by scale of measurement. From their study in this area, Klien and Probst (2005) comment that the aforementioned utilization of taxonomic reasoning is useful but not completely efficient in this regard. From the point of view that in a non taxonomic structure a concept does not have to occupy a fixed position in a static hierarchy, Klien and Probst (2005) suggest that its position in the hierarchy can be dynamically inferred based on an existing concept and role definitions using subsumption reasoning. By this reasoning the advantage is to enable searching for unknown information sources (Klien and Probst 2005). This theory can be viewed as more realistic in conceding that there is no clear and set boundary separating two features. These are often subjective and relative to the application and context to which they are represented.

For example in a hydrographical model terms such as water level, water body and discharge are formalised on the domain level. It is stated that every water body has a water level and a discharge and that these qualities can be observed and measured, thus the general description of 'water body' provides an entry point for the semantic search. How measuring and representing is done for a specific water level measurement can then be formalised at the application level.

The separation of concepts for data presentation from geospatial concepts is a crucial requirement for consistent, implementation-independent ontologies.

2.6 Ontology Driven Geographic Information Systems

"Certain communities make a commitment to ontologies, although the structure used to represent the ontologies is flexible and can be opposed to those seeking standardization" (Fonseca et al. 2002a, 5). According to Fonseca et al. (2002a) despite initiatives such as Spatial Data Transfer Standard (SDTS 2008), the Spatial Archive and Interchange Format (SAIF) and Open GIS, the use of standards as the only means to achieve integration is not widely accepted.

The use of ontologies in GIS development provides "a dynamic and flexible approach to flexible information exchange that allows partial integration of information when completeness is impossible" (Fonseca et al. 2002a, 5). The use of

explicit ontologies will contribute to improved information systems. Since every information system is based on an implicit ontology, when we make the ontology explicit we avoid conflicts between the common-sense ontology of the user and the mathematical concepts in the software, and conflicts between the ontological concepts and the implementation (Fonseca et al. 2002b). Thus, ontology driven geographic information systems can act as a system integrator independently of the model. The five-universes-paradigm for modelling a computer representation (Fonseca et al. 2002a) is used to understand the role of ontologies in geographic data modelling. This paradigm includes:

1. The *physical universe*, which comprises the objects and phenomena of the real world that will be modelled in the computer,
2. The *cognitive universe*, in which geographic phenomenon in the physical world is captured by the cognitive system of a person and is classified and stored in the human mind,
3. The *logical universe*, includes the formalization of the conceptualizations of world in the human mind and gives explicit formal structures, the ontologies that are part of the logical universe,
4. The *representation universe*, where a finite symbolic description of the elements in the mathematical universe is made, reference systems and conceptualizations such as fields and objects are part of this universe,
5. The *implementation universe* used to map the elements from the representation into structures implemented in computer language.

This is often summarised to a four-universe model in which the domain ontology consists of the logical universe and the representation universe, in the form of an ontological vocabulary. Bishr (1997) suggest that the use of semantic translators in dynamic approaches is a more powerful solution for interoperability than can be achieved by using the current approaches which promote standardization.

Fonseca et al. (2002a) describes a process of ‘binding’ that is the result of the process of linking the results of scene interpretation to geographic features. This system allows expert and non-expert users to choose how much involvement they want to take in the image classification process. The main objective is to allow a single

feature in the GIS to have more than one description. Matching features found in the images which are then used as classes in the ontologies (Figure 2.1).

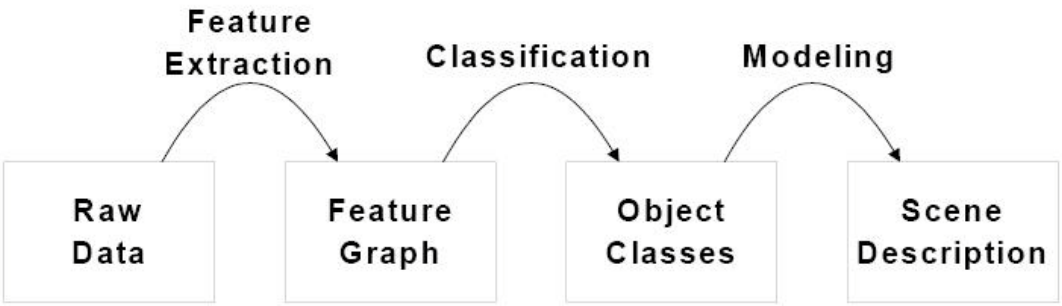


Figure 2.1: Feature Extraction of Remotely sensed images (Fonseca et al. 2002a).

A similar approach can be applied to non-image based classification of geographic features. Consider the representation of hills and mountains, in a Digital Elevation Model (DEM). Cartographically, the boundaries of geographic features such as hills and mountains are only annotated via a name and a height above sea level. So how would someone searching for hills or mountains interrogate a cartographic database? A landform is defined as a region in the continuum of variation of the surface of the earth. Physiography (also known as land surface characteristics) is the study of the physical features and attributes of the earth's land surface. The detection of the physiographic features of a terrain is the first phase involved in the classification of the various landforms of the terrain. According to the findings of Dinesh (2007) the main reason why physiographic features are more suitable to be considered as fuzzy objects is that the assignment of any location to a physiographic feature is not necessary stable under repeated observation at different scales. Landforms such as mountains can be identified using the height above a certain elevation as a membership function, with membership increasing with height. Geomorphological landforms are generally viewed as boolean objects. However, research by (Dinesh 2007) has shown that landforms are more suitable to be viewed as fuzzy objects. Dinesh (2007 16) gives the fuzzy membership of a physiographic feature μP at location as the average of boolean memberships of that feature over the scales of measurement. Scale and subsequently context therefore is an essential attribute when describing geographic phenomena. “What a person calls a mountain may be called a

hill by a person from another area where mountains are more spectacular" (Mark, Turk and Stea 2007, 3).

3 RETRIEVAL ARCHITECTURE

3.1 Introduction

This chapter outlines the main aspects involved in the information retrieval mechanism, from the specification of a query, or what is required by the information seeker to the retrieval of the actual data as it is represented in its raw format. An understanding of what is involved from getting from a user's query to a search result will highlight the areas in which information flow 'bottlenecks', mismatches or imprecision occur and can then be identified, with the view that they may then be subsequently improved.

The terms data retrieval and information retrieval are often used synonymously, however there are clear distinctions between them. The differences between the two dichotomies are indicative of the complexities that exist both in data storage structures but also the range of available architectures.

Stubinz and Whighli (2007) defined information retrieval as the science of locating, from a large document collection, those documents that fulfil a user specified information need and is primarily concerned with the retrieval of (usually text) documents and matching the contents of documents with terms in a user query. Data retrieval algorithms are deterministic, and demand an exact match between the query specification and the contents of the database (Larson 1996). The variations between the two retrieval architectures are vague and overlapping and in this respect certain methodologies in information retrieval are researched such that they may be applied to data retrieval.

Due to both its size in terms of data, and popularity the HTML document retrieval architecture and the methods used for searching these sources in engines such as Google and Yahoo have received the most attention in terms of research in information retrieval using the platform of the WWW. This model can be applied to database applications, although some restructuring and rethinking of certain concepts is necessary.

3.2 Key Processes in Retrieval

The architecture of information retrieval regardless of the data type or format of the information consists of basic tasks common to retrieval of all information. In a development of an information retrieval system de Vries and Wilschut (2004) outlines these processes resulting in the so-called retrieval model.

1. Choose an appropriate scheme to represent the documents,
2. Model the query formulation,
3. Select a ranking function.

The dependencies between each step are evident, the disparity between the query and source data are resolved by the formulation of the query which relies on the format of the data source (in this case text documents) to meet a basic criteria, to which further comparison can then be made. The result is an 'ideal' document to which the query is best represented. Picard (2000) outlines these processes in a document based information retrieval system (Figure 3.1). Figure 3.1 illustrates the agreement which is sought between what exists (data source) and how it is represented, to that which is requested by the user.

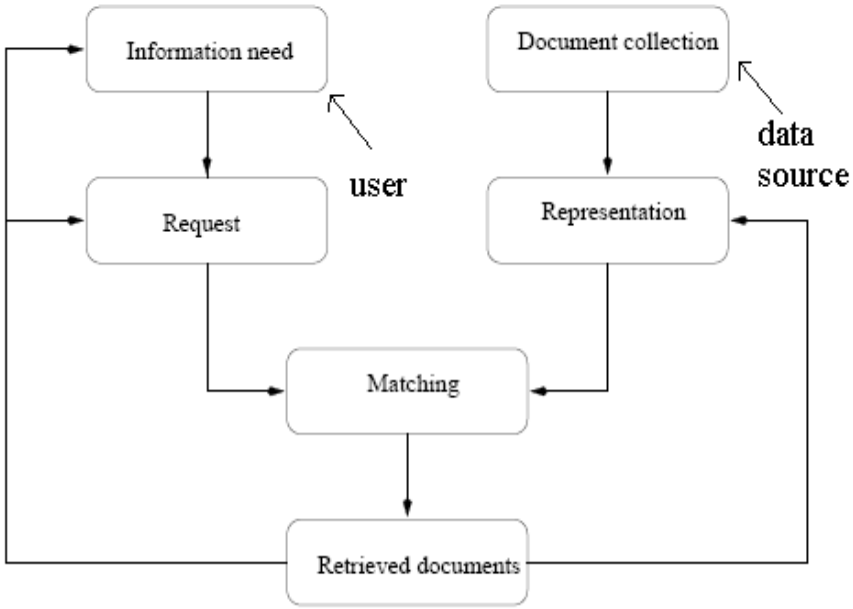


Figure 3.1: Information retrieval model (Picard 2000)

The process of matching varies between architectures and is dependent on the nature of the source data, which determines whether results of a search can be broken down and ordered by the degree to which they agree with the user's query or simply treated as matching or not matching. This is an important consideration in circumstances with potentially hundreds to thousands of matching results. The procedure for employing results ranking as it is known is discussed later in this chapter.

There are two ways of approaching the design of an information retrieval system, to modify the way that the data are represented to search the manner of the query, or to modify the way the query is represented in order to suit the data source. Welie and Veer (1999) observed that user interface design has never really been integrated into software engineering methods, even though a 1992 study by Myers (1995) found that an average of 48% of the code of applications is devoted to implementing the user interface and about 50% of the implementation time is devoted to implementing the user interface portion. Interface design is inherently tied to the usability of the system and as such governs the users perception of the underlying data.

According to Harter (1986, 2) an information retrieval system consists of a "device interposed between a potential user of information and the information collection itself," and contains four major components:

1. The database,
2. The *communication channel or interface*, which has both a physical component that facilitates interaction,
3. A *conceptual component* that gives the user guidelines on how to interact with the information structure and search mechanisms,
4. *The user*.

3.2.1 User Interface and Query Representation

The motivations for interface design should adhere to defined principles that allow visualisation and provide support for the information seeking processes itself in which the key design issues are functionality and usability. As such an interface must provide support for *query specification* and the *presentation* of results as well as conceptual support for the process as a whole.

In the traditional online search environment such as Google or Yahoo the searcher is obliged to break down a search request into distinct concepts in the form of keywords. For example, try typing more words into a search request in Google and it is likely less rather than more results will be returned. The informed searcher also thinks how the concepts and terms associated with them correspond to the document representation stored in the database (Efthimiadis 1996).

Without a specific understanding of the information that is being sought and how it is likely to be represented the inexperienced user can often become lost and frustrated. Additional support may often be provided via *relevance feedback*, or a *starting point* such as list overviews in which the user chooses items from a listed collection of names to search, examples and automated source selection which involves unsupervised clustering often combined with a graphical display to support browsing. Several of the information retrieval systems reviewed in the course of this research provided alternative interfaces for both novice and experienced users. The latter via provision of minimal keyword or key variable searching and the former, the so-called 'advanced search' feature allowing more specific searching via the inclusion of several field choices selected from database categories.

Interface design according to Hansen (1998) is a study that encapsulates the manner in which a query is specified, or how the underlying data structure is represented to the user. Hansen (1998) suggested that *interface design* on the WWW must contend with an increasingly large end user population and that the growth in research in information retrieval is increasingly moving towards an interdisciplinary approach to interface design in information retrieval and the Human Computer Interaction (HCI). Ergo that the formulation of the interface should be modelled from the perspective of the particular user in a certain discipline and their uniquely defined characteristic requirements, level of knowledge, experience and expectations.

Expectations of a user interface are often based on previous experiences acquired using different information retrieval systems and reflect the users' mental model of the information retrieval system (Hansen 1998). Studies conducted in this area by

Hansen (1998) add support to the theory that there is a period of time in which a user formulates their own understanding of an information retrieval system, learning which keywords and language formats are likely to yield the most relevant results, such that they are able to achieve the best utility for their individual purposes based on past experience and their conceptual perspective of the system.

Agent technologies are a growing field of research designed to complement an individual users' particular search criteria and language, based on utilizing knowledge derived from past workloads. Agent technologies are defined by Finnin, Nicholas and Mayfield (1998) as autonomous and adaptive, goal directed processes that cooperate with humans and computers to achieve tasks, by applying sophisticated domain knowledge and recognizing underlying goals and intentions. Agent theories are based in cognitive science, database and knowledge base technology and distributed computing. In information retrieval agent based technologies such as FAB (Balabanovic and Shoham 1997) use web-based documents which apply adaptive information retrieval techniques formed to a users 'profile' derived from his/her previous search results. The user's profile may also be adapted based on feedback as to how much they like certain results via relevance feedback.

The process of *relevance feedback* is one in which users identify relevant documents in an initial list of retrieved documents, and the system then creates a new query based on those sampled relevant documents. Algorithms for automatic relevance feedback have been studied in information retrieval for more than 30 years (Salton 1983), and the research community considers them to be thoroughly tested and effective (Croft 1995). Relevance feedback applies the logic that terms related to those selected by the use, and which successfully discriminate between relevant and irrelevant documents are useful in discriminating between documents themselves. This is a time consuming process, as it requires considerable input from the user in determining which documents are relevant and those that are not and is impractical although effective in most circumstances.

“The input of multiple disciplines is needed since computer science does not have all the expertise needed for the design of truly ‘usable’ systems” (Welie and Veer 1999, 1). Furthermore Welie and Veer (1999) suggest that integration can be achieved using ontologies based on this expertise to address the integration.

There exists a variety of alternatives for query specification, ranging from Boolean queries to free-text queries and non-textual queries, which complement the software architecture for each different interface style, such as command lines, forms and menus. Natural language offers the most intuitive method of query specification as opposed to a word by word restricted interface and is generally treated as a so-called ‘bag of words’ with its roots in artificial intelligence and linguistics (Smeaton 1995; Chandalia and Srihari 2006). This method stems from the need to automatically process text based data to cope with the modern phenomenon of information overload (Strzalkowski 1999).

The knowledge domain describes context and decisions with regards to information design as such developers should always consult the knowledge domain, as “the knowledge domain is the basis for the usability design” (Welie and Veer 1999, 6). Inevitably the digital representation of reality is an abstraction and simplification of the real world filtered in its complexity to the goals of the system. Often, as it is with maps there is an inferred or stated (by means of a legend) explanation of the symbols used and their meanings.

Graphical interfaces are increasing in popularity from a user perspective as hardware constraints are becoming less of an issue when considering the processing and memory power required to render images, but also because they overcome the boundaries of language. Graphical interfaces are attractive because they can overcome language barriers between cultures, but also, within the constraints of ones own language where the precision in choice of words is often less accurate than that which can be expressed by pictures or by imagery. Such interfaces focus on recognition instead of recall and are well suited to a domain of non-expert and infrequent users. In addition, graphical interfaces have other benefits in being able to

provide direct manipulation as well as continuous representation of objects where textual representation can be limited to discrete representation.

Interface design to spatial data concerns itself with the internal representation of the data and how to convey the representational structure with the users own beliefs about space, place and features of a landscape (Figure 3.2).

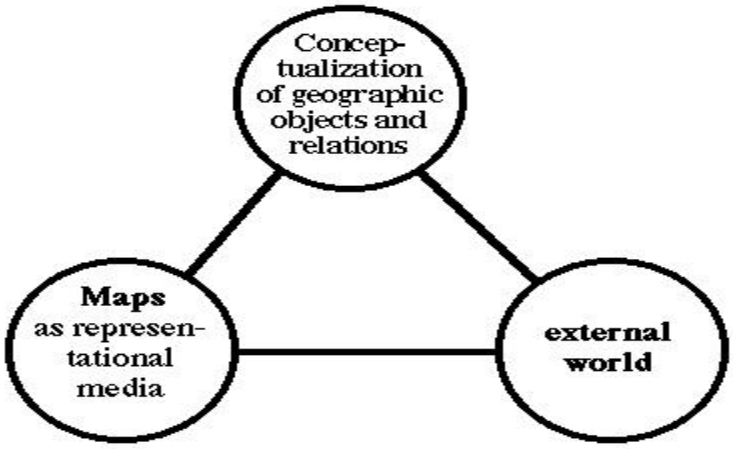


Figure 3.2: Map Interface Paradigm (Goodchild and Kemp 1990)

Goodchild and Kemp (1990) describe interface design to spatial data that often involves the consideration of frames of reference:

- Egocentric frames of reference are those in which objects are always represented in their relationship to the individual.
- An environmental frame of reference uses a local point as reference and moves when the individual moves from one local area to another.
- A global frame of reference is constant irrespective of the location of the individual.

Interface design to data of a spatial nature is unique in that it supports query visualisation and specification methods of a characteristically graphical nature. The sophistication of modern computing hardware means that most basic systems can facilitate the display of high resolution satellite ortho imagery and maps capable of displaying an array of cartographic symbols and features such as roads, terrain and

service information. However, computers deal with limited models (discrete representations), which are often not in the preferred from of interaction, in fact the models may be in an uninterpreted form for example raw images from remote sensing. Modelling means chunking, simplifying an abstracting from the real world as we experience it.

Goodchild and Kemp (1990) report that common deficiencies in spatial interface design are the:

- the lack of a dynamic view,
- the weakness in representing fuzziness, and
- an unbalanced load across available sensory channels.

Spatial data is unique in that there is a strong need to show processes which are encumbered by the predominantly static images available to most GIS. As such the difficulty of spatial interface design is how a process can be mapped to the interface. Fabrikant (2001, 1) states that "the use of the spatial metaphor has become popular to depict complex database content in information visualisation communities." For example, the symbol for first aid is commonly associated with a cross, water and related services are usually coloured blue, however these may vary between groups. The value of the metaphor is to formalise the issues of information space design is threefold:

1. Geometric reduction of database complexity (e.g. cartographic generalisation) is formalized using basic geographic concepts, such as identity, location, €magnitude and time.
2. The concept of Benediktine space is introduced to propose a structured approach for the semantic generalisation. Benediktine spaces are semantic constructs that preserve properties of mapped entities and functional relationships between entities represented in an information space.
3. An information user centred view of design is adopted.

A review of the literature with respect to interface design for spatial data finds common agreement between researchers that geographic space is more than absolute

as represented by Euclidian geometry or topology but include experiential properties and socially constructed meanings.

Sources of error in spatial reasoning are a result of a lack of consensus between groups whether it is between the designers and users, or between disparate user groups as to how peoples explore and speak and think of space and place. In addition, there exists a difficulty with representing fuzziness, the distortion caused by non-discrete atomic objects in the real world.

How people reason about space is dependent on many factors these are important research considerations towards gaining an understanding in interface design. Mark, Turk and Stea (2007), refer to such factors as ethnophysiographic dimensions which include grammar, topography, climate, vegetation, lifestyle, religious beliefs, historical factors and place names.

The problem of spatial reasoning in user interface design is most evident when employing the use of natural language processing in the use query. This complexity of using language to describe spatial query is best illustrated in the use of prepositions to convey spatial relations, which is subject to complex and hidden rules. Some examples are the use of terms such as: in, on, between, across, and near. We say the car is near the house but not that the house is near the car, or across the lake rather that along the lake.

Current interface functionality for natural language query processing usually involves such processes as language identification, spell checking, phrase detection, numerical conversion, tagging, query transformation and results generation (Dittenbach, Merkl and Berger 2003). Sometimes, the words are reduced to their stems (stemming), in order to reduce a word to its root variant. For example ‘walking’ ‘walks’, ‘walker’ are reduced to ‘walk’, and ‘computing’ and ‘computer’ are reduced to ‘compute’. These methods attempt to capture or isolate a specific concept that can be express in a number of terms. Frequent words with little semantics are left out in a process known as stopping. A ‘stop-word’ library containing words which have little search value such as ‘the’ and ‘and’ can be used

to eliminate these words and therefore minimize search time when scanning a document.

Techniques from natural language processing may be applied to identify phrases. Within this paradigm further techniques are often necessary to reduce a phrase to its key usable terms or to add meaning to ambiguous terms or phrases. The resulting features of these procedures are called the indexing terms. Berger, Dittenbach and Merkl (2003, 8) observed, "that when formulating a query in natural language, users are more specific than compared to keyword based searches."

Natural language processing is popular in question answering systems such as FAQ's (Frequently Asked Questions) and question template systems where users enter a question relating to a problem in order to find a solution based on a database of answers from similar problems. Natural language processing involves complex algorithms derived in order to provide solutions for advanced users attempting to interrogate the large unorganised textual datasets that dominate the WWW. Compound terms such as "New York", word couples or grammatical constructs such as "due to" and noun phrases have been investigated as methods of extracting context and meaning from words. The *lexical affinity* is the tendency of groups of words or phrases to occur together frequently (Terra 2004). Syntactic patterns such as *noun noun* or *noun prep noun* are also used, because words alone are often not synonymous with a particular concept (Nie 2001). To illustrate, Nie (2001) uses an example of the occurrence of 'data base system' following the *noun noun*, pattern in a document. Both the correct term which is 'data base', is derived but also 'data system' and 'base system' which by the *noun noun* theory also alludes to the concept of the query being requested.

The two primary problems associated with unstructured textual, corpus analysis or natural language processing in the derivation of context is that of polysemy (Frietag et al. 2005) and synonymy (Deerwester et al. 1990). Polysemy is also known as the problem of 'word sense disambiguation' due to the fact that several words have many meanings. For example, the word 'bank' (which has 18 senses) can be used in the sense of 'a place to keep money' or in the topographic sense such as 'the bank of

river.' Synonymy is inversely the problem that there are many words that may be used to describe the same concept. For example, the geographic feature knoll could also be called a small hill, a mound, a butte or hillock.

Natural language prepositions are inevitably dependent on scale and sensitive to context. For example the preposition 'north of' does not indicate exactly how far north or how north. Research involving the incorporation of natural language into user interface design reveals the difficulties in design which attempts to accurately reflect the way humans view the world, think and reason about space. As will be illustrated later in this paper, space and spatial features have different meanings to different groups of people depending on cultural regional and utilitarian factors which inhibit the goal of the a universal interface, if one is even possible.

3.2.2 Named Entity Recognition

A form of the 'word sense disambiguation' problem is the extraction of persons, places and domain specific attributes of entities. Place names are not unique. Research by Vestavik (2004) indicated that there are 26 places in Norway sharing the name *Lade*. Consider for example a query of the place name London in the World Gazetteer (World Gazetteer 2007), which returns 7 results. 1. London: England, 2. London: Ohio, USA. 3. London: Kiribati, Kiritimati. 4. London: Arkansas, USA. 5. London: California, USA. 6. London: Kentucky, USA 7. London: Ontario, Canada. Furthermore place names change over time. For example Istanbul formerly Constantinople or Oslo once called Christiana. Similarly, there are different ways of referring to the same place, due to language variations, legacy historical conventions or colloquialisms. For example '*The United Kingdom'* is also known as '*Great Britain,' Rome* is referred to by its inhabitants as *Roma* and more abstractly Paris is also referred to as the *city of light*. These are collectively referred to as the problem of *synonymy* or *metonymy* (Buscaldi, Rosso and Arnal 2005). The concept of synonymy makes information retrieval problematic in that there is no one correct or universal term to describe a specific phenomena in this regard query expansion is a method used to elicit additional terms which may denote a concept for information retrieval to take place.

3.2.3 Query Expansion

"Most casual users of information retrieval systems type short queries. Research has shown that adding new words to those queries via *blind feedback*, without using input from the user improves the performance of such queries" (Efthimiadis 1996). In Figure 3.3 Efthimiadis (1996) illustrates some of the options available in this part of the information retrieval architecture, which are available to designers and programmers. These take the form of being fully manual, fully automatic or interactive (a semi automatic process). Through analysis of current methods these three processes can fall under the umbrella of being based either on the results of the initial query request or on a 'knowledge structure' which can be either collection dependent or independent.

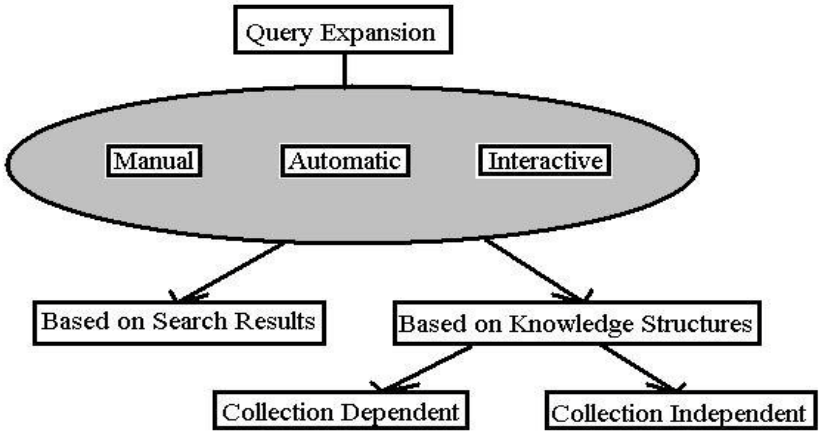


Figure 3.3: Strategies for query expansion (Efthimiadis 1996)

When conducting a search for information, the search user must formulate an understanding of what he or she intends to find. For example the type or format of the document, text-based, a journal, a news article, a book etc. These are the subjective personal requirements that must be articulated in the expected language or format of the source document. The questions must be answered as to 'what are good terms', 'what are the best terms', if these prove too difficult then 'where can the *query expansion* terms be found?'

Robertson and Jones (1994, 4) described the concept of term ‘weighting’ or ‘selectivity’ in information retrieval as a term’s ability to “pick any of the few relevant documents from the many non-relevant ones.” Current research defines query expansion as the process of supplementing the initial query with additional terms as a method for improving retrieval performance. In information retrieval research and texts, the process of automatic query expansion is often referred to as the ‘magic’ of the system. “It may be necessary to consult a thesaurus, a list of subject headings, a dictionary, or a classification system and its index to aid in the selection of terms. This effort usually requires specialised training or experience on the part of the searchers because the results for the inexperienced and those who do not consult the search aids will most likely be poor” (Efthimiadis 1996). Inexperienced users cannot be neglected because each information seeker is in some form a novice, otherwise they would not be seeking information.

The online search can be conceptually reduced to a two-stage process: (1) initial query and formulation and (2) query reformulation, which may be manual automated or interactive, based on search results or based on a knowledge structure (Efthimiadis 1996). For example, a user searching for information on the Claremont Football team might enter a query such as *Claremont.* After expansion, the query might be *Claremont football WAFL tigers*. It is then necessary to decide how useful these expanded terms actually are and how can they be represented and ranked. This is necessary to impose structure upon what may become an unwieldy process when there are many possible results returned. The results of a search are quantifiably ranked by the degree to which the user defined search word matches the given search information keywords. This process is known as *indexing*.

Existing models include methods such as pseudo relevance feedback (Abberley et al. 1999), web-based expansion, interactive elicitations from the user searchers, and expansion approaches based on query clarity. Nambiar and Kambhampati (2004) suggested a method of ‘query relaxation’ to generate more results. This is done by identifying the weakest attributes in tuples and removing them from an SQL query. The weakest attributes are defined as those which have ‘the minimal effect on binding with other attributes’ as identified by *Approximate Functional Dependency*.

Pseudo relevance feedback is an automated procedure that assumes the top-ranking documents following a search are relevant. It extracts the most representative terms (the most frequently occurring) from these documents and assumes they are relevant. Research building on current methods for improving search precision must consider how searchers select query terms. What criteria do searchers use to select the terms and what kinds of relationships are there between the original query terms and the terms selected by the users?

Language modelling, for example using synonymy, has been explored as a method of query expansion in information retrieval (Bai et al. 2005; Buscaldi, Rosso and Arnal 2005; Cohen 2000) in which the document to query relationship is not viewed as a one to one pair-wise relationship but in context and between a set of terms and another term. That is, rather than simply attempting to match one word with another, instead methods and additional sources to attached meaning to a word are used so that if an initial match is not found similar results can be returned. These models are similar to *pseudo relevance feedback* methods. However, they operate on the query side of the formulae, where pseudo relevance feedback performs query expansion using terms from the returned results within a result document, Language modelling uses terms from the query itself in addition to expansion terms via a data source of term relationships.

The two key elements, which need to be considered when applying any form of query expansion, are the source, which will provide the terms for the expansion, and the method that will be used to select terms to be used in the expansion (ranking algorithm). Simply put, automated query expansion of format supplements an initial query with additional related terms, such that if the original term yields a null result the additional terms are used to expand the likelihood of obtaining relevant matching results.

Examples of the so-called *collection-independent knowledge structures* suggested by Efthimiadis (1996) include:

- Domain-specific (manually constructed) thesauri or a searching thesaurus,

- Global (general purpose) thesauri, such as Roget's or WordNet,
- Dictionaries and lexicons such as Collins' dictionary.

Stenmark (2003) explored with success how a manually constructed semantic network, defined as a graphical depiction of topics and concepts of a specific area, in a corporate intranet environment can be applied as a basis for term expansion (Figure 3.4).

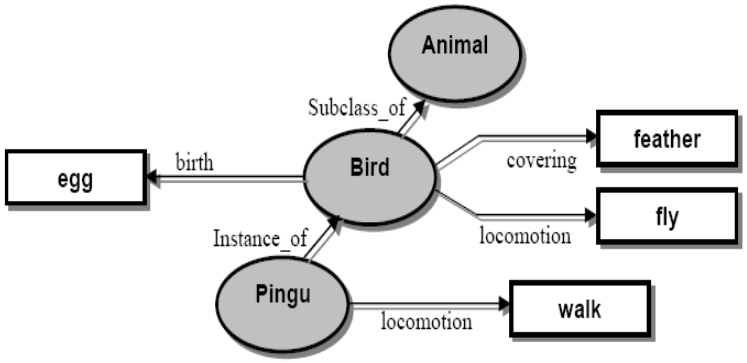


Figure 3.4: A typical semantic network with classes and attributes (Stenmark 2003).

The creation of the Stenmarks' (2003) semantic network is based on heuristic knowledge from ten domain experts from various business backgrounds. In Figure 3.4 (above) the class of bird is defined by its characteristics covering: feathered and locomotion: fly, and birth: egg. It is defined as a subclass of animal. Therefore a bird is an animal. The Pingu is an instance of the class bird, similarly to the way that bird is a member of the class animal, but the instance Pingu has no further sub classes.

Stenmark's (2003) implementation is a simplified model of the class hierarchy based on the *specialise* and *generalise* concepts of inheritance to use only synonymous terms for query expansion. (Figure 3.5)

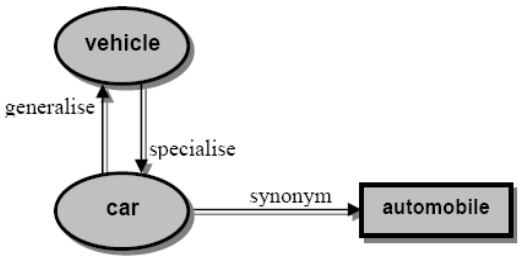


Figure 3.5: Simplified model of a semantic network (Stenmark 2003)

Stenmark (2003) reported that by adding an additional query via the 'AND' conditional statement resulted in a reduction of 12% in the number of documents recalled and an improvement in the increase in precision from 3% to 22%.

Research by Stenmark (2003) confirmed that adding additional terms to a query can improve search results. However, alternate sources suggested that adding additional terms to standard keyword searches often has the inverse effect by returning fewer results (Markowetz et al. 2005; Qui and Frei 1993). Primarily, because the frequency of term occurrences in documents discriminate poorly between relevant and non-relevant documents (Qui and Frei 1993).

In the actual retrieval of items from the database, the algorithms used for matching between the query and the index elements (or database contents) are based on the particular retrieval model. Information retrieval (documents) models lead to a class of retrieval algorithms that are probabilistic in nature, and may involve the actual calculation of probabilities and the use of statistical inference methods.

3.2.4 Results Ranking

The process of results ranking, also known as indexing, is that of assigning a hierarchical order to the search results of a query. This process is subject to the shape and form of the dataset that can affect the discretisation of a concept, the effect of these on the steps in a retrieval model are highlighted by Larson (1996) (Figure 3.6).

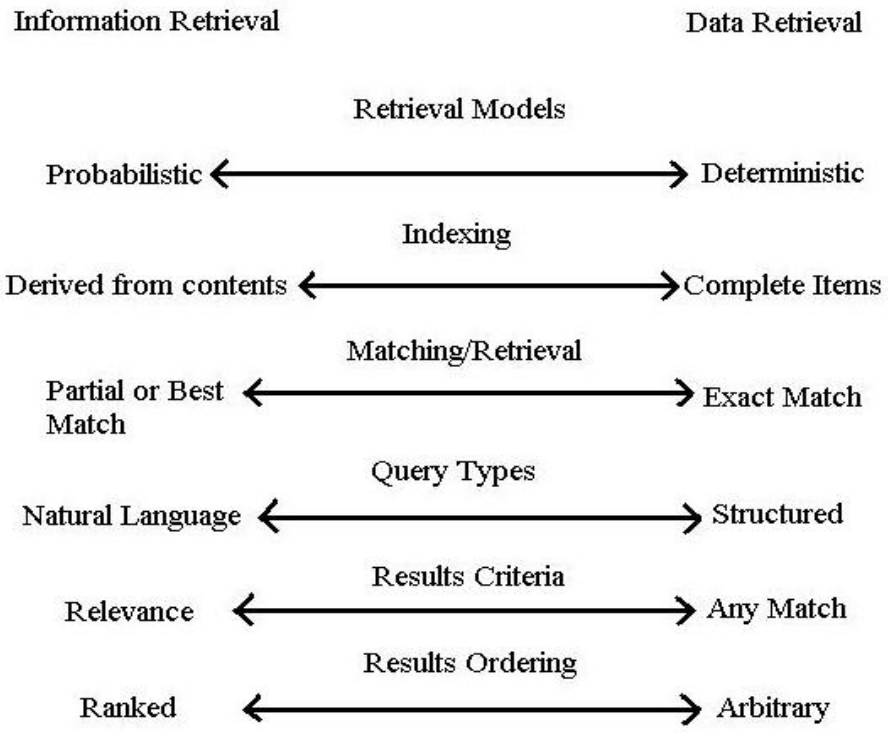


Figure 3.6: Differences between data and information retrieval (Larson 1996).

In the information retrieval architecture ranking functions are often based on the frequency of occurrence of query values in documents (term frequency, or TF). However in a database context, where data are characteristically categorical, term frequency is irrelevant as tuples (database records and joins) either contain, or do not contain, a query value (Chaudhuri et al. 2004).

Information retrieval in database applications is characterised and limited to *Boolean* or two valued logic (Fuhr 1990) resulting in either a 'hit' or 'miss' scenario. In this scenario a 'hit' will result in too many answers and a 'miss' in no results at all. Information retrieval allows the use of partial and best match algorithms based on probability theory, in which the results can be described as a sensitive error response in data retrieval compared with an insensitive one in information retrieval.

3.2.5 Probabilistic Ranking in Information Retrieval

Document based information retrieval is able to employ more sophisticated ranking functions than the retrieval of information from databases. In the information

retrieval of documents, there exist a number of paradigms in contemporary literature for results ranking, examples include those based on probability (Nahm and Mooney 2001) or concept association graphs (Chandalia and Sriharu 2006). These algorithms are characterised by inductive inference and are typically an approximation of similarity between the search query and the data collection. De Vries and Wilschut (2004) defined this form of probabilistic ranking algorithm as a statistic "about the distribution of the indexing terms over the document collection" which can be described using probability theory (Equation 3.1).

$$Bayes' Rule: p(a \mid b) = \frac{p(b \mid a)\, p(a)}{p(b)} \tag{3.1}$$

For example, for a document collection *D*, we require a probability that a document *t* matches the terms of a query *Q*, so that any document *t* can be ranked within the collection *D* in terms of its relevance R to the query *Q*. Therefore from (Equation 3.1) $p\,(a \mid b)$ is the probability of the query term *b* existing in a document *a*. This is then dependent on the probability of a document *b* given a query *a*, $p\,(b \mid a)$ which is a product of the number of times a term $p\,(a)$ appears in the collection $p\,(b)$.

A further demarcation between the concepts of information retrieval and data retrieval is in the query specification which is complete for data retrieval but incomplete in information retrieval. This is evident in the language used to describe results, 'matching' in the former instance and 'relevant' in the latter. The distinction between the two becomes increasingly vague at the query level (or implementation stage) where query terms for data retrieval are largely described as artificial and query terms for information retrieval tend toward natural language.

Recent research (Agrawal et al. 2003; Chaudhuri et al. 2004; De Vries and Wilschut 2004) focused on the migration of this ranking algorithm to a database environment using such methods as *query reformulation* and *automatic ranking* to solve the 'many answers' problem that is characteristic of Boolean database queries. Where information retrieval (document retrieval) lends itself to inductive inference, data retrieval is by nature, deductive and exact as Database Management Systems (DBMS) do not sufficiently support content: "this is in contrast to document based

information retrieval which is relatively unstructured in comparison" (De Vries and Wilschut 2004, 1). Chaudhuri et al. (2004) outlined a 'domain independent' solution to ranking answers to a database query when many tuples are returned and which can be further customized for different applications. In information retrieval there are three traditional sources for the derivation of weights, which represent the strength of a document with respect to a query and thereby used to rank the documents: collection frequency, term frequency and document length (Robertson and Jones 1994).

3.2.6 Inverse Document Frequency

Inverse Document Frequency (IDF) is based on the philosophy that frequently occurring words convey less information about user's needs than rarely occurring words, and thus should be weighted less. Its use is well documented in information retrieval journals (Buscaldi et al. 2005; Chaudhuri et al. 2004; De Vries and Wilschut 2004; Gillam and Tariq 2003; Robertson and Jones 1994) pertaining to document collections. This ranking algorithm has applicability in database applications, by substituting documents with tuples (t).

Collection frequency weights (also known as inverse frequency weights) are defined by Robertson and Jones (1994, 4) as follows, for term $t(i)$:

Given:

n = the number of documents terms $t(i)$ occur in

N = the number of documents in the collection

The Collection Frequency Weight (CFW) for a term given by Robertson and Jones (1994) is given in equation.

$$CFW(i) = \log N - \log n \tag{3.2}$$

The within-document frequency or term frequency is the number of times a term occurs within a document; the rational being that the more often a term occurs in one document, the more likely it is to be important for that document. A term that occurs the same number of times in a short document and in a long one is more likely to be more valuable for the former (Robertson and Jones 1994). By combining the three

weight sources query terms, *local* terms (Equation 3.3) and *global* term occurrences (Equation 3.4) a terms weight for a collection can be derived.

$$DL(j) = \textit{the total of term occurrences in document } d(j) \tag{3.3}$$

$$NDL(j) = \frac{DL(j)}{\textit{Average DL for all documents}} \tag{3.4}$$

The weight of any given word or term (Equation 3.7) in the vector space model is a component of the Term Frequency (TF) and IDF (Equation 3.6) given by Nahm and Mooney (2001).

$$\textit{Term Weight} = \textit{Term Frequency}\,(TF) \times \textit{Inverse Document Frequency}\,(IDF) \tag{3.5}$$

$$IDF = \log(\frac{D}{df_{ik}}) \tag{3.6}$$

$$w_{ik} = tf_{ik} \times \log(\frac{D}{df_{ik}}) \tag{3.7}$$

where:

tf_{ik} is the term frequency,
D is the number of documents in the database
df_{ik} is the number of documents containing term i

The resulting weights are treated as coordinates in vector space, and the dot product for the terms in each document is then compared with the query terms (Nahm and Mooney 2001).

The vector space model (Salton 1983) for ranking search results is an abstraction in which each document can be thought of as a vector. Each document or so-called ‘bag of words’ can be represented as an array of 10,000 counts. This array can be thought of as a point in 10,000-dimensional space where the “distance” between two vectors is the “similarity” of two documents, and a query is represented as a short document

and each document is defined as the cosine of the angle between the document and the query in vector space.

The vector length for each document (Equation 8) is given by Garcia (2006) as:

$$| Document_i | = \sqrt{\sum_i w}^2_{i,j} \tag{3.8}$$

Similarly for the query is:

$$| Query | \sqrt{\sum_i w^2_{Q,j}} \tag{3.9}$$

Then the dot product is:

$$Q \bullet D_i = \sum_i w_{Q,j} w_{i,j} \tag{3.10}$$

The similarity between the query and a document is then calculated. Upon which, the individual documents are then ranked (Equation 3.11) given by (Nie 2001; Garcia 2006):

$$\text{Cosine} \quad \theta_{\text{Di}} = Sim(Q, D\text{i}) = \frac{(\sum_i w_{\text{Q,j}} w_{i,j})}{(\sqrt{\sum_j w^2_{Q,j}} \sqrt{\sum_i w^2_{i,j}})} \tag{3.11}$$

3.2.7 Probabilistic Ranking in Data Retrieval

Agrawal et al. (2003) proposed a method to generate probabilistic rankings by "mimicking" the information retrieval solution by applying the TF x IDF by extending TF x IDF concepts to databases that contain a heterogeneous mix of categorical as well as numerical data based on the frequency of occurrence of attribute values in the database. The domain independent approach (Chaudhuri et al. 2004), applies the probabilistic information retrieval model where each tuple in a single database table is treated as a document. This model focuses on the many answers problem and obviously only tuples that satisfy this condition are considered. Each tuple *t* is partitioned into two parts, $t(X)$ and $t(Y)$, where $t(X)$ is the subset of values corresponding to the attributes in *X*, and $t(Y)$ is the remaining subset of values corresponding to the attributes in *Y*. *X* is the set of attributes specified in the query and *Y* is the remaining set of unspecified attributes (Chaudhuri et al. 2004).

$$Score(t) = \frac{p(t|R)}{p(t|D)} = \frac{p(X,Y|R)}{p(X,Y|D)} = \frac{p(Y|X,R)}{p(Y|X,D)} = \frac{p(Y|R)}{p(Y|X,D)} \quad (3.12)$$

Where *D* represents the collection and *R* represents the set of relevant tuples.

From (Equation 3.12) above, any quantity not involving *Y* is treated as a constant because for the many answers problem only tuples satisfying the query condition are considered of which all contain X values. Therefore (X | R) is removed from the numerator. The resulting final equation (Equation 3.12) above is the ideal set the user has in mind, but was only partially specified in the query. This accounts for the relevant set of tuples which must be ranked between what has been specified and what is ideal.

This theory is best illustrated by an example. Consider a query within an inventory of aircraft in which the requirement is for a Surveillance role with the condition Capacity = 1 and Price = Low. Other possible desirable search criteria may be that Feature = Range or Feature = Radar whilst Features such as Armament may be less desirable. Therefore *p* (Feature =Range | *R*) and *p* (Feature = Radar | *R*) may both be high but *p* (Feature =Armament | *R*) may be relatively low.

Furthermore if there is a general abundance of selected aircraft with Radar as a Feature as compared to Range i.e. *p* (Feature=Radar | Capacity=1, Price=Low, *D*) is larger than *p* (Feature=Range | Capacity=1, Price=Low, *D*)). The final rankings would be aircraft with Feature = Range followed by aircraft with Feature = Radar followed by aircraft with Feature = Armament.

In addition Chaudhuri et al. (2004) continued the derivation of a probabilistic weight by including in this formulae assumption with regards to the associations or independence between attributes. Given a query *Q* and a tuple *t*, the *X* (and *Y*) values within themselves are assumed to be independent, though dependencies between the *X* and *Y* values are allowed. The assumption of limited conditional independence is in many cases false, as in the aforementioned example there is no dependency between low capacity aircraft and range.

Research by Agrawal et al. (2003) reported findings that while IDF similarity works well for some database ranking applications sometimes; its effectiveness is quite limited. In certain instances the relevance of data values for ranking may be due to other factors in addition to their frequencies. Hence ranking functions need to also consider values of unspecified attributes. This is an observation, which is also applicable to information retrieval in the Internet domain.

For purely categorical attributes Agrawal et al. (2003) states that each database tuple (and query) can be treated as a small document and defines a similarity function between tuples and queries. For database applications the IDF therefore is calculated for every value t in the domain of attribute A_k .$IDF_k(t)$ is be defined as $\log(n / F_k(t))$,

where n = number of tuples in the database

$F_k(t)$ = frequency of tuples in database where $A_k = t$

For any pair of values u and v in A_k 's domain, let the quantity $S_k(u,v)$ be defined as $ID_k(u)$ if u = v, and 0 otherwise (Agrawal et al., 2003).

Consider tuple T = <$t1$,…,tm> and query Q = <$q1$,…,qm> (i.e. the latter has a C-condition of the form "WHERE $A1$ = $q1$ AND … AND Am = qm"). The similarity between a tuple T and a query Q is defined in Equation (3.13) by Agrawal et al. (2003) as:

$$SIM(T,Q) = m\sum_{k=1} S_k(t_k, q_k) \qquad (3.13)$$

i.e., similarity between a tuple T and a query Q is simply the sum of corresponding similarity coefficients over all attributes in T.

This is similar to the information retrieval ID formulae although there are some notable differences. The first being that there is no longer a need for the TF, the local or document frequency as each term or attribute can only occur in a tuple once. As an example of indexing in a structured database, for example, a database used to record the available stock of tree species in a nursery, whereby a typical search query could be Australian plants with maximum growth height less than 10 cm. This would inevitably produce many results the first of which would be those matching the exact criteria Australian plants less than 10 cm; the second would be plants less than 10 cm and then Australian plants. This is because there are likely fewer occurrences of plants below 10 cm and hence its IDF is higher than Australian plants which are likely to be many. In addition, Agrawal et al. (2003) observed that databases contain numeric as well as categorical information, and therefore, there is need for a restructuring of the TF and IDF concepts with regards to the results ranking paradigm when applied to database driven IR algorithms.

A simple solution is to discretise the domain of numeric attribute A into buckets, effectively treating a numerical attribute as categorical. Income tax salaries are an example of a bucketing approach, which groups the level of tax you will pay depending on a predetermined range. If numeric data are to be categorised it is important that the frequency of a numeric value should depend on nearby values, implying that cut off points for categories should be natural rather than imposing boundaries on the data which discriminate between two numeric values whose differences are minimal. For example, if we request for a home in a realtor database with price $300k and 10 bedrooms, the price is less important for ranking purposes (there may be many houses priced close to $300k, even if few have exactly that price) than the number of bedrooms (relatively fewer homes have around 10 bedrooms) (Agrawal et al. 2003).

Agrawal et al. (2003) commented that most bucketing approaches are problematic because:

- Inappropriate bucket boundaries may separate two values that are actually close to each other
- Determining the correct number of buckets is not easy.

- Values in different buckets are treated as completely dissimilar, regardless of the actual distance separating the buckets.

There may be instances where the relevance of an attribute value may be due to factors other than the frequency of its occurrence. In such scenarios the current prevailing literature suggests leveraging workloads, using the Query Frequency (QF) similarity philosophy. According to QF similarity, the importance of attribute values is directly related to the frequency of their occurrence in query strings in workload.

The query model assumes that values for all attributes are specified in a query thereby restricting similarity calculations only to the attributes specified by the query. It is suggested by Agrawal et al. (2003) that only when numerous tuples have the same similarity score that missing attributes be used to break ties.

3.3 Conclusions

The various phases involved in retrieving data and the interaction between the main modules of an information retrieval architecture are introduced, whose design centres upon the goal of harmonizing via semantics what is available with what is desired. In an ideal world this would simply comprise a one to one match. However as this is not the case, there are many terms which can be used to describe the same concept, and these may or may not be directly reconciled between source and query. The main hurdles in achieving parity between user and data source are in the information retrieval phases of query expansion and results ranking.

The word matching approach of the vector space model is limited by its directness, in that relevant documents will often not be found simply because the correct word has not been supplied in the query. For example, a document containing the word ‘UNIX’ will be overlooked in a query specified for ‘operating systems’ (Nie 2001). Where conceptually there is an obvious match this is not reflected in either the query or the document. Each concept has been downgraded to a *word*. An *inferential* approach is proposed as a solution to automatically obtaining conceptual matches. As suggested by Nie (2001) the realisation of the vector space model for information retrieval in “classical algorithms assumes a strict correspondence between words and

meanings which is not the case, a meaning may be expressed by different words and words can be used to express different meanings in different contexts" Nie (2001, 8). Information retrieval using unstructured text, natural language and documents, uses various methods in order to derive additional meaning from words, additionally incorporating the concept of degrees of similarity, which can be used to structure results into an ordered structure of relevance or ranking.

In this regard the advantage of the vector space model is that it assigns quantified importance to terms, by the number of times they occur in each document and in the entire set of documents. However, this model assumes that the terms are independent. In addition, it is highly computationally expensive, as it requires that each document is completely searched, and must be searched again each time a new word is added to it. Current trends designed to augment this approach are to add metadata in the form of specifically chosen keywords via XML to documents, via automated metadata creation software or via human means.

It has been shown by Efthimiadis (1996) that there are a number of ways to expand an unsuccessful query in the information retrieval model via means of automated, manual or semi automatic processes. These can be based either on returned results or additional data sources. These however, do not generically apply to the data retrieval model.

Employing the probabilistic ranking methods of the vector space model has been shown to be possible and desirable by Chaudhuri et al. (2004) and Agrawal et al. (2003) via the use of additional attributes and the frequency of tuples to databases records, to solve for the 'many answers problem' in which a database query either returns many results or none at all. Applications of techniques for query expansion in a data retrieval model negate the use of expansion based on results and in this respect further investigation is required in this area, particularly with regards to the use of external supplementary sources.

4 KNOWLEDGE STRUCTURES

This chapter presents an overview of the use of terms used in modern ontology as distinct from philosophical or classical ontology and distinguishes the differences between the two views. As such, the first goal outlined is a detailed description of the distinction between ontology (singular, sometimes called the big O) as a philosophical study, to the use of ontologies (multiple) as a formal modern tool via philosophical definitions, as a pragmatic or realist, nominalist conceptualization. That is not to adopt the view of nominalists in their rejection of the universals, as the goal of usability and interoperability is to attain universal understanding. This is best articulated by Mark and Turk (2003b) who state “ontology is an objective, realist account of the true nature of the landscape, and would be universal across human languages and cultures, although some languages or cultures may emphasize some of these universal properties and ignore others” (Mark and Turk 2003b, 14).

Ontology has several uses. In information systems it can act as a bridge that can integrate terminology between users of varying disciplines and expertise, between different data sources, and between a user and data. This chapter elaborates on the definition of ontology in the terms of a linguistic model, or vocabulary for a specific domain of interest, where a linguistic model is used, as a tool consisting of lexemes and semantics, as opposed to a mathematical or logical one consisting of symbols and operators. In this model nouns correspond to concepts or phenomena in the geographic domain and their relationships to other objects within a hierarchy or taxonomy.

“Formal ontology is an abstraction of the formal features that characterize all scientific areas whilst material ontology is a statement of the necessary and sufficient conditions for something to be a particular kind of entity within a given domain” (Peuquet, Smith, and Brogaard-Pederson 1999, 14). A domain is defined in contrast to an epistemological study, by being restricted to a specific field or scope of study, for example, medicine, botany or geography. The evaluation of ontology touches on epistemology in that (Peuquet, Smith, and Brogaard-Pederson 1999, 16) state that “we cannot know what the world is like via any simple method, we can only

gradually move closer to the truth." However, the comparison of ontological proposals requires a specific form of knowledge, namely knowledge of the world of the sort that is provided by science and also by common-sense experience. In the geographic domain phenomena are examined in terms of their intention. Intention being inherently related to an individual's belief about a phenomenon differs from the traditional or philosophical view of ontology which is not a description of how we conceptualize the world but a description of the world itself. The implication of this view is the acknowledgement that there is no one true reality, but indeed several.

4.1 Introduction

Ontology is a term which appears in a variety of disciplines, such as Software Engineering, Information Systems and Computer Science. As such it is important to delineate the sometimes subtle differences between these academic fields. In the view of Parnas (1998) computer science and software engineering can be seen as two separate but complimentary disciplines. Computer science is the study of the properties of computation in general, while the principal focus of software engineering is the design of specific computations to achieve practical goals. As such software engineering is the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software. The discipline of software engineering encompasses the techniques and procedures regulated by the software development process, with the purpose of improving the reliability and maintainability of software systems. The effort is necessitated by the potential complexity of those systems, which may contain millions of lines of code. Since the 1940's software engineering has evolved from writing machine code on cards, to macro assemblers and interpreters until the first compilers emerged enabling more efficient programming. Modern object orient programming languages such as PHP, C++ and Java were developed to allow more reusability and maintainability of complex code, based on data rather than processes. The fundamental concept of Object Oriented (OO) programming languages is that of the class, whose definitions formulate the abstract characteristics of a thing or object denoted by its attributes(fields or properties, elevation, composition etc), and methods(operations and features). Techniques of OO Software include encapsulation, modularity, polymorphism and inheritance. Information systems originated as a sub-discipline of

computer science in an attempt to understand and rationalize the management of technology within organizations. A computer based information system is defined by Langefors (1973, 1) as “a technologically implemented medium for recording, storing, and disseminating linguistic expressions, as well as for drawing conclusions from such expressions.” It is apparent that the disciplines of software engineering and information systems are intimately intertwined, and the term computer science with regards to discussion involving ontology is often used to encompass the two.

Guarino (1998, 1) states that “in some cases, the term ontology is just a fancy name denoting the result of familiar activities like conceptual analysis and domain modelling carried out by means of standard methodologies in software engineering.” Traditionally these methodologies involve gaining an understanding of the problem domain, as a design phase carried out prior to the development of a conceptual model to which the software, its inputs, outputs and transactions are describe based on the domain logic. For example, a car parts inventory may describe a car in terms of is components, engine type, wheels, chassis, etc. Each of these parts may consist of a unique serial number, manufacturer and year, and whether it is used or new. These factors could then be used as a basis for determining the sale price of the part. In software engineering and information systems the direct result of this study is a model of entities and their relationships between each other, which can be used in the task of designing a database or objects (OO programming) in the implementation phase.

Section (3.2) highlighted the problem of abstraction and loss of information that occurs in moving from a real world to a digital representation of that world. From the architectural perspective, there is the centrality in the role that an ontology can play in an information system, leading to the use of ontology-driven information systems. Ontology driven information systems can play a central role by impacting the main components of an information system: information resources, user interfaces, and application programs, and is base on some agreement between user groups and designers about a perception of the world. As such ontology refers to “an engineering artefact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words” (Guarino, 1998, 2).

Guarino (1998) observed that in information systems and software engineering the majority of ontologies present their own methodological and architectural peculiarities. From a methodological perspective, the main peculiarity found by Guarino (1998) is the adoption of a highly interdisciplinary approach, where philosophy and linguistics play a fundamental role in analysing the structure of a given reality at a high level of generality and in formulating a clear and rigorous vocabulary (Guarino, 1998). As such a conceptualization is defined in Guarino (1998) by <D, R> where *D* is a domain and *R* is a set of relevant relations on *D*. According to Guarino (1998) a conceptualization is semantic and inherently intensional, it can be applied to model any aspect of the real world, so that one can state the intended meaning of terms used to explicitly indicate the relevant relations so that there can be no ambiguity. It is typical of modern software design theory which emphasizes the user as being integral to system design as opposed to having to adapt to the systems representation of reality (Section 3.2). “In the domain of information systems and software engineering ontologies, truth, or fidelity to reality, are not really issues, instead, the conceptualizations embedded in the information system must be faithful to the conceptualizations held by the clients or users” (Mark and Turk, 2003b, 4).

4.2 Philosophical Background

The term ‘ontology’ is being used with increased prevalence in the field of information retrieval (Guarino 1998; Smith and Mark 2000). Historically *ontology* has its origins in philosophy (Probst 2007; Cocchiarella 2001). At the highest level of abstraction ontology is a model of reality and the concepts and the relationships amongst the objects that exist within it.

The actual term, first appeared in the early 1600s (Øhrstrøm, Uckelman and Schärfe 2007) but the concept is thought to date to early Greece and the works of Aristotle and Plato as the ‘study of logic’ with regard to the rules for discussion and reasoning (Cocchiarella 2001). Later, evolving towards the formal ontology’s as a part of science as opposed to philosophy, as a structured system of categorical analysis or primary theory, applied to domains such as naïve physics and living things (Carnap

1950). Whilst there was little application for the logic of ontology in the 19th century it has become a major field of study in academia for the 21st century primarily in the integration of databases and the definitions of terms.

Ontology in the classical sense takes what contemporary philosophers would term a universal approach. It is the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality. It is as such an objective view of reality. “In contemporary terms this is the so-called domain independent or ‘general ontology’ which promotes an objective ‘ideal’ in which concepts have only one meaning” (Thellefsen, 2004, 5).

“Domain-dependent ontologies concern categorically closed regions of being” (Poli, 2003, 1). The disadvantage of ‘general type’ systems is that they are usually only implicitly and informally defined with the consequence that the ontology is affected by a high level of vagueness and ambiguity (Sowa, 2000).

The *interpretive view* is that concepts are based on human cognition and consequently promote a *constructivist* approach whereby concepts are constructed to serve a purpose, including certain individual and social values and interests (Thellefsen 2004). Universal or neutral (in terms of intension) knowledge information structures are in many ways unattainable, because of the nature of concepts and consequently will always reflect a point of view dependent on factors such as time, place and purpose or perspective of interest.

4.3 Logic and Metaphysical Categorization

Aristotle first presented logic in the form of ontology in his so-called *topics*, *metaphysics* and the *categories* as a study of the common properties of all entities, and the categorical aspects in which they can be analysed. Until this time ‘a formal logic of propositions’ was only conceived of as a set of rules for arguing, such questions as the nature of existence. Aristotle’s doctrine of categorization assumes that all entities can be classified into a limited number of ultimate classes. Essentially it outlines a logical way of viewing the world through a common model, in which new experiences could be classified and understood within this recognized

vocabulary. It is the overarching requirement for such a common model or world view which is seen as the holy grail of research to solve the usability issues in GIS.

Philosophical ontology asks questions such as ‘what is a bird?’ or ‘what is a star?’ more pertinently to this research, ‘what is a mountain?’ The way that these questions are answered reflects how we perceive and interact with the world.

Logic based ontology is a logical theory (Copi 1979). The terms of the terminology, whose semantics are to be specified, appear as names, predicate and relation symbols of the formal language. Logical axioms and definitions are then added to express relationships between the entities, classes, and relations denoted by those symbols.

Aristotle’s *topics metaphysics* and the *categories* it is presumed constitute an exhaustive division of ‘kinds’ from which everything in the world can be understood. In logical terms importantly, and in a broader sense it can also be the study of what might exist (Smith 2003, 1). There are various philosophical schools of thought with relation to how logic within such different models views reality and categorisation. The essential objects of philosophical or scientific models are belief systems as to how the world is perceived which can then be formalised into axioms. From the early works of Aristotle and Plato several, alternate perspectives have emerged within these constructs of viewing, categorising and communicating with and about objects and events in the world.

“Through the axioms and definitions, the semantics of the terminology is specified by admitting or rejecting certain interpretations” (Bittner, Donnelly and Winter 2005, 4). As such the axioms of logic based ontology specifies meaning by rejecting interpretations that do not conform to the intended use of the underlying terminology.

Grenon (2003) suggested that the main philosophical approach to contemporary knowledge representation is *pragmatic conceptualization*. The philosophy of *pragmatic conceptualization,* purports that knowledge representation should be a representation of reality and not systems of representation of concepts and models.

Ontology is classically differentiated from logic as ‘concepts abstracted directly from physical reality,’ whilst logic is about concepts abstracted wholly from the ‘material’ content of first intension, as well as about such categorical concepts as individual, proposition, universal, genus, species, property, etc. and so-called syncategorematic concepts such as negation. “So-called second intensions have a foundation in real entities, but ‘exist’ only in the mind, which is not to say that they are subjective mental entities” (Grenon, 2003, 5). In this manner Grenon’s (2003) *pragmatic conceptualization* as a description for modern knowledge representation can also be described as a ‘first order intensional logic’.

In classical first order logic intension plays no role. It is extensional by design since primarily it evolved to model the reasoning needed in mathematics, whose equations can be said to model intensions in terms of relationship. For example, an intensional definition of ‘prime minister’ might be the most senior minister of a cabinet in the executive branch of government in a parliamentary system; whereas an extensional definition, also called a denotative definition, would be simply a list of all past and present prime ministers.

Extensional relations reflect a particular state of affairs, which are categorical structures. For instance in the ‘blocks world’ they represent a particular arrangement of blocks on the table (Guarino 1998). In the view of Armstrong (1997), whenever a *particular x* instantiates a *universal u*, there is a state of affairs with *x* and *u* as its constituents. Moreover, it is only in this case that there is a state of affairs with *x* and *u* as constituents. For a *particular x* to instantiate a *universal u* is for it to be the case that there is a state of affairs with *x* as its *particular* constituent, and *u* as its *universal* constituent. There are relational states of affairs as well as monadic ones: they have more than one *particular* constituent.

4.3.1 Intensional Logic

Meanings are often called *intensions* and things designated *extensions*. Contexts in which extension is all that matters are naturally *extensional* while contexts in which extension is not enough are *intensional*.

The terms intension and extension are derived from Carnap (1958) whilst Frege (cited in Berkovski, 2006) used the terms ‘sense’ and ‘reference’. However expressed, and with the variation from author to author, the essential dichotomy is that between what a term *means* and what it *denotes*. Intension as defined in (Carnap 1958; Fitting 2001) as being ‘non-rigid designators’ which designate different objects under different circumstances or by functions from different domain views or with regards to the ‘ontological commitment’ of the domain. Where to be less ‘ontologically committed’ implies an emphasis on upper level ontological terms and generalisation, as opposed to granularity and concept depth.

Alonzo Church (1956) introduced the notion of *concept*, where anything that is the sense of a name for something can serve as a concept of that something. Simply formalised this is saying something of how intensions behaved, without saying what they were. It is thus explicit that concepts are language independent, and might even be uncountable.

4.4 Model Theoretic Semantics and Linguistic Ontology

Modern era philosophers such as Rudolf Carnap suggested that one crucial project in philosophy is to develop *frameworks* that can be used by scientists to formulate theories of the world (Carnap 1956). Such frameworks are as such formal languages that have a clearly defined relationship to experience or empirical evidence as part of their semantics. A conceptualization can thus be viewed as an intensional semantic structure, which encodes the implicit rules constraining the structure of a piece of reality or the action of building such a structure (Guarino 1995). Therefore in accordance with Gruber (1993) in that, ontology is a logical theory, which gives an explicit partial account of a conceptualization; the aim of which is to define which primitives provided with their associated semantics, are necessary for knowledge representation in a given context. In addition there are several methods of representing a knowledge system through logical formalisms such as first order predicate calculus that has the advantage of allowing explicit representation of the objects in the relevant domain of discourse (Grenon 2003).

4.4.1 Predicates

The ontology of a theory or of a formal model is tightly connected to the use of predicates (concepts, attributes, features, relations) applying to the entities or relating the entities, and to criteria of identity and distinction that allow us to recognize an entity. Consider the statement taking the form P (*a*), where *a,* is the subject and P is the predicate where the subject denotes an object, i.e. something that is independently saturated (distinctly identifying attributes), ontologically self-sufficient and complete, and which does not need to be determined further.

The predicate is something that is intrinsically unsaturated and which requires the noun to acquire completeness. A predicate is either a relation or the Boolean valued function that amounts to the characteristic function of such a relation. As such predicates correspond to concepts and universals are domain neutral concepts. By means of normalization we obtain a situation of the type F (P), where the original predicate appears in the guise of a noun in the subjective position.

Non-saturation thus exists in the purely dispositional state of cognitive capacities. The exercise of these capacities informs the mental act consisting in the combined application and mutual saturation of a referential concept and a predicative concept. The non-saturatedness of the concepts also means that they may stand as logical subjects by means of their individual representations.

4.4.2 Nominalism

Nominalism is best understood in contrast to the form of *realism* advocated by some medieval philosophers. This form of *realism*, which is quite distinct from realism in the modern sense, holds that if we use descriptive terms such as "green" or "tree," the forms of those concepts really exist, independently of the world in an abstract realm. Such thought is associated with Plato, for instance, *nominalism,* holds that ideas represented by words have no real existence beyond our imaginations.

Nominalism is the preferred paradigm for contemporary ontology as language. *Nominalization* refers to the use of a verb or an adjective as a noun and as such *nominalists* reject the reality of universals. In this case P is no longer unsaturated as

it was in P (*a*), but it has the same subjective characteristics as *a*, except that corresponding denotatively to *a*, is an entity in the universe of discourse.

Objects denoted by *nominalized predicates* are intensional entities, or in other words, properties and relations, which have their own, abstract form of individuality. A concept as such, in that it lacks individual characteristics, cannot be part of any universe of discourse. With the nominalization of P in the sense of 'the concept P' one obtains an individual term in the theory of concepts that denotes not concepts but special objects (for a given domain).

4.5 Modern Ontology in Philosophical Context

The process of ontological theorizing as described in the work of Quine (1953, cited in Smith and Mark 2001) seeks to elicit ontologies from scientific disciplines. Also known as the study of *internal metaphysics* or the study of ontological commitments of specific theories or systems of beliefs. This theorising takes the form of the logical methods in the analyses of a system and the so-called ontological commitments or presuppositions embodied in a scientific discipline that represent the intension of the concepts in the domain (Mark, Skupin and Smith 2001, 2).

Philosophical ontology has come to be replaced by the study of how a given language or science conceptualises a given domain. Gruber (1993, 2) defines ontology in the context of its use in computer science as a "specification of a conceptualization."

In information systems and software engineering development, an ontology is a representation of some pre-existing domain of reality which; 1. Reflects the properties of the objects within its domain in such a way that there obtains a systematic correlation between reality and the representation itself. 2. It is intelligible to a domain expert, and, 3. Is formalised in a way that allows it to support automatic information processing (Pressman 2001). As such the process of domain analysis is the identification, analysis and specification of common requirements from a specific application domain, typically for reuse on multiple projects with the application

domain (Pressman 2001). In an Object-Oriented analysis this involves the definition of semantic classes (based on customer requirements) from which a taxonomy and the relationships which connect these classes can then be modelled (Pressman 2001).

To the computer scientist, the aforementioned domain of reality is that which can be represented, including the processes abstract in as much as they may only exist as specific events within a small window of time. Peuquet, Smith, and Brogaard-Pederson (1999) state that the fundamental problem of data representation is that of inferring past processes from observed spatial data. The time space between the observation of a spatial data representative and the processes and forces that determine its formation, are often greatly varied in time scales. For example a river created by seasonal flash flooding, which may only exist for weeks at a time. From the perspective of the computer scientist (information scientist or software engineer) the key objectives are modelling information flow, and information structure. Such modeling processes are often limited to the domain or the community to which they are designed, and as such have limited potential for reusability or expansion to a wider community of users. Ontologies in the information systems sense are formalizations of conceptualizations, suitable for implementation in information systems and software models, used to capture knowledge in a given domain. The philosophy of *conceptualism* lends itself to the representation of ideas giving a major role to epistemology (general ontology or domain neutral ontology), philosophically the study of being and existence and foundations of metaphysics. In information systems the conceptualization involved could be a philosopher's "ontology" as described in, or it could be an epistemology, a collection of beliefs about the landscape and knowledge of the landscape, possibly containing imaginary or false conceptualizations.

4.6 Taxonomy

Common to most ontologies if not all is a taxonomic structure that employs an inheritance relation (also known as the IS-A relation or the Hypernymy/Hyponymy subsumbtion relation) between concepts. Taxonomies play an important role in conceptualising knowledge. “In information science the ontology stipulates the taxonomy that forms the basis of a data dictionary used in building an information system” (Mark and Turk, 2003b, 2). Aristotle organised ‘kinds’ or ‘categories‘

hierarchically in the form of a tree, the lower nodes of which were called ‘species’, and the upper nodes ‘genera’. Similarly in the way that in Section 3.2.3 Stenmark (2003) modeled bird as a subclass of animal (genera) and pengu as an instance (species) of bird. A taxonomy as a product of a process, is by definition *empirical* in nature and a classification hierarchy based on the is-a relation, in which “the greater the inclusiveness of a category, the higher its level of abstraction” (Rosch 1978). The terms hypernym and hyponym are the inverse of each other for example as the hypernym of poodle is dog, then a hyponym of dog is poodle, as such they are transitive relationships between sets of synonyms (Fonseca 2001). Via the IS-A relationship hyponyms of concepts inherit the properties of the concepts to which they are hyponyms an in define additional attributes. In the *empiricist* philosophy “the non experienceable is not real or at least not recognisable” (Wyssusek, Schwartz, and Kremberg 2001, 2). The taxonomy is a structure that imposes a conceptualization of the world an as such its correspondence with a truth can often be disputed, therefore an ideal taxonomy of the real world is widely considered unreachable. Smith (2001) stipulated three principles required for an ideal taxonomy:

1. A taxonomy should take the form of a tree in the mathematical sense.
2. A taxonomy should have a basis in minimal nodes, representing lowest categories in which no sub-categories are included.
3. A taxonomy should be unified in the sense that it should have a single top-most or maximal node, representing the maximum category.

The first principle guarantees that the tree corresponds to a connected graph without cycles where the nodes of the tree represent categories at greater and lesser levels of generality, and branches connecting nodes represent the relations of inclusion of a lower category in a higher (Smith 2001). Defined in this way a category may never be subordinate to more than one other category in the tree. Each category much capture the general features pertaining to its subordinates, and each subordinate must specify its differentiating features. As Smith (2001) notes this definition may be to strict to be applied to real world knowledge:

”Certainly it is useful for some purposes to employ taxonomies

which depart from the tree structure by placing a given category simultaneously on a number of separate branches within a hierarchy in such a way that it inherits information from each branch" (Smith 2001, 12).

The third and final principle states that a single omnipresent category should exist that subsumes all subordinate nodes. The justification is that a taxonomy with two maximal nodes would be in need of completion by some other node representing the union of the two.

"Otherwise it would not be one taxonomy at all, but rather two separate and perhaps competing taxonomies, the claims of each would need to be considered in their own right." Smith (2001, 12)

For geographic entities, categories may in part reflect similarities and discontinuities in the landscape, but to some extent are projected onto the landscape by human cognition and language. Mark and Turk (2003b) defined the study of ethnophysiography as having objectives similar to ethnobiology, which studies folk names and categories for plants and animals, but differs in important ways. Where ethnobiology often uses scientific taxonomy as a baseline for assessing folk categories for plants and animals; the variation in landforms and waterbodies is not constrained by mind-independent natural kinds in any obvious way. As such the approach of ethnophysiography by (Mark and Turk 2003b) to geographic ontology is not based on a scientific taxonomic analysis of similarities of geographic features, but rather an examination of categorization of an inorganic natural domain. There are significant differences between Ontology in the domains of biology or physics as compared with Ontology in the geographic domain. An important aspect of the ontology of a domain is the categories to which entities can belong, and the relations among those categories (Mark and Turk 2003b). According to Rosch (1978) categories are central to cognition. Ontology identifies observable properties or attributes, such as size, shape, or curvature, in the geographic domain however, the problem of Ontology must contend with identity and categorization by human cognition via language. In botany for example it is easy to delineate the singularity of an entity. A particular plant species has obvious physical boundaries however

geographic features are often human imposed concepts describing phenomenon in continuous space.

4.7 Mereology

A partonomy involves classification based on the part-of relation and is distinct from taxonomy that is based on similarities (is-A). The theory of parts and the whole, or mereology, as it concerns formal ontology is concerned with the pure theory of independence and non-independence.

Mereology is the general theory of collective sets, or parts, which are generally classified into extensional and intensional. The extension of a term is the set of objects that that term denotes. The term extensional is ontologically monistic as universals by definition are non-extensional. As it were to being in a specific domain or pertaining to a singular concept, every object that exists is an object; the parts of objects are objects and compositions of objects are objects as well.

Mereology is formally defined by an axiomatic formulation of reflexivity (Equation 4.1), antisymmetry (Equation 4.2) and transitivity (Equation 4.3) in which Schneider (2002) defines parthood (P) within a possible world of maximal particulars containing maximal elements of equivalent class (CM^{ϵ}_{p}) under a relationship of compossibility. Genesereth and Fikes (1992) explain that to constrain the meaning from all possible worlds fictional or not, a greater number of expressions must be written to resolve any ambiguity. Mereology assumes a further relation between counterparts, namely compossibility, by leaving out any mention of worlds by "*collapsing* accessibility and counterpart relations onto each other" (Schneider, 2002, 30). In mereology, a part Pxy is true when x is any sort of part y, then:

$$CM^{\epsilon}_{p}\ P\ xx \rightarrow P xx \tag{4.1}$$

$$P xy \wedge P yx \rightarrow x = y \tag{4.2}$$

$$P xy \wedge P yz \rightarrow P xz \tag{4.3}$$

Reflexivity (Equation 4.1) is the rule of part hood that requires everything be a part of itself by definition. Antisymmetry (Equation 4.2) is the proposition that the

relationship of part hood is singular, that is, two distinct things cannot be a part of each other. Finally, the axiom of transitivity (Equation 4.3) requires that any part of any part of a thing is itself part of that thing.

Intensional mereologies as such distinguish the entity into independent and non-independent parts. The former are termed ‘pieces’ and they effectively assume the denomination ‘part’, while the latter are called ‘moments’. For example, in the two expressions, *the capital city of Poland*, and *Warsaw*. *Warsaw* being the label of an individual is independent of state and time moments. The former clearly illustrates the dependency of the time moment, inferring that Warsaw has not always been the capital of Poland and is therefore *intensional*. Moments may therefore be equal to the whole of which they are moments. However, the concept of equality should be understood in the sense of *indiscernibility*. The possible indiscernibility of the moments from the whole should not be confused with the possible identity of the part (as distinct from the proper part) with the whole. A part may even be the whole itself, while a moment can at most coincide with the whole; or in other words, it may be discernible from the whole but it is nonetheless distinct from it.

4.7.1 Geographic Partonomy

Whether one views space as *absolutist* or *relational* has been the focus of intense debate in philosophy (Nerlich 1994, 4), and has immediate consequences for a representational theory in GIS. Ontology in the geographic domain must effectively quantify, prioritise and account for spatial relationships and their representations. In GIS the choice described by Casati, Smith and Varzi (1998) is between a discrete categorisation of independent individuals (a sort of container) between objects, or as a contiguous representation. The argument is essentially one of how to move from singular regions, or objects in space to geographic space in its entirety.

Smith and Mark (1996) suggest a mereological theory of parts and wholes is essential in defining a geographic ontology. In this domain, spatial, morphological and functional, intersection theory and geometrical nature must be considered.

In information science the value of creating partonomies in the geographic domain is in geographic queries, as partonomic relationships reveal the interdependence

between phenomena at different levels of detail, this relationship adds context to a query and can be exploited in query expansion. Unlike taxonomies an object can be part of more than one object as such multiple partonomies are useful in database transformation from one level of detail to lower levels of detail. Research by Smith and Bittner (2000) classifies granular partitions along three mereological axes; the first, by the degree to which a partition represents the mereological structure of the domain it is projected to; the second by the degree of completeness and exhaustiveness with which a partition represents reality; and third by the degree of redundancy in the partition structure.

4.8 Topology

Topology comes from the Greek word 'topos' meaning the *study of place*, and is a large branch of mathematics that includes many subfields. Mark, Smith and Tversky (1999) define topology in the geographic domain as the theory of boundary, contact and separation to describe the relationships between geographic features. Topologically two or more regions or features can be defined as contiguous if they share a border.

"Topology by itself is too fixed and is thus a very limited criteria for GIS to model spatial relationships" (Casati, Smith and Varzi 1998, 8). In ontological terms nearness is not perceived as a relationship among entities in space but rather an attribute of them, depending on other attributes and an (implicit) measurement of distance. Casati and Varzi (1997) stated that the limitations of topology are significant even at a fairly elementary level, "a long way before functional features become important for classifying shapes or providing an analysis of such relationships as containment" Casati and Varzi (1997, 10).

4.9 Mereotopology

Smith and Mark (1998, 2) assert that the problem in using set-theoretical models to account for geographic categories is that some members are better examples of a class than others. As such Casati, Smith and Varzi (1998) suggests mereology, as a

powerful alternative to set theory, and further, that it is especially suited for spatial and geographic representation.

However Casati and Varzi (1997) stated that mereology alone is too weak as it is ontologically neutral and in itself not sufficient to describe the domain of geography without consideration of topology. "We need mereology because topology is mereologically unsophisticated. We need topology because mereology is topologically blind, and we need ontology because both topology and mereology – even if we try to relax or supplement the primitives, are intrinsically incapable of making sense of important ontological distinctions" (Casati and Varzi, 1997, 20).

Mereotopology may thus be viewed in this context as consisting of two independent but mutually related components: a mereological component, concerned with the concept of part hood (or overlap), and a topological component, concerned with the concept of wholeness (or connection). Varzi (1998) defined the axioms for boundaries in mereotopology such that a more flexible approach can be achieved to manage the classification and taxonomy of the data by describing the *partonomy* of the data in meaningful ontological concepts:

x is a boundary part of $y =_{df}$ if every part of *x* is connected to the complement of *y*.
x is a boundary of $y =_{df}$ if *x* is a boundary part either of *y* or of the complement of *y*

"Mereological part-hood (P) and topological connection (C) are important axioms for defining mereotopological primitives required in order to recast set theory and point-set-theoretic topology in nominalistic terms" (Casati, Smith and Varzi, 1998, 5). Part-hood therefore must be a partial ordering while connection must be symmetric, reflexive, and monotonic relative to part-hood.

Identity in mereology as a limit case is symbolized as a binary predicate '≤' the primitive of which (Casati, Smith and Varzi 1998) suggested should include a three-place relation involving a temporal parameter. The mereotopological axioms of part-hood and connection are thus given by (Casati, Smith and Varzi 1998) which add to the mereological axioms of reflexivity, antisymmetry and transitivity, the axioms of

connection reflexivity and connection symmetry. *Part-reflexivity* (Equation 4.4) means everything is part of itself. *Part-antisymmetry* (Equation 4.5) means two distinct things cannot be part of each other. *Part-transitivity* (4.6) means any part of a part of a thing is itself part of that thing. *Connection-reflexivity* means everything is connected to itself. *Connection-symmetry* means everything is connected to anything, to which its parts are connected, or a thing is connected to a second thing, the second is connected to the first.

$$x \leq x \tag{4.4}$$

$$x \leq y \wedge y \leq x \rightarrow x = y \tag{4.5}$$

$$x \leq y \wedge y \leq z \rightarrow x \leq z \tag{4.6}$$

4.10 Conclusions

This chapter has briefly covered ontology as a structure which is the manifestation of a conceptualisation of the world which can have many points of view that are open to endless debate. An ontology can be viewed as a set of beliefs about the world, and is often accompanied by a taxonomic structure which orders concepts into a hierarchy. Where a taxonomy is a structure which relates concepts via the hypernym/hyponym relation (IS-A), mereology is slightly more complex and concerns itself with the study of independence between parts. In addition topology and mereotopology have been discussed as structures which describe the relationships between geographic features. Whether such structures such as is-A and part-of are a natural result of human cognition is in the context of this paper not as important as how this structure can express a conceptualisation of a domain of interest. This chapter has examined at an abstract level how the world can be decomposed into entities which can be organized into structures without examining the how and why people or different groups of people cognize their environment. This is represented in the language used by different communities to describe their environment and is the subject of the next chapter.

5 GEOGRAPHIC CATEGORIES

The issue of interoperability dominates the information age as the internet has facilitated the globalisation of data and accessibility to datasets. Ontology is a well recognized tool in research toward the development of theories and techniques to facilitate interoperability in GIS. As such recognizing and understanding the various philosophical traditions and worldviews is important when discussing the use of ontology for interoperability. This chapter is a review of the current methods, techniques and views to which spatial information can be categorized.

The study of whether there are certain cognitive universals affecting the way we form geographic categories, has been a long standing scientific and philosophical argument. According to Rosch (1978) people do not recognize various objects by classifying them under some concept, but by comparing them to prototypes that serve as concrete examples of the category in question. Furthermore, there are no sufficient and necessary criteria by which to determine whether an object belongs to a specific category or not, and, thus, belonging to a category is not a simple either/or question.

The complex reality of geography is best described by a continuous model, but this runs contrary to our human need for categorization which Rosch (1978) describes as “not an arbitrary product of historical accident or whim but of a result of psychological principles of categorization, which are subject to investigation” (Rosch, 1978, 1). The formation of categories and category systems is according to Rosch (1978, 2) “to provide maximum information with the least cognitive effort”. The structure of the information then asserts that the perceived world comes as structured information rather than as arbitrary or unpredictable attributes. Thus, “maximum information with least cognitive effort is achieved if categories map the perceived world structure as closely as possible” (Rosch 1978, 2).

The interdisciplinary nature of depicting abstract geographic categories is supported by Fabrikant (2001) who suggests that most spatializations or information spaces are generated by researchers outside of GIScience and cartography. The use of use of conceptualizations of a domain (COAD’s) is suggested by Mark and Turk (2003a) to overcome the difficulties of interdisciplinary discussion regarding the key aspects of

ontology. In recognizing rather than dispelling any particular belief or alternative definition about ontology Mark and Turk (2003a) suggest the use of COAD's which reflect the different ways in which a domain of interest can be categorized, and the various levels of existence and awareness about the world. The following chapter is an analysis of some of the predominant approaches to the geographic categorization similar to the research of Mark and Turk (2003a) with the intention of understanding the many ways in which geographic categories are created.

5.1 Geographic Categories in the Representation Universe

Geographic data can be modeled as objects or fields. "Most geographic information scientists will generally accept the distinction between 'entities' in the real world and 'objects' as representations of them" (Kemp and Vckovsky 1998, 2). The object model represents the world as a surface occupied by discrete, identifiable features with a geometrical representation (point, line or polygon) and descriptive properties. The field model views geographic reality as a set of spatial distributions over geographic space, for example climate and vegetation cover is usually represented in this way. Objects are primarily identified by their non-spatial and temporal characteristics and then attributed with their spatial (and temporal) extension. The measurement and representation of fields usually firstly identifies the spatial and temporal component (the element of the domain) and then associates the (non-spatial) field value (Kemp and Vckovsky 1998).

Depictions of complex and abstract data domains, that is involving the transition from the cognitive and logical universes to the representation universe (Fonseca et al. 2002a) are often based on a geographic metaphor, and are commonly referred to as spatializations or information spaces (Fabrikant 2001). "The shift from the logical to the representation universe is where a finite symbolic description of the elements in the logical universe is made so that we can apply operation such as search, query and retrieval on them" (Fonseca et al. 2002a, 6). It is in this transition where the areas of ambiguity concerning whether certain phenomenon are fields or objects and distinction between the two is resolved.

“Similes and metaphors are the verbal expression of similarity, and by understanding the natural groups prevalent in complex environments then we may be able to understand better other entities within the same natural group” (Holt 2001, 1). “Metaphors are often used to describe spatial information because a metaphor can display reality better than reality itself” (Kardos, Moore and Benwell 2003, 3). “Spatial metaphors therefore, structure many of our complex ideas about life, for example ‘on the one hand’, ‘on the other hand’ are spatial metaphors” (Holt 2001). Holt (2001) states that spatial similarity can be used to help retrieve:

- Cases according to their geographic location
- Cases according to their topographical relations
- Cases according to their abstract values

The use of metaphors will be discussed further in the following chapters. Fabrikant (2001) applies the process of spatialization to region and scale in which regions are spatial taxonomies, based on perceived similarity/difference of a phenomenon’s characteristics; that is, regions are partitions in space. In this situation Fabrikant (2001) applies scale as the granularity (or ‘mesh size’) of the partitioning. The finer the partitioning scheme, the higher level of detail can be identified, and the more regions will be seen in an environment.

Fabrikant (2001) used the technique of latent semantic indexing via the vector space model (Section 3.2.6) to project geographic regions obtained from Reuter’s news stories onto Euclidian information space (Figure 5.1). In the diagram, documents and terms that are closely associated semantically are placed near each other in Euclidean vector space. In the example, individual news articles are landmarks (regional centers) created by geographic terms and term co-occurrences within the document which are projected into a Euclidean information space in which the connections between nodes are based on the semantic relationships between them. The nodes are distributed using a Voronoi tessellation whereby the node locations are used as region centroids to partition the information space and to capture the cognitive concept of center-periphery.

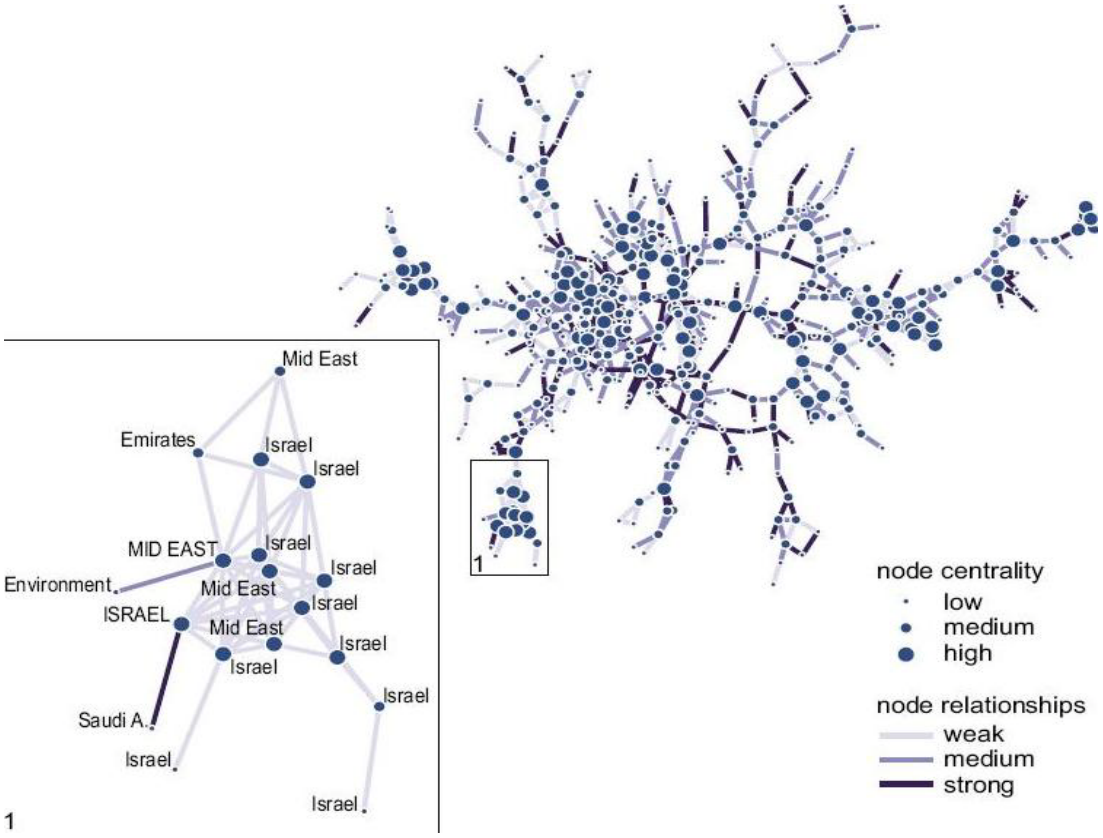


Figure 5.1: A Semantic neighbourhood created via Latent Semantic Indexing of Reuters news article content (Fabrikant 2001).

Fabrikant (2001) extends the concept of information visualisation in the Reuters news space from a two to a three dimensional space by incorporating density into the semantic model. Fabrikant (2001) shows how a hierarchically embedded cluster of terms of aggregated spatial units based on the similarity of document content can serve as a method to depict scale dependent topic regions in news space and identify the level of granularity at which the information space can be explored (Figure 5.2). Using latent semantic indexing the occurrence matrix generated by the occurrences of each term in a document, describes a relation between the terms and some *concepts*, and the relation between those concepts and the documents. Thus terms and documents are then indirectly related through the concepts they are composed of.

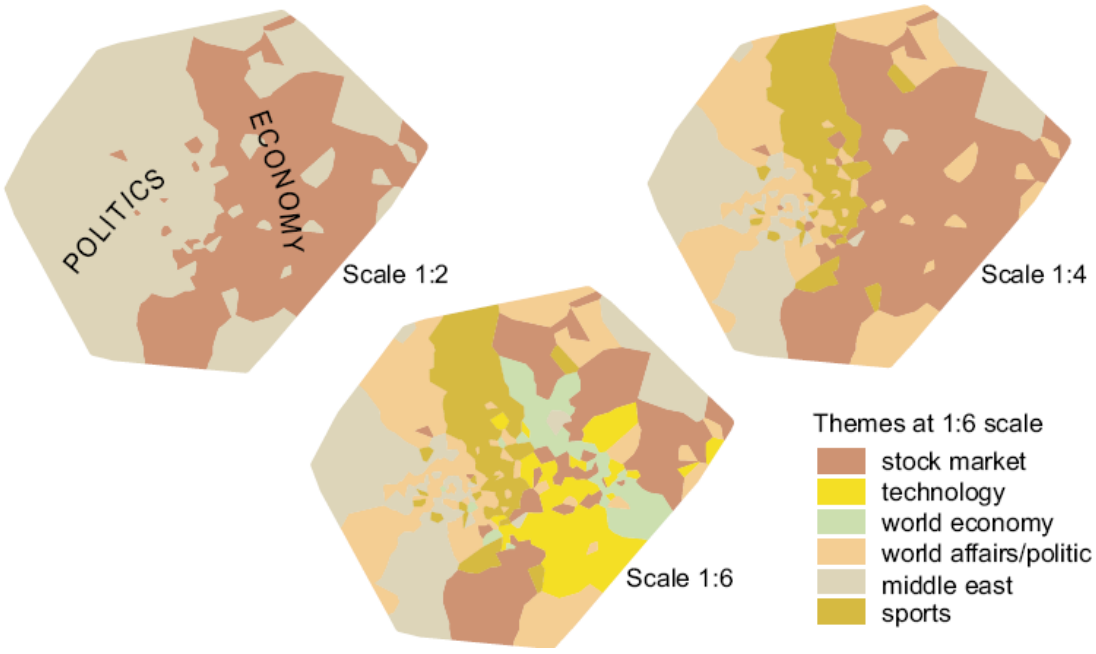


Figure 5.2: Varying levels of semantic detail in three dimensional space (2D geographical space and the third dimension concept density) created via latent semantic indexing (Fabrikant 2001).

The visualisation method of Fabrikant (2001) allows exploration based on varying scales of concept density, where the largest scale (1:2) is conceptually less complex than at scale 1:6 in which less frequently occurring terms such as technology and world economy are visible. Intrinsically connected to the notion of identity is the concept of the boundary. “Objects do not merely have constituent parts, they also have boundaries, which contribute as much to their ontological make-up as do the constituents they comprehend in their interiors” (Smith and Varzi 2000, 1). Boundaries cannot exist in isolation from their hosts; however boundaries can be both physically real and imposed by cognition. Region and scale can thus be more formally described by axioms of mereotopology. “Discontinuities (i.e., boundaries) in the environment separate zones of relative homogeneity (i.e., regions)” (Fabrikant 2001, 2). The functional regions used by Fabrikant (2001) are semantic partitions of space, which only exist dependent on humans’ cognitive capabilities of representation. These functional regions contain semantically similar entities (i.e.,

fiat objects) separated from different entities by fiat boundaries. Political and economic boundaries are examples of the fiat kind.

Bona fide boundaries refer to genuine discontinuities, a partitioning of a landscape based on tangible, physical entities which exist regardless of human cognitive agency such as continents, rivers, and forests (Smith 1995). According to Fabrikant (2001) bona fide objects are implicitly organized in nested hierarchies, simply by applying the size rule (spatial extent) and the containment principle. For example, a soil sample is part of a patch of land, the patch is part of a region, the region is part of a continent, and so on. An organizational chart of a business corporation, or the chain of command in an army are nested examples of semantic hierarchies. These cognitive constructs match the ontological primitive 'part-whole' in mereotopology. At a global political scale we have a hierarchy of regions to the world is delineated into countries (e.g. Australia) which consist of states (Western Australia) which have shires (Donnybrook) etc.

Fiat boundaries are useful to their creators as they impose a reference frame which allows operations to be applied to large datasets. In geography, a reference frame is usually based on a formalized coordinate system. "Once the construction details of a chosen frame of reference are known, scale change can be identified and measured" (Fabrikant 2001, 2). Generally defined, scale relates to levels of organization, or hierarchies. Features on the same level of the hierarchy therefore have the same scale. In GIScience, scale can be referred to as (1) the level of detail (resolution) of a phenomenon under study (e.g., 30m sampling interval of a digital elevation model), (2) as level of abstraction or spatial extent (e.g., a footprint at 1:200,000 scale), (3) as a level of human point of observation (i.e. body space, geographic space), or (4) as a purely semantic level, for example nested enumeration units, delineated by political boundaries (Fabrikant 2001).

Whilst the term fiat is often used to refer to boundaries associated with legal and political processes it also refers to products of social conventions. People often speak of inner boundaries even in the absence of corresponding spatial discontinuities or

intrinsic qualitative differentiation. For example fleets of ships, pairs of shoes, mountains, bays and coves, which require further investigation.

5.2 Humanistic Geography

'Naïve geography', according to Egenhofer and Mark (1995), stands for the 'instinctive and spontaneous' body of knowledge that people have about the surrounding geographic world, central to which are the areas of spatial and temporal reasoning. Naïve geography is also described in Mark, Egenhofer, and Hornsby (1996) as "an untutored set of beliefs about geographic phenomena" (Mark, Egenhofer, and Hornsby 1996, 1). Primary naive beliefs relate to the mesoscopic phenomena in the realm that is immediately accessible to perception and action: beliefs about tables and boats, tabletops and snow, neighbourhoods and streets (Smith and Mark 2001). As such the primary theory is that of 'common sense' which we find in all cultures and in all human beings at all stages of development. "Secondary theories are those that relate to *folk* or pre-scientific beliefs at a different economic scale and social setting" (Smith and Mark 2001, 597). Where as 'folk' refers to the conceptualization of landscape for any particular cultural/language group.

Similarly, humanistic geography seeks to understand the cognitive processes that underpin the human behaviour behind spatial cognitive decisions. As such humanistic geography expands the cognition of the environment from an individual perspective to a social one, in the form of language as a social construct. The philosophy of environmental determinism assumes that people struggle for survival in a narrow context of environmental controls, emphasizing the distinctively human themes of meaning, value, goals, and purposes. This approach also has helped develop ties between geography and psychology, placing greater emphasis on the role of the individual as someone who shaped, as well as responded to, the conditions of the physical and social environment, stressing the links between perception, decision making, and behaviour. Smith and Mark (2001) contend that if places have a special ontology, landscape terms themselves make individual commitments to specific landscape entities. This suggests a preconceived notion regarding the nature of landscape objects as defined by language terms. Research by Burenhult and

Levinson (2008) asks how landscape features are selected as nameable objects and poses the question, as to whether such universal categories exist in the cognition of landscape features. Burenhult and Levinson (2008) suggest that contrastive cognitive styles are reflected in language, and these play a causal role in inducing community wide consensus for one style over another. On the other hand, Smith and Mark (2001) contend that features such as ‘mountain’, ‘cliff’ and ‘river’ presume the existence of such things that one might reasonably consider universal concepts. Investigation of these arguments is central to the study of geographic cognition as such how different views or perceptions are influenced by cultural factors can enable a better understanding of how geographic information can be accessed between different groups.

5.2.1 Ethnophysiography

Research by (Mark and Turk 2003a; Mark and Turk 2004) take a regional domain view of the study of the cognitive process of geographic categorisation via the study of geographic cognition via language across several ethnic groups known as ethnophysiography. The term is derived from ethno being *folk* categorizations and physiography being physical geography and examines the similarities and differences in conceptualizations of landscape held by different language and/or cultural groups. Ethnophysiography is related to the study of ‘place’ and ‘place attachment’, and examines how these significances are tied into the traditional beliefs, often embedded in creationist stories.

Work by researchers in the language and cognition group of the Max Planck Institute for Psycholinguistics (MPI) in Nijmegen has extended the range of ethnophysiography case studies and strengthened its linguistic basis, via case studies of landscape terms in ten languages in a variety of geographic locations (Mark, Turk and Stea 2007). According to Mark, Turk and Stea (2007, 2) the data, “point to considerable variation within and across languages in how systems of landscape terms and place names are ontologised.”

Ethnophysiography dimensions include topography, climate, vegetation, lifestyle, beliefs, historical factors and grammar. Ethnophysiography also examines emotional and spiritual bonds to place and landscape, and the role of landscape features in

traditional knowledge systems. The ethnophysiographic hypothesis is that people from different language groups and cultures have different ways of conceptualizing the landscape, evident by the different terminology and ways of talking about and naming landscape features.

Ethnophysiographic study of the Yindjibarndi, an aboriginal Australian group, of the Pilbara region in North Western Australia reveals there are no permanent or even seasonal rivers or creeks in Yindjibarndi country (Mark and Turk 2003a). The country of the Yindjibarndi is characterized by relatively low relief (elevation differences), with rolling hills and extensive flats. In this geographic region larger watercourses have running water only after major precipitation events, usually associated with cyclones (hurricanes). “Between major precipitation events, rivers continue to 'run', however, the water is underground, beneath the (usually sandy) surface” (Mark and Turk 2003a, 4). Further analysis of the region indicates that unlike many areas of inland Australia, there are no significant intermittent or seasonal lakes in the region inhabited by the Yindjibarndi group. Mark and Turk (2003a) describe permanent pools occurring where the lie of the land and the geology cause the water table to break the surface of the ground, and permanent sources of water including permanent pools along the channels of the Fortescue and other larger rivers. In addition there are some permanent small springs, and soaks where water can be obtained by digging. The primary contention of ethnophysiography is that the local factors constituting the ethnophysiographic dimensions is reflected by the language terms for geographic features and ergo the cognition of these features by its inhabitants. The definitions and semantics of Yindjibarndi terms for water-related features from the Mark and Turk (2003a) study indicates several important differences between the Yindjibarndi and English languages regarding the conceptualizations, the ontology, of water in the landscape.

Mark and Turk (2003a) find no distinct terms for different sizes of ‘wundu’ (a stream bed or channel with flat cross profile and low gradient long profile), as there are for streams in English (river, creek, brook, etc.). English has a variety of terms for concave topographic features and the watercourses that often run along them, however, in arid and semiarid regions, the distinction between topographic and

hydrological features is not accommodated. The English language has been extended to refer to geographic features in arid or semiarid areas, words such as ‘river’ or ‘creek’ which are often still used, even if water is seldom present at the surface. “English speakers normally indicate frequency of flow by adding adjectives such as ‘seasonal’, ‘intermittent’, or ‘ephemeral’ to nouns that canonically refer to flowing water features” (Mark and Turk 2003a, 7). The research by Mark and Turk (2003a) indicates that the language of the Yindjibarndi has no specific terms for such features. Possibly as result of the seasonal rainfall of the region and the spiritual significance and importance of water, the permanence of water features is a matter of categorical predication in the Yindjibarndi ontology, that is, otherwise similar entities that differ only in being permanent or temporary apparently are considered to be different kinds of things and are referred to using different terms. In contrast, to the Yindjibarndi “the English language makes permanence a matter of accidental predication and denotes it by using adjectives” (Mark and Turk 2003a, 16). “In English, water features are first classified by flowing or not, and then by size, such that size is often of categorical importance (e.g., lake vs. pond)” (Mark and Turk 2003a 16). In addition Mark and Turk (2003a) find no terms for water-related features in Yindjibarndi that are differentiated only by size, size where specified, is indicated by generic modifiers for large or small.

Similar findings by Burenhult and Levinson (2008) indicate there are no direct equivalents between terms such as ‘mountain’, ‘cliff’ and ‘river’ in the Yeli Dnye language (isolate, Island Melanesia). A dried up watercourse (or wadi) may be thought of as a river but consists only of the bottom and banks which are perhaps not part of a ‘river’ as English speakers think of it. In comparison of the terms *mountain* and the Ye´liˆ Dnye term *mbu,* (Burenhult and Levinson 2008) the latter is unspecified as to size, being applicable to features of varying magnitude (‘mountains’, ‘hills’ and even ‘crab mounds’ on the beach), and only encodes that the feature has a conical shape. In Tzeltal of Mayan, Mesoamerica, ‘witz’ similarly translates as both ‘mountain’ and ‘hill’, as does the Kilivila (Austronesian, Island Melanesia) word ‘koya’. On the other hand, several ‘mountain’ terms in Marquesan (Austronesian, Polynesia) are all described as denoting large-scale convex features and categorizing according to shape or location of the feature. In Seri (isolate,

Mesoamerica) the word *hast com* is defined by its ‘substance’ (stone) and posture (lying). In Lao of Tai, Mainland Southeast Asia, while the word *phuu2*, is translated as ‘mountainous terrain’ which encodes an elevated landmass; it does not refer to a conical or similar unit. The findings of Burenhult and Levinson (2008) report that in a cross linguistic assessment of concave geomorphic features, a category corresponding to English ‘valley’ is not universally present in the sample of regional languages, even though the regions where the languages are spoken contain these features. Kilivila, Lowland Chontal (isolate, Mesoamerica), and Tzeltal are described as having a ‘valley’ term comparable to that of English. The Marquesan term for valley is the same as that for ‘river’ and ‘village’. Lao and Ye´liˆ Dnye lack a ‘valley’ term, the closest equivalent terms meaning things like ‘gradient’ or ‘bottom of inclined plane.’

Furthermore, research by Burenhult and Levinson (2008) show that some languages make no lexical distinction between ‘water’ as a substance and ‘water’ as a landscape feature for example the Jahai language in Mon-Khmer on the Malay Peninsula. Some languages make a distinction based on size, Tzeltal and (to some extent) Lao being cases in point; others less so, like Lowland Chontal and Seri. Ye´liˆ Dnye segments running water into three distinct portions of a drainage system, two of which make up what would normally be translated as ‘river’ and one which falls outside the denotation of the English term (the flow of river water across a lagoon, between the river mouth and the reef opening). ≠Akhoe Hai//om (Khoisan, southwestern Africa) makes no distinction between permanent watercourses and dry riverbeds which only sporadically contain running water.

The research by Burenhult and Levinson (2008) reveals insights into ontological diversity and constraints as to whether categories are driven by natural human intellectual interest or by utilitarian considerations. “’Utilitarianists’ argue that lexical categories reflect practical consequences of knowing certain category distinctions, related to cultural practice and the functional affordances of referents” (Enfield 2008, 2). Whereas according to Enfield (2008) ‘intellectualists’ would argue that lexical categories reflect people’s innate interest in the natural world, combined with the perceptual discontinuities supplied by ‘nature’s plan’” Enfield (2008, 2).

Enfield (2008) is of the utilitarian position with the subtle difference that word meanings are derived primarily by the utility of the word and not the words referent as the word's function is to achieve social and conceptual coordination. Enfield (2008) argues that lexical categorization at large cannot be directly explained by the natural or utilitarian salience of entities, but must be understood through the utility of words in conversation. Research by Enfield (2008) suggests that data from conversation show how people use words to strategically bring certain kinds of ideas into discourse and not (primarily) to 'map words onto the world'. The utilitarian position holds that lexical distinctions will reflect the affordances of referents for the communities who use the languages. The claim is that "communities will linguistically recognize those categories of entity which are of practical importance to members" (Hunn 1982, cited in Enfield 2008, 17).

According to research by (Enfield 2008) many rural Lao speakers practice a wide range of fishing techniques, involving dozens of different forms of traps, lines, nets, and other devices. Each fishing device is appropriate for a different range of aquatic environments depending on its design and manner of the deployment of each device. As such "an accustomed practitioner of these fishing techniques will be able to categorize the aquatic environment based on these practices" (Enfield 2008, 18). The word-utility hypothesis (Enfield 2008) illustrates the cultural rationale for categorizing the landscape. Different water features show differential utility as determined by a suite of affordance properties, including the design of different fishing devices, human access to different aquatic environments, and presence of certain target types of aquatic life. For example using 'giant upright basket trap' (toum thoong) requires deep water at the banks of the largest rivers whilst a fishing with a cast net (he`e`) requires waist-deep water and gently flowing water. By this rational, a range of non-identical physical environments will be defined by patterns of human behavior as alike with respect to their suitability as places where one may fish using a particular fishing technique. Two otherwise rather different looking pieces of geographical reality are therefore functionally equivalent with respect to their suitability for fishing using a certain device. As such geographic categories are formed by Lao fisherman by "grouping disparate features of the landscape as alike for certain purposes and not for others" (Enfield 2008, 20). Enfield (2008) illustrates

the theory of word-utility using the example of a recorded conversation between two people; Speaker A: Where did Miss Oi go? Speaker B: To look at them collect reeds. Results of research by Enfield (2008, 24) indicate that other members of the group consistently replied that Miss Oi must have been going to the swamp (no`o`ng3) as part of the knowledge of the Lao people of the word swamp (no`o`ng3) is that it is a place to collect reeds. Similarly, Enfield (2008) further illustrates word-utility of geographic terms in conveying distance in conversation. Lao Speaker: 'You have to climb up cliffs and mountains to go there.' As such the use of the terms, mountain and cliff in the Lao culture implies that the travelling distance is far.

Enfield (2008) suggests that linguistic categorization is more complex than mere categorization because it involves the public constitution of categories, as such the patterns of word usage constitute a word learners' basis for constructing semantic hypotheses. Enfield (2008) suggests that words can enable or cause our 'disattention' to differences between single instances in the act of communication within a social group. Word usage patterns as such, constitute a word learners' basis for constructing semantic hypotheses, which go on to become the (effectively) fixed and conventional semantic representations which linguists are in the business of describing. For example, Enfield (2008) suggests Lao speakers will have formed a concept of 'sea' not because they have encountered any worldly referent of the word, but because they have encountered the word itself. "There a community's convergence on particular meanings for particular words is not a direct effect of perception or cultural practice, but is a secondary effect of those things, mediated though language use. Whereby perception and cultural practices affect the conventionalization of word meaning only as far as they determine or constrain the things people say and the way they say them" (Enfield, 2008, 27).

5.2.2 Semantic Similarity Measures versus Categorisation

A critical point for ethnophysiography is that natural inorganic domains are not organized by nature into kinds in the same way that biological entities are so organized. Burenhult and Levinson (2008) state that landscape features do not for the most part come presegmented by nature. Mark and Turk (2003c) suggest that the answers to the question as to whether humans view the landscape as continuous space or as discrete objects, are revealed in the context of the ethnophysiographic

dimensions of a region. Unlike higher plants and animals, which to some large degree are grouped into species by nature, landforms more properly belong to continua (Mark and Turk 2003c, 3). In this regard according to Mark and Turk (2003c, 3) water is certainly ontologically distinct from land, but the sizes and shapes of lakes or islands do not naturally fall into discrete categories with absent intermediate cases, in the same way that kinds of trees or birds fall into such groups. Many disciplines and specialist lines of research have been interested in human understandings of landscape, for example archaeology, anthropology, psychology, philosophy and of course cognitive geography. These domains and the understanding of the value of landscape features to these domains directly correlate with how landscape features are selected as nameable objects. How translatable are landscape terms across varying languages and cultures.

5.3 Conclusions

This chapter has examined the nature of geographic categories, which can vary in both physical scale and conceptual scale. Much thought has been invested as to how human beings cognize their environment, and whether scalable categories are a natural result of cognition. Scalable categories such as political boundaries and legal boundaries have an obvious administrative value; however it has been shown that the categorization of natural geographic features varies greatly between social groups. This presents a problem for semantic integration and interoperability because the motivations used to describe features is dependent on highly localised cultural variables evident in the language used by social groups. Facilitating interoperability between groups requires an understanding of these factors. Furthermore due to the diverse nature of geography and social groups it is apparent that correlation between terms may not be either direct or symmetric. As such a paradigm for the comparison between terms is necessary in order to build a knowledge base which reflects and supports our current understanding of the geographic world and the many number of ways it is perceived. In addition the way in which data is stored in the representation universe must also correlate to its perception in the cognitive universe. This would suit the interests of both users and designers.

6 METHODOLOGY

The application of knowledge and knowledge structures to information and data retrieval is part of an inferential approach that Broder (2006) refers to as the third generation of information retrieval models. In the evolution of the commercial search engines this is preceded by the first generation approach which relied solely on, ‘on page’ textual data and second generation approaches which began to use more sophisticated methods such as connectivity methods and mathematical models. In Section 4 it was indicated that information retrieval can benefit from the use of external data sources in the provision of both additional terms to supplement a user query. Corpus based similarity methods such as IDF (Section 4) return additional terms via the occurrences of terms within a document. If a term occurs frequently and in conjunction with another term in that document, they are according to this method deemed to be similar. The disadvantage of this approach is that it relies on the document accurately reflecting everyday word usages. The goal of the third generation approach is to move from the randomness of syntactic matching to semantic matching. The so called inferential approach is characterised by semantic analysis and focuses on the need or meaning behind the query. Knowledge structures (Section 5) such as Ontology build on philosophical theories of how people perceive the world. Structures such as topology, mereology and mereotopology have been used to model the concepts which confront the issue of conceptualising the geographic domain. These may manifest themselves at a representation level or a cognitive one. It is less obvious whether people naturally decompose geographic phenomena into parts and wholes even though terms such as ‘estuary’ infer both a real physical ‘connection’ and implicit cognitive connection between a fresh water river and the ocean. As such the issues of connection and boundary dominate how we perceive geographic phenomena; furthermore the scale of observation has also been shown in Section 5 to be an important factor which shows how these issues are related to the concept of identity. In this vain Section 5 investigated through the research of Ethnophysiography, the origins of words and word usages, the patterns that indicate the greatest degree of commonality between user groups. The proposed methodology is thus a manifestation of several motivations. The first is to integrate the terminology used by different user groups; by moving from the document based approach and its aforementioned shortcomings, by employing an approach which

incorporates more structured, less random, supplementary data. The second is to evaluate how well this structure represents and incorporates the varying conceptualisations of the aforementioned geographic domain.

6.1 Introduction

There are several alternative approaches in determining lexical similarity and ergo context and meaning from words. Words represent concepts in human language but the mapping from words to concepts is many-to-many. That means one concept may be represented with many different words (synonym) and one word may represent many different concepts (polysemy). This chapter discusses the predominant methods to assess semantic similarity which can be broadly grouped into four main approaches; ontology based, dictionary based, corpus based and information theoretic, although some examples may also use a combination of these methods. Determining the quality of a solution is a task which is particularly hard to quantify for these tasks, primarily due to the fact that the outcome is not one which can always be predetermined. Typically a user may have a general idea of what a query is likely to return but by the very nature of the inferential approach the terms returned via any of the aforementioned query expansion methods are more than likely to be unknown and hence there is an element of knowledge discovery involved on the part of the user. Approaches in the determination of semantic similarity or relatedness are usually motivated towards issues of word sense disambiguation (Section 2) or malapropism detection (spelling errors). This research is more concerned with the analysis of each method with respect to the criteria based on its applicability to terms and concepts specific to the geographic domain, and thus quantifying the degree of similarity of terms is also a reflection of how terms are related.

These methods may not have been developed with geographic applicability as the sole criteria but rather within the context of human language as a whole. In a comparison and evaluation of each of these methods and given the range and depth of possible terms used by various groups to describe geographic phenomena the criteria for judgement does not regard the completeness of any one method as necessary if it is even possible. Indeed it is understood and expected that language is a fluid and dynamic medium in which its speakers take liberties, the use of slang and

jargon for example. At the time of this research a body of terms encompassing geographic terms describing phenomena which integrates the terms used across different languages and cultures does not yet exist (to the author's knowledge). To this end, the potential of each method is analysed given the soundness of its conceptual and structural logic, and for the richness of terms that it contains.

In addition, how terms are related by the various similarity measures is explored as an indicator of the granularity of the structure. Thus the ability to add to a knowledge base of terms for a given concept can be seen within a larger context whose maintenance and development would have wide applicability to the GIScience community.

Placing a numerical value on the similarity of terms has the value that it allows a judgment to be made on the quality in terms of relatedness between any two terms, and secondly it allows new terms to be addressed in terms of similarity to existing concepts when there are a fixed number of target terms. In this respect it has been a difficult task for researchers to quantify the relatedness of two terms in any measurable scale. In this area it has been determined that responses from the survey of a human test group provide the best indication of the quality of a measure to place a metric on the relatedness of two terms (Resnik 1999; Nuno 2005).

6.2 Research Questions

For the task of information retrieval it is important to understand how words represent concepts of a geographic domain and how can the words we use to describe geographic space relate to each other to reflect the underlying principals of the concepts they denote (Section 5). In accordance with Objective 1, the many or no answers problem was identified as the primary encumbrance of efficient information retrieval in a database environment and it has been suggested that the use of additional digital sources can be incorporated into the retrieval mechanism to supplement a query. Given the aforementioned shortcomings of corpus based approaches in this regard, the inferential approach is developed by exploring ontological approaches in determining meaning behind the query to achieve Objective 2. In Section 5 the primary beliefs of groups of people when conversing

about geographic phenomena was discussed (Objective 3). This research therefore continues by investigating the implications of this geographic domain specific knowledge in existing non spatial knowledge structures, in order to achieve Objective 4. The application of knowledge structures can be used to expand search results and allow search users to find information of a nature which is conceptually similar but encoded in a different manner by the diversity and complexity of language. In accordance with this aim, more specific revised research questions are added and new questions are posed:

1. How can structures and measures of relatedness and similarity lead to improved precision and recall? (Objective 4)
2. Do the current structures adequately represent the diversity and richness of geographic terms? (Objective 3)
3. Are the primary abstractions present in conversation about geographic features and phenomena discussed in Section 5 represented in the current structures?
4. Is there is a unifying conceptual salience which supports a shared human experience when communicating about geography?
5. What measures are necessary to create, maintain and/or enrich a geographic semantic knowledge structure?

The structural relationships that bind concepts (Section 4) and the origins of words and word referents (Section 5) have been broadly discussed. The following sections examine the ways word groupings can be formed by abstract relationships; these abstract relationships have origins in the word usages and referents in the communities which use them in everyday language. This chapter builds on objectives 3 and 4 and examines the methods by which to study the complex patterns that exist in word groupings and investigates the use of geographic ontology in WordNet, a non spatial knowledge structure. Within the constructs of a lexical structure such as WordNet, there are several measures which can be applied to determine the similarity of words. WordNet as a handcrafted resource, is similar in nature to a thesaurus, has theoretical groundings in ontology, and includes word definitions akin to a digital dictionary. Within this complex environment a measure of lexical

similarity can reveal different levels or granularity between abstract concepts as well as an intermingling, or meshing between an abstract concepts word representation.

6.1.1 Similarity and Relatedness of terms

In the prevailing literature the terms semantic similarity and semantic relatedness are often used, as such it is important to draw a distinction between these terms at an early stage. This distinction is best explained using an example given by Resnik, (1999, 1):*"... cars and gasoline would seem to be more closely related than, say, cars and bicycles, but the latter pair are certainly more similar."*

According to Budanitsky and Hirst (2006) semantic relatedness is a more general concept than similarity; by which similar entities are semantically related by virtue of their similarity (e.g. bank–trust company), but dissimilar entities may also be semantically related by lexical relationships such as meronymy (car–wheel) and antonymy (hot–cold), or just by any kind of functional relationship or frequent association (pencil–paper, penguin–Antarctica, rain–flood).

"The assessment of semantic similarity among objects is a basic requirement for semantic interoperability" (Rodriguez, Egenhofer and Rugg 1999, 1). Most speakers are aware that in the use of language two words may often be used interchangeably to communicate the same idea (synonymy). However other relationships such as meronymy and hypernymy can also be used to structure language to describe concepts based on many a number of salient properties. Where the synonymy of terms would infer equality in terms of similarity assessment accounting for these more complex relationships requires further stretching of the imagination. Rodriguez and Egenhofer (2004) defined geospatial semantic similarity to support the identification of objects that are conceptually close, but not identical and as such highlight the importance of similarity assessments in retrieval of geospatial data in settings such as digital libraries, heterogeneous databases and the WWW.

6.1.2 A Semantic Neighbourhood

A term which is often used in discussions regarding semantic similarity and relatedness is *semantic neighbourhood*. Broadly speaking it is a grouping of terms

which are of particular importance in the description of a concept and may include the use of several relationships such as meronymy, hyponymy etc. Schwering (2004) defined the semantic neighbourhood of a concept *A* as all concepts C_i which have a smaller distance to concept *A* than the radius of the neighbourhood and the concept *A* itself (Figure 6.1). According to Schwering (2004) if two concepts have the same relation to the same concept in a neighbourhood, they can be considered as similar. The semantic neighbourhood of a geographic object suggests that 'lakes' and 'ponds' are both categories at some basic level or with the same conceptual 'neighbourhood', which would distinguish the two objects by size. According to Schwering (2004) in order to set up these neighbourhoods, some basic connections like a common vocabulary between ontology's must already explicitly exist. The distance between two concepts is measured along the shortest path in the net of unidirectional arrows that represent the hierarchical relations. The size of the neighbourhood can be defined by its radius. Thus, if the radius equals 1 (Figure 6.1), only the immediate neighbours, *surface water body*, *riverbed* and *river mouth* are in the neighbourhood of the concept *river*.

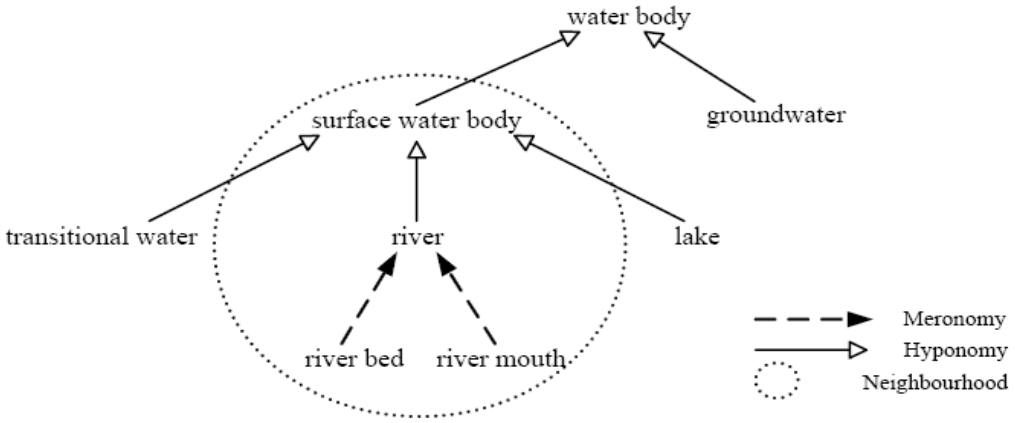


Figure 6.1: Semantic Neighbourhood "River" with Radius=1 (Schwering 2004)

Is-a, part-of, and next-to are relationships, which Schwering (2004) suggests must all be considered in a measure of similarity assessment. In this study the *Next-to* relationship is unique as it captures relationships between geographic features which is dependent on heuristic inference. For example a river flows into the sea. Similarly,

according to ecological data sources, floodplains are periodically flooded areas next to rivers.

6.1.3 Semantic Granularity

Fonseca et al. (2002b, 2) defined semantic granularity as “addressing the different levels of specification of an entity in the real world”, as opposed to spatial granularity, which deals with the different levels of spatial resolution or representation at different scales. Spatial granularity, with reference to spatial databases, implies that granularity is related to the level of distinction between the elements of a phenomenon that is represented by a dataset. In GIScience this notion refers to the variation of representation of geographic objects across varying scales. “For instance, if an urban settlement is perceived at a small scale, the level of detail is usually small enough for an entire city with all its complex internal structure and boundaries to be represented as a point or as a simple polygon on a map. If the same city is perceived at a larger scale it becomes necessary to represent its internal structure with more detail, for instance depicting blocks, square, major streets and buildings” (Fonseca et al. 2002b, 16).

Semantic granularity on the other hand can be seen as a hierarchy of nested terms resulting from a decomposition of like terms within a semantic neighbourhood. The operations for changes in the level of semantic detail are generalisation and specialisation. In the process of generalisation of a class with a certain level of detail, a new class with less detail is generated. Specialisation operates in a manner inverse to generalisation by converting a more general class to a more specific one. Fonseca et al. (2002b) illustrated the benefits to be gained exploiting the generalisation operation, to improve the number of results in geographic query. That is by using a more general class of a given concept, for example ‘body of water,’ will retrieve more objects than by using a single term at a lower ontological level such as ‘lake’ or ‘pond’. The more general term would include ‘lake’ and ‘pond’ as retrieved results. Semantic granularity is represented in language by the relationships of hypernymy and hyponymy.

6.1.4 Semantic Relationships

Hyponymy is transitive and asymmetrical (Lyons 1977), and since there is normally a single superordinate, it generates a hierarchical semantic structure, in which a hyponym is said to be below its superordinate. “Such hierarchical representations are widely used in the construction of information retrieval systems, where they are called inheritance systems” (Touretzky 1986, cited in Miller 1990, 8): a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate. For example, *maple* inherits the features of its superordinate, *tree*, but is distinguished from other trees by the hardness of its wood, the shape of its leaves, the use of its sap for syrup, etc. This convention provides the central organizing principle for the nouns in WordNet.

6.2 Ontology Based Approaches (WordNet)

Common to most ontological approaches is the use of a structure such as a taxonomy (Section 4.7) which relates concepts to one another via the transitive Hypernym/Hyponym IS-A relation. In WordNet (refer Section 1.5) English nouns, verbs, adjectives and adverbs are organized into synonym sets (also called synsets), each representing one underlying lexical concept. Different relations such as hyponymy, hypernymy, meronymy link the synonym sets. Nouns are the most common word type in WordNet. Some nouns refer to classes, by which the membership in those classes determines the semantic relation of hyponymy. Other nouns however refer to particular individuals also known as instances. There are 141,691 nouns in WordNet and 7,671 synsets tagged as instances. WordNet was not initially conceived as an ontology but as a description of lexical knowledge and therefore the ontological distinction between classes and instances has been made explicit. The distinction between classes and instances as of Version 2.1 of WordNet Miller and Hristea (2006) suggests allows it to be treated via the WordNet nouns as a semi-ontology by ignoring all nouns tagged as instances.

For example ‘women are numerous’ does not imply that any particular woman is numerous, but rather that the class of women has numerous instances. The noun woman denotes a class, whilst a proper noun, for example Margaret Thatcher denotes

an instance of that class. Similarly, the noun river denotes a class, whilst a proper noun such as the 'Avon River' denotes an instance of the class of river. The distinction between nouns and instances or concepts and individual is not always clear in WordNet, but has become so in later versions as WordNet has moved to convert the WordNet hierarchy to a more ontological structure. Version 2.1 incorporates the so-called 'unique beginners' or top-level elements such as *entity*. There are characteristics which all words denoting instances share. Firstly, they are all nouns and denoted as such in the WordNet synset field, 'ss_type' with the value *n*. Second, they are proper nouns and as such should be capitalised. Finally, the referent is a unique entity that implies it should not have hyponyms. Place names, which are considered instances in WordNet, include geographic regions that do not have well-defined political boundaries, as well as cities and states, islands and continents, rivers and lakes, mountain peaks and mountain ranges, seas and oceans, planets and satellites, stars and constellations.

In WordNet nouns are organized in lexical memory as *topical hierarchies*. Each of these lexical structures reflects a different way of categorizing experience; attempts to impose a single organizing principle on all syntactic categories would, according to Miller (1990, 3) badly misrepresent the psychological complexity of lexical knowledge. In the words of Miller (1990, 3) the disadvantage of this organization comes with the realization that syntactic categories differ in subjective organization, which emerged from the first studies of word associations. Fillenbaum and Jones (1965), for example, asked English-speaking subjects to give the first word they thought of in response to highly familiar words drawn from different syntactic categories. The price of imposing this syntactic categorization on WordNet is a certain amount of redundancy that conventional dictionaries avoid—words like 'back', for example, turn up in more than one category. "The advantage is that the fundamental differences in the semantic organization of these syntactic categories can be clearly seen and systematically exploited" (Miller 1990, 3).

Lexical semantics begins with the recognition that a word is a conventional association between a lexicalized concept and an utterance that plays a syntactic role. According to Miller (1990) the aforementioned definition of 'word' raises at least

three classes of problems for research. First, what kinds of utterances enter into these lexical associations? Second, what is the nature and organization of the lexicalized concepts that words can express? Third, what syntactic roles do different words play? According to one definition “two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made” (Leibniz cited in Miller 1990, 6). By this definition, true synonyms are rare, if they exist at all. For a definition of synonymy in terms of truth values, thus making synonymy a discrete matter, thus two words are either are synonyms or they are not. According to research by Miller (1990) popular thought is that synonymy is best thought of as one end of a continuum along which similarity of meaning can be graded. In addition Miller (1990, 7) states that “theories of lexical semantics does not depend on truth functional conceptions of meaning; semantic similarity is sufficient and according to such measures of similarity, semantically similar words can be interchanged in more contexts than can semantically dissimilar words”.

6.3 Measures of Similarity and Relatedness in and Ontology

6.3.1 Path Length or Edge Counting

A simple way to compute semantic relatedness in a taxonomy such as WordNet is to view it as a graph and identify relatedness with path length between the concepts: “The shorter the path from one node to another, the more similar they are” (Resnik 1995a, 3; Budanitsky and Hirst 2006, 5). According to this definition an ontology containing the concepts and links (Figure 6.2), it can be said that dog and cat have a semantic distance of 4 links. In the prevalent literature the concept of ‘path length’ is differentiated from the method of ‘edge counting’ which is identical with the exception that it relies entirely on hypernymy and hyponymy (IS-A) relationships whereas the ‘path length’ approach is inclusive of the relations of hypernymy, hyponymy, holonymy, and meronomy. The edge based approach is thus a measure of similarity as opposed to relatedness (path length approach). Thus from the example (Figure 6.2) *Cat* is more similar to *Feline* than it is to *Carnivore* or *Canine*. “In the case, in which similarity is limited to the use of hypernym/hyponym links, the opposite of this measure corresponds to a measure of similarity” Nuno (2005, 54). Accordingly the similarity between *Cat* and *Dog* is therefore 0.25(1/4).

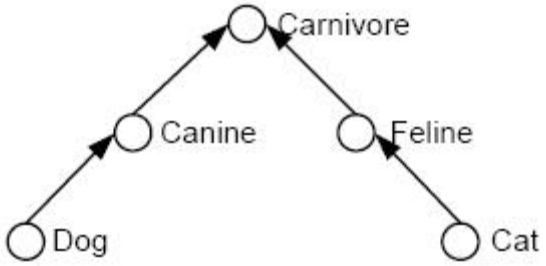


Figure 6.2: The relevant portion of an ontology which can be used to extract the semantic distance (Nuno 2005)

According to Lee, Kim, and Lee (1993 cited in Budanitsky and Hirst 2006, 5) "in the context of semantic networks, shortest path lengths between two concepts are not sufficient to represent conceptual distance between those concepts when the paths are restricted to IS-A links, the shortest path length does measure conceptual distance." Despite its apparent simplicity, a widely acknowledged problem with the edge counting approach is that it typically "relies on the notion that links in the taxonomy represent uniform distances" (Resnik 1995a, 1), which is typically not true: "there is a wide variability in the 'distance' covered by a single taxonomic link, particularly when certain sub taxonomies (e.g., biological categories) are much denser than others" (Resnik 1995a, 1). For instance, in WordNet, the link *rabbit ears* IS-A *television antenna* covers an intuitively narrow distance, whereas *white elephant* IS-A p*ossession* covers an intuitively wide one. In this regard Jiang and Conrath (1997, 5) note that "in a more realistic scenario, the distances between any two adjacent nodes are not necessarily equal." It is therefore necessary to consider that the edge connecting the two nodes should be weighted. The approaches discussed below are attempts undertaken by various researchers to overcome this problem by scaling the taxonomy (Budanitsky and Hirst 2006).

6.3.2 Depth Relative Scaling

Sussna's Depth-relative Scaling (Scriver 2006) is an approach to scaling based on the observation that sibling-concepts deep in a taxonomy appear to be more closely related to one another than those higher up. The scaling factor reflects the

observation that the deeper we are in the taxonomy the more subtle the distinctions between concepts become, consequently branching in deeper parts of the ontology should not represent semantic leaps as large as the branching performed in higher (more abstract) parts. The method was originally conceived for the task of word sense disambiguation, an the result of the research shows improvements where multiple sense words can be disambiguated by finding the combination of senses from a set of contiguous terms which minimizes total pair wise distance between senses.

Each relation *r* has a weight or a range [min_r; max_r] of weights associated with it: for example, *hypernymy*, *hyponymy*, *holonymy*, and *meronymy* have weights between min_r = 1 and max_r = 2.4 The weight of each edge of type *r* from some node c_1 is reduced by a factor that depends on the number of edges, $edges_r$, of the same type leaving c_1:

$$wt(c_1 \rightarrow_r) = max_r - \frac{max_r - min_r}{edges_r(c_1)} \tag{6.1}$$

The distance between two adjacent nodes c_1 and c_2 is then the average of the weights on each direction of the edge, scaled by the depth of the nodes:

$$dist_S(c_1, c_2) = \frac{wt(c_1 \rightarrow_r) + wt(c_2 \rightarrow_{r'})}{2 \times \max\{depth(c_1), depth(c_2)\}} \tag{6.2}$$

where *r* is the relation that holds between c_1 and c_2 and *r'* is its inverse (i.e., the relation that holds between c_2 and c_1), which Sussna assumes is symmetrical. The semantic distance between two arbitrary nodes c_i and c_j is thus the sum of the distances between the pairs of adjacent nodes along the shortest path connecting them (Scriver 2006; Budanitsky and Hirst 2006).

6.3.3 Leacock and Chodorow

In their measure of semantic similarity, Leacock and Chodorow (Scriver 2006) calculate the similarity between two concepts by finding the length of the shortest path that connects them in the WordNet taxonomy. The length of the path found is scaled to a value between zero and one and the similarity between the terms is then

calculated as the negative logarithm of this value. The similarity (simLC) of two terms c_1 and c_2 is thus calculated by:

$$\text{simLC}(c_1, c_2) = -\log \frac{\text{len}(c_1, c_2)}{2 \times \max_{c \in \text{WordNet}} \text{depth}(c)} \tag{6.3}$$

Where length is the length of the shortest path between the two synsets (using node-counting) and D is the maximum depth of the taxonomy (Budanitsky and Hirst 2006).

6.3.4 Resnik's Information Based Approach

Information based approaches use a combination of corpus statistics and conceptual taxonomy to calculate the similarity between two terms. Given a multidimensional space upon which a node represents a unique concept consisting of a certain amount of information, and an edge represents a direct association between two concepts, the similarity between two concepts is the extent to which they share information in common (Scriver 2006). In a hierarchical concept space such as WordNet, this common information 'carrier' is identified as a specific concept node that subsumes both of the two terms in the hierarchy. This so-called 'super-class' is thus the first class upward in this hierarchy that subsumes both classes, the information content value of this super-ordinate class as such defines the similarity between the two terms. The value of the information content of a class is then obtained by estimating the probability of occurrence of this class in a large text corpus. Resnik (1995a) used the noun frequencies gathered from the one-million-word Brown Corpus of American English to generate these values. The key characteristic of his counting method is that an individual occurrence of any noun in the corpus "was counted as an occurrence of each taxonomic class containing it" (Budanitsky and Hirst 2006, 21). For example, an occurrence of the noun *nickel* was counted towards the frequency of nickel, as indeed was its hyponym coin, and so forth. The probability of encountering an instance of the concept is given by:

$$p(c) = \frac{\sum w \in W(c) count(w)}{N} \quad (6.4)$$

where $W(c)$ is the set of words (nouns) in the corpus whose senses are subsumed by concept c, and N is the total number of word (noun) tokens in the corpus that are also present in WordNet (Budanitsky and Hirst 2006).

The degree of similarity between two concepts c_1 and c_2 is:

$$sim_R(c_1, c_2) = \max_{c \in S(c_1, c_2)} -\log p(c) \quad (6.5)$$

Where $S(c_1,c_2)$ is the set of concepts that subsume both c_1 and c_2. This method is characterized by the way in which the number of edges separating terms is downplayed, and does not factor in the calculation of similarity. In this method, as the node's probability increases, its information content or its informativeness decreases. If there is a unique top node in the hierarchy, then its probability is 1, hence its information content is 0. Therefore the higher one travels up the hierarchy to find the most common subsumer of two terms then the less similar those terms will be.

6.3.5 Jiang and Conrath's Combined Approach

Jiang and Conrath (1997) use a combined model that is derived from the edge-based notion by adding the information content as a decision factor. Jiang and Conrath (1997) argue that the strength of a child link is proportional to the conditional probability of encountering an instance of the child concept c_i given an instance of its parent concept p: $P(c_i \mid p)$. By definition the quantity $P(c_i \mid p)$ is:

$$P(c_i \mid p) = \frac{P(c_i \cap p)}{P(p)} \quad (6.6)$$

Like Resnik (1995a) the Jiang and Conrath (1997) method counts every instance of a child as an instance of its parent. Therefore the equation can be reduced to:

$$P(c_i \mid p) = \frac{P(c)}{P(p)} \tag{6.7}$$

$$\text{distJC}(c_i \mid p) = \text{IC}(c_i) - \text{IC}(p) \tag{6.8}$$

$$\text{distJC}(c_i \mid p) = \text{IC}(c_i) - \text{IC}(p) \tag{6.9}$$

$$\text{distJC}(c_1 \mid c_2) = \text{IC}(c_1) + \text{IC}(c_2) - 2 \times \text{IC}(\text{lso}(c_1, c_2)) \tag{6.10}$$

$$\text{distJC}(c_1 \mid c_2) = 2\log P(\text{lso}(c_1, c_2)) - (\log \times P(c_1) + \log P(c_2)) \tag{6.11}$$

The Jiang and Conrath (1997, 8) method uses the sum of the individual distances between the nodes in the shortest path. As the lowest super-ordinate (lso) shared by c_1 and c_2 does not have a parent in the path, this node is excluded from the summation (Scriver 2006).

6.3.6 Mapping between Languages

Wu and Palmer (1994) reveal the importance of determining meaning from terms in a study involving the translation of verbs from English to Mandarin. Wu and Palmer (1994, 136) map verbs from English and Mandarin verb compounds into concepts from which a hierarchy of terms can be created and thus used to compare the similarity of terms between languages. Wu and Palmer (1994, 134) note that the nature of the lexical selection task in translation obeys Zipf's law in that, for all possible verb usages, a large portion is translated into a few target verbs, while a small portion might be translated into many different target verbs. Fine grained translations between terms is achieved by the creation of concepts in which verbs from both languages can be mapped and the decomposability of terms within these concepts can then be used to define a scaled metric for what they call conceptual similarity between a pair of concepts c_1 and c_2 in a hierarchy. This strategy of *projecting* verbs from both languages onto a common conceptual structure, from which a similarity measure is defined and based on this conceptual representation, allows a target lexical item to be put in correspondence with a source item that most closely carries the same meaning. In Wordnet the Wu and Palmer (1994) method

calculates relatedness by considering the depths of the two synsets in the WordNet taxonomy, along with the depth of the Least Common Subsumer (LCS). The method proposes that the similarity (simWP) between a pair of concepts c_1 and c_2 can be formulated as:

$$\text{simWP}(c_1, c_2) = \frac{2 \times \text{depth}(\text{lso}(c_1, c_2))}{\text{len}(c_1, \text{lso}(c_1, c_2)) + \text{len}(c_2, \text{lso}(c_1, c_2)) + 2 \times \text{depth}(\text{lso}(c_1, c_2))} \quad (8.2)$$

The depth ($lso(c_1, c_2)$) is the 'global' depth in the hierarchy whose role is that of a scaling factor. The same equation recast from similarity to calculate distance is:

$$\text{distWP}(c_1, c_2) = 1 . \text{simWP}(c_1, c_2) \quad (8.3)$$

$$\text{distWP}(c_1, c_2) = \frac{\text{len}(c1, \text{lso}(c1, c2)) + \text{len}(c2, \text{lso}(c1, c2))}{\text{len}(c_1, \text{lso}(c_1, c_2)) + \text{len}(c_2, \text{lso}(c_1, c_2)) + 2 \times \text{depth}(\text{lso}(c_1, c_2))} \quad (8.4)$$

This formula returns a value between 0 and 1. The score can never be zero because the depth of the LCS is never zero (the depth of the root of a taxonomy is one). The score is one if the two input synsets are the same. Based on this similarity measure a correct lexical choice can be achieved, even when there is no exact lexical match from the source language to the target language.

6.3.7 Dictionary Based Approaches

Dictionary based approaches use a machine readable dictionary to discover relations between concepts. Similarity can be achieved via comparison of word definitions. For example, in a dictionary (Kozima and Ito 1997) the terms *car* and *bus* are associated by the term *vehicle* that occurs in each term's definition, which could then be placed in the set: taxi, railway and airplane. As WordNet contains both relations in the form of synsets and definitions in the form of glosses it can be used as both an ontology and a dictionary.

Dictionary based approaches to semantic similarity calculate the overlap of the content words in the definition of two words. Or use distributional similarity as an approximation of semantic distance between the words in two glosses. For example the definitions of chicken and turkey both include the terms food and meat:

- chicken - the *meat* from this bird eaten as *food*
- turkey - the *meat* from a turkey eaten as *food*

Dictionary approaches use word definitions to create a node for every headword and link this node to the nodes corresponding to all the headwords in its definition, in this manner *turkey* and *chicken* would both be linked to *food* and *meat.*

In the extended gloss overlap measure Banerjee and Pedersen (2003) use a comparison of the definitions for each WordNet synset rather than an examination of the paths created by the synsets. According to Banerjee and Pedersen (2003) while WordNet provides *explicit* semantic relations between synsets, via *is–a* or *has–part* links, such links do not cover all possible relations between synsets. For example, WordNet encodes no direct link between the synsets *car* and *tire*, although they are clearly related. Banerjee and Pedersen (2003) observe that the glosses of these two synsets have words in common, and as such the extended gloss overlap method attempts to overcome these shortcomings in WordNet. Relatedness of terms is computed by means of the overlap of the words in the definitions of the headwords (terms being compared) and related concepts in the WordNet hierarchy. Banerjee and Pedersen (2003) suggest that 'phrasal overlaps' - sequences of words that appear in different glosses are often indicative of a strong relationship between concepts. As such Banerjee and Pedersen (2003) calculate relatedness by counting the number of words that co-occur in the glosses of different synsets and assign a higher value to a phrasal overlap of *n* words than to an overlap of *n* words that are not in sequence. Specifically, a phrasal overlap of *n* words is assigned the value n^2, whereas *n* shared words that do not belong to a phrasal overlap are assigned the value *n*. For example, in WordNet the glosses (definitions) for drawing paper and decal are as follows:

- *drawing paper* – 'paper that is specially prepared for use in drafting'
- *decal* – 'the art of transferring designs from specially prepared paper to a wood or glass or metal surface.'

As the phrase ‘specially prepared’ appears in both glosses, it contributes a score of 2x2 = 4. The word ‘paper’ also appears in both glosses, and contributes a score of one, for a total score of five (Scriver 2006).

Banerjee and Pedersen (2003) extend this function to include the glosses of not only the target synsets, but also of their nearest neighbours in the semantic network. For each relation type *r*, they define a function *r* (s_1) that returns the gloss of the synset related by *r* to s_1. For example, the function hypernym(s_1) returns the gloss of the hypernym of the synset s_1. If s_1 is connected to more than one synset by the relation type *r*, then *r* (s_1) returns the concatenation of the glosses of each related synset. In addition, Banerjee and Pedersen (2003, 3) also define a function named gloss (s_1) that returns the gloss for the synset s_1. Given the non-empty set of RELPAIRS = {(gloss,gloss), (hype,hype), (hypo,hypo), (hype,gloss), (gloss,hype)}, the relatedness function of two terms is calculated by:

relBP (s_1, s_2) = score(gloss(s_1), gloss(s_2))+score(hype(s_1), hype(s_2)) + score(hypo(s_1) hypo(s_2)) + score(hype(s_1), gloss(s_2)) + score(gloss(s_1), gloss(s_2))

Where score (t_1, t_2) is a function that returns the overlap score of two strings t_1 and t_2.

6.3.8 Functional Relationships from Word Definitions

Kavouras, Kokla and Tomai (2005, 147) have made some inroads in the area of affordance based reasoning for, via the extraction of verb patterns in WordNet glosses. For example the pattern for the extraction of the semantic relation *purpose* is given, “if the verb used within the definition is post modified by a prepositional phrase with the preposition *for*, then a *purpose* relation is created with the head(s) of that prepositional phrase as the value” (Kavouras, Kokla and Tomai 2003, 147). For example, a *purpose* relation is extracted from the definition of the word, *canal*. From WordNet is defined as ‘a manmade or improved natural waterway used for transportation.’ The algorithm searches the definition of the word for the verb *for* and automatically assigns a conceptual link based on the word followed by it. Therefore *canal* can be assigned the concept (purpose) of transportation. This method of extracting geographic information is used by Kavouras, Kokla and Tomai (2005,

148) to analyse the definitions of geographic categories into a set of semantic relations with their corresponding values. This formalised semantic information is further used to explicitly disambiguate categories by explicitly and objectively identifying similarities and heterogeneities between them. More specifically, if the methodology for extracting semantic information is used in the analysis of the category *lake*, which from WordNet is defined as, 'a body of water surrounded by land', then three semantic relations can be extracted:

1. Hypernym with value 'body'
2. Material with value 'water'
3. SURROUNDED BY with value 'land'

According to Kavouras and Tomai (2004, 189) "expression of spatial reference and geographic concepts are common to all natural languages, in that humans are spatially referenced 'objects' and are forced to interact with space in every aspect of their existence." Subsequently, there is a wealth of linguistic expressions about space in natural languages, which can be exploited. As such, natural languages have deployed figurative expressions when describing spatial objects, such as metaphors with great expressive power. Kavouras and Tomai, (2004) use an analysis of tourist guides from the WWW as an example of how semantics reveal how textual descriptions encode geospatial semantics, which can be used to for data integration, and which could potential be used to supplement the WordNet knowledge base. Kavouras and Tomai (2004) categorised geospatial semantics into descriptions of location and descriptions of motion. Descriptions of location, include linguistic terms such as the category of prepositions, which accompany the *locative* verb, such as *stand, over, at, dangles, or southern.* Descriptions of motion, describe location in space using a 'vector' exemplified by the *source* and *goal* of the motion, which include phrases such as *runs through, continues westward, along, or continues,* etc (Kavouras and Tomai 2004). Kavouras and Tomai (2004) report in the findings of this research that the linguistic elements tend to differ widely over natural language cultures and suggest that further exhaustive analysis will give a clear account of how humans reason about space and express it through language.

Kuhn (2002) creates finer-grained conceptual spaces, by using functionality as the primary structuring relationship between terms. Kuhn (2002) emphasizes the semantic role of function (affordances provided by an instance of the category) in mappings among conceptual spaces in order to achieve total (rather than partial) semantic mappings of similarity. The functional relations are captured in WordNet (such as the sheltering function of a house) most prominently in its glosses (definitions), but, they are treated outside the formal apparatus of hypernymy and hyponymy relations. Further examination by Kuhn (2002) of the concept hierarchies of the two concepts 'houseboat' and 'boathouse' reveal that the basic *affordances* (shelter and transportation) derived from the glosses of the concepts occur at a low level in the taxonomy. As such glosses at lower levels tend to use functional descriptions ('used to store boats', 'for travel on water'), while those higher up in the hierarchies are not only more abstract, but also formulated independently of function. Kuhn (2002) introduces *image schemata* to compensate for the lack of functional description higher in the taxonomy where the concepts are too abstract and less plausible. These image schemata include *containment*, *surface*, *path* and *contact*, in a similar way in which Wu and Palmer (1994) map verbs to domains such as *change of state, action, motion and contact.* Kuhn (2002) relaxes the restrictions of the taxonomy created from this functional mapping, by allowing for a concept to inherit from multiple parents where functions are the units of inheritance. As such the function of a boathouse to shelter boats, or of a houseboat to shelter humans, can be described as the result of different morphisms (structure-preserving mappings) from the theories of houses and boats to those of boathouses or houseboats.

6.4 Evaluation of Current Research

WordNet is constructed based on a partitioning of nouns via a set of semantic primes, or unique beginners' (Figure 6.3). The features that characterize a unique beginner are inherited by all of its hyponyms. Each unique beginner constitutes the root node of a separate hierarchy; as such these multiple hierarchies correspond to relatively distinct semantic fields, each with its own vocabulary.

{*act, action, activity*}	{*natural object*}
{*animal, fauna*}	{*natural phenomenon*}
{*artifact*}	{*person, human being*}

{*attribute, property*}	{*plant, flora*}
{*body, corpus*}	{*possession*}
{*cognition, knowledge*}	{*process*}
{*communication*}	{*quantity, amount*}
{*feeling, emotion*}	{*shape*}
{*food*}	{*state, condition*}
{*group, collection*}	{*substance*}
{*location, place*}	{*time*}
{*motive*}	

Figure 6.3: List of the 25 unique beginners' for WordNet nouns (Miller 1990)

The organization of syntactic categories, is by the authors own admission (Miller 1990) subjective, and as such WordNet is by no means a complete reference and depending on the purpose or point of view, the relationships between concepts is one which is also open to debate and analysis. Some unique beginners' for example {plant, flora} are more developed than others, in addition Banerjee and Perdersen (2003) mention that there are several links such as *car* and *tire* which are omitted in WordNet. As a consequence dictionary based methods which utilize the Wordnet word definitions (glosses) such as the extended word overlap (Banerjee and Perdersen 2003) can be useful in order to compensate or supplement WordNet relationships by plugging gaps in the taxonomy. Kavouras, Kokla and Tomai (2005) show via the analysis of word definitions that the word 'lake' has three potential attributes for abstraction, 'body', 'water' and 'surrounded by land.' Understanding how a word can have several abstractions is the key to interoperability, as some word abstractions may be more important to one group of people than another. In a retrieval scenario a user based design requires synchronisation of user vocabulary and the word meaning. The relationships between words in a corpus based approach are often the result of the purpose or subject of the document. The ontological approaches discussed represent a manually alignment of terms, whilst these links can be used for query expansion it is necessary to examine this alignment.

As described in Section 5 different communities infer different properties and usages from geographic features and as such it is necessary to examine how and what abstractions are encoded in WordNet. This concept is developed in research by

Kavouras and Tomai (2004) who suggest an organisation of geographic concepts into the upper level concepts of location and motion, by identification of spatial descriptors, which may vary amongst cultural groups. Kavouras, Kokla and Tomai (2005) build on this in later work via the extraction of affordance or functional based relationships via the extraction of terms following verbs in the definitions of spatial terms. If the encoding of words are based on affordances and culturally distinct factors as suggested by Section 5 then abstractions based on affordances, is a many to many relationship (Figure 6.4). As such the abstraction of a word based on this affordance abstraction will also be many to many and can therefore words and word groupings based on abstraction relationships can be used for query expansion.

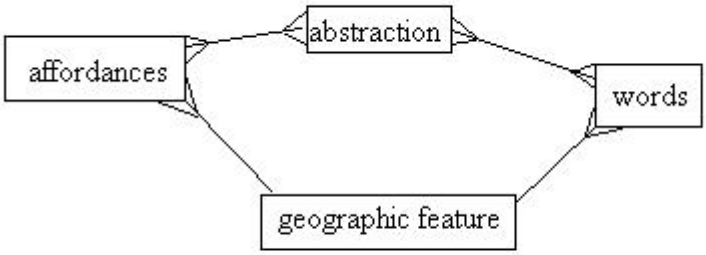


Figure 6.4: Proposed alignment of geographic concepts based on word usages.

In order to determine the nature of the relationship between words and word referents of geographic features and how their abstraction relationships are encoded in WordNet, the similarity measures of Resnick (1995a), Leacock and Chodorow (Scriver 2006) and Jiang and Conrath (1997) are used. A retrieval implementation using the WordNet relationships is a natural evolution to document based retrieval methods, however the salient concepts in 'geographic words' and their relevance to the affordance abstractions and ethnophysiography research (Section 5) must be identified. To reveal what mechanisms relate 'geographic words,' the aforementioned similarity measures are applied and the WordNet structure for geographic terms is mapped.

The use of this methodology may be applied to create new taxonomies and links which previously did not exist. Such a methodology continues in the vain of research

of Wu and Palmer (1994) in which concepts were successfully projected from taxonomies to achieve integration between English and Mandarin language words.

6.5 Conclusions

This chapter has reviewed the most prevalent methods in contemporary literature for determining the relatedness (all relationships) or similarity (using only IS-A relationships) of terms. Some of the methods examined use an ontology such as WordNet with a predetermined structure of 'handcrafted' relationships between terms known as edges via relationships such as hypernymy, hyponymy or meronymy, to determine similarity. This represents a more structured approach to the approach of corpus based similarity measures and presents a source of terms which has the potential to be used for query expansion in information retrieval. Analysis of the ontological approach of WordNet requires explicit representation and identification of the relationships between relevant terms, therefore the generalisation/specialisation relationships between terms should be mapped to reveal the mechanism behind word groupings.

When using the so called path based measures which utilize the WordNet structure via its synsets it is acknowledged that "the shortest path lengths between two concepts are not sufficient to represent conceptual distance between those concepts, however when the paths are restricted to IS-A links (edge counting), the shortest path length does measure conceptual distance" (Lee, Kim and Lee 1993 cited in Budanitsky and Hirst 2006, 17). Words in the hierarchy become more similar the deeper in the hierarchy they are located. As such similarity measures generally attempt to impose weights on these relationships either as a function of the depth of the taxonomy or via additional information such as word occurrences in supplementary data such as the Brown Corpus of American English. These weights have the utility of quantifying how much one term varies from another as such can provide a useful gauge for the possible substitution of terms for a user. Further analysis of word groupings and word relationships using the similarity measures outlined is proposed.

7 IMPLEMENTATION AND EVALUATION

This chapter outlines the processes and algorithms by which the WordNet database is used and how its respective fields and tables are used to form an ontology, this ontology is exploited to derive additional meaningful terms for query expansion. It is expected that the utilization of this structure will lead to improved search recall and precision over the more prevalent corpus based methods. This research proposes that geographic phenomena entail unique determinants in word usage as such the word clusters of geographic terms in the WordNet structure are mapped in order to clearly expose these determinants. Analysis of the effectiveness of the retrieval implementation is based on using the similarity methods discussed in Section 6 to reveal the factors contributing to word groups rather than via direct comparison with corpus based search methods. The relationships determined from these mappings are evaluated based on the ethnographic research and affordances discussed in Chapters 5 and 6.

7.1 Overview of Implementation and Evaluation

Concepts denoted by words in the databases are related via the synset relationship to form a hierarchy, which is used to manage and organize knowledge in a network structure. This structure is then used to manage and organise the relevance of terms from gazetteer test datasets from which the metric of semantic relatedness is determined to quantify the similarity of the retrieved term and the query term. The similarity metric employed is based on the node distance measure which is based on the number of edges between terms in the taxonomy. This method however does not describe the variation between terms based on their depth in the tree structure. The node based measure is supplemented using the most prominent similarity measures obtained using the software of Pedersen and Michelizzi (2009). The secondary objective of the implementation is to expose and analyze the WordNet taxonomy for geographic concepts based on the research of ethnophysiography (Sections 5) and in accordance with the research questions posed in Section 6 particularly where directly relevant to the retrieval process. To recap, these are to determine how the retrieval process may be improved via the abstraction of salient features of geographic concepts which support a shared human experience when communicating about

geography. These are largely present in the more abstract terms (hypernyms) which subsume two concepts. Determining what higher level abstractions are present in taxonomy of geographic concepts in WordNet is the basis for identifying if the unifying concepts present are in accordance with the ethnophysiographic concepts illustrated in Section 5.

7.1 Data

Two types of data are required for the implementation, the WordNet database to provide terms for query expansion, and a test data set to test the effectiveness of the implementation. Two test datasets were used in testing; the gazetteer of Australia database for the Perth area and the gazetteer for British Columbia, Canada.

7.1.1 WordNet

The WordNet database is available in a number of formats, online and downloadable in raw formats for conversion for use in conjunction with a number of programming language environments. In this case the *Prolog* to *Mysql* conversion of WordNet provided by Android Technologies (2007) was used. The main tables are outlined below (Figure 7.1).

Database name:-> wn_pro_mysql
Show tables;

wn_antonym
wn_attr_adj_noun
wn_cause
wn_class_member
wn_derived
wn_entails
wn_gloss
wn_hypernym
wn_hyponym
wn_mbr_meronym
wn_part_meronym
wn_participle

wn_pertainym
wn_see_also
wn_similar
wn_subst_meronym
wn_synset
wn_verb_frame
wn_verb_group

Figure 7.1: WordNet 2.1 tables.

Of the tables present in WordNet (Figure 7.1) only three are required for implementation; these tables and their fields are displayed in Figure 7.2.

TABLE: WN_SYNSET
synset_id (decimal)
w_num (char)
word (char)
ss_type (char)
sense_number (int)
tag_count (int)

TABLE: WN_HYPONYM
synset_id_1 (decimal)
synset_id_2 (decimal)

TABLE: WN_HYPERNYM
synset_id_1 (decimal)
synset_id_2 (decimal)

Figure 7.2: WordNet database table relations used in implementation.

The creation of the network structure is based on understanding these three tables, which use the unique table keys as pointers between nouns, and hypernyms or hyponyms. The table WN_SYNSET includes the fields such as word which is of a

character type form such as ‘rock’ or ‘mountain’. For this implementation, the purpose is to restrict the data to terms of a geographic nature. This is achieved by the field ‘ss_type’ which denotes the type of word e.g. noun (n), verb (v), adverb (a) etc. For this implementation only nouns are used.

7.1.2 Retrieval Data

Two source geographic datasets were used to test the retrieval mechanism. The first obtained from Geoscience Australia (The Australian Government 2008) and is based on a local subset of the gazetteer of Australia for the Perth, Western Australia region (GEODATA_TOPO250K_TILE_DATA). The second dataset, the gazetteer for British Columbia, Canada (British Columbia Geographical Names 2008) was obtained from the government of British Columbia, Canada. These data consist of geographic features encoded with a unique name, when available, and geographic location using Geodetic Datum of Australia (GDA94), and WGS84 coordinates respectively. Each geographic feature is grouped according to a predefined ‘feature type’ (Appendix A). The features included for this region were created from several ESRI shape files, instances of polygon feature types were reduced to point features by using the centroid of the original polygon and exported via comma separated values file to MySQL database. Therefore all features in the database can be referenced in a uniform manner via coordinates in latitude and longitude. The gazetteer for British Columbia, Canada (Appendix B) contained significantly more features 186 feature types compared to the 20 feature types used in the Perth, gazetteer. The structure of the resulting MYSQL database table ‘feature’ consists of the fields:

1. Unique id,
2. Feature Type,
3. Feature Name,
4. Latitude WGS84,
5. Longitude WGS84.

The unique id is the primary key or index denoting a unique record in the database. The feature type can include any of the 20 or 180 features mentioned above, which may or may not have a name. For example Perth international airport has a feature type: ‘airport,’ and a feature name: ‘Perth international airport’, with a location in

easting and northing. There can be many airports in the database but there is only one airport named “Perth international airport” with a unique id and unique easting and northing. “Feature type” therefore serves as a class and each table is denoted by its fields: unique id, feature name, easting and northing that are attributes which denote an instance of that class. Each dataset was used independently of the other.

7.2 Detailed Design

As a proof-of-concept the use of a lexical ontology of geographic terms given by WordNet is created firstly so that a user’s query can be expanded to retrieve more results than would normally be possible in a traditional Boolean search environment (refer Section 3.2.4). This structure serves a second function, which is to order the results based on the relevance to the original query term via semantic relatedness.

7.2.1 Design Considerations

The design of the architecture is therefore centred on the facilitation of the visualisation, of the aforementioned concepts of query expansion and relevance ranking. The interface is composed of the three query terms: feature type, relation and feature instance. Due to the presence of instances in the source or search data set, the specification of a query can be taken further to assume that a user is searching for a specific instance of a class. As previously mentioned an instance is denoted by a unique position in easting and northing and possibly a name. As the source data has geographic coordinates the results of a search can be displayed graphically giving the position of the instance on a map.

For graphical display a base point (feature instance), and a relation (relation) of proximity are necessary to control the geographic scale of the search. Therefore a typical search would consist of:

Lakes (feature type) near (relation) Perth (feature instance).

The term relation consists of a ‘drop down list’ where the user is given a choice between near, north of, east of, south of etc. This query specification paradigm has been adopted in a number of document based spatial retrieval applications such as

SPIRIT (Spatial Information Retrieval on the Internet) (Jones, Alani and Tudhope 2003).

7.2.2 Software Platform/Hardware Architecture

All datasets were converted from their native formats to MySQL format, of which dynamic access via the Internet environment is facilitated using server side language PHP (Hypertext Preprocessor). AJAX or Asynchronous JavaScript and XML (Extensible Mark-up Language) were developed in 2005. The AJAX XMLHttpRequest object allows JavaScript functions to call database objects via PHP without the necessity of forms, which can be cumbersome and time consuming.

As discussed in Section 3.2.1 the role of the interface is two fold: query formulation and results display. A query from the user in HTML is sent to the PHP server (Figure 7.3) which interrogates the MySQL database to determine if the feature and feature instance entered by the user exists in the gazetteer. If matching data is found the results are displayed using JavaScript and AJAX in Google Maps.

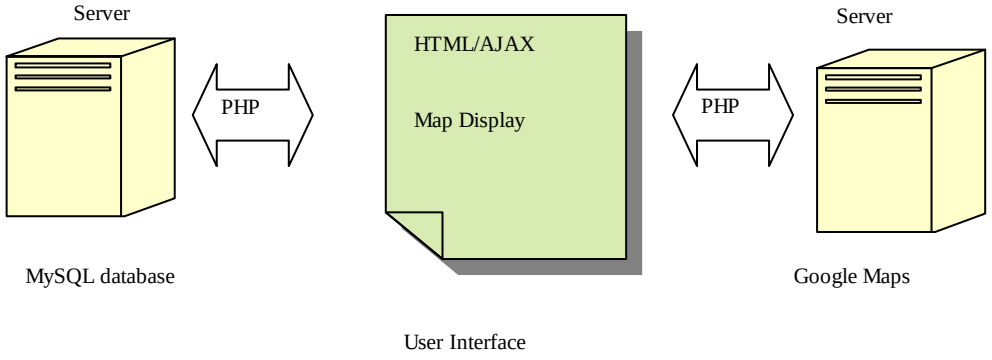


Figure 7.3: Web Implementation Architecture

7.3 Search Algorithm

The WordNet database stores relationships between words via the presence of the unique id's of two words in a table (i.e. hypernym, or hyponym) denoting the form of the relationship between the pair. If the id's of two words are present as a pair in a table record then a relationship between the two words exists as described by the name of the table. The following Section describes how the relationships between

words in the WordNet database are dynamically formed at run-time to create a lexical structure to which ontological semantic similarity for query expansion and indexing can be applied.

7.3.1 Context Diagram

The diagram below (Figure 7.4) shows how the two data sources, WordNet and the test gazetteer/s are linked and interrogated via the user query. A query is specified using the three terms of feature type, feature instance and relation (which is set to *near* by default). The feature instance is used to specify a base position from which the query is stipulated, and the relation is used to spatially constrain the results returned. Possible values for the relation field include north-of, south-of, west-of, east-of. These values do not have any bearing on the analysis on the results obtained which are analysed from an entirely conceptual perspective, as such only the default option *near* is used in the analysis. The feature type is not used to directly query the gazetteer database; rather a match is first sought from the WordNet table WN_SYNSET (1), if a match is found then the synset_id of the word is used to search the WN_HYPERNYM and WN_HYPONYM tables for matches (2). This is a recursive process which is performed until no further matches are found (3). The results of this process is an array of synsets which must then be matched to the WN_SYNSET table in order to retrieve the corresponding word terms. The resultant array of words is used to find matches with the field feature type in the gazetteer (5).

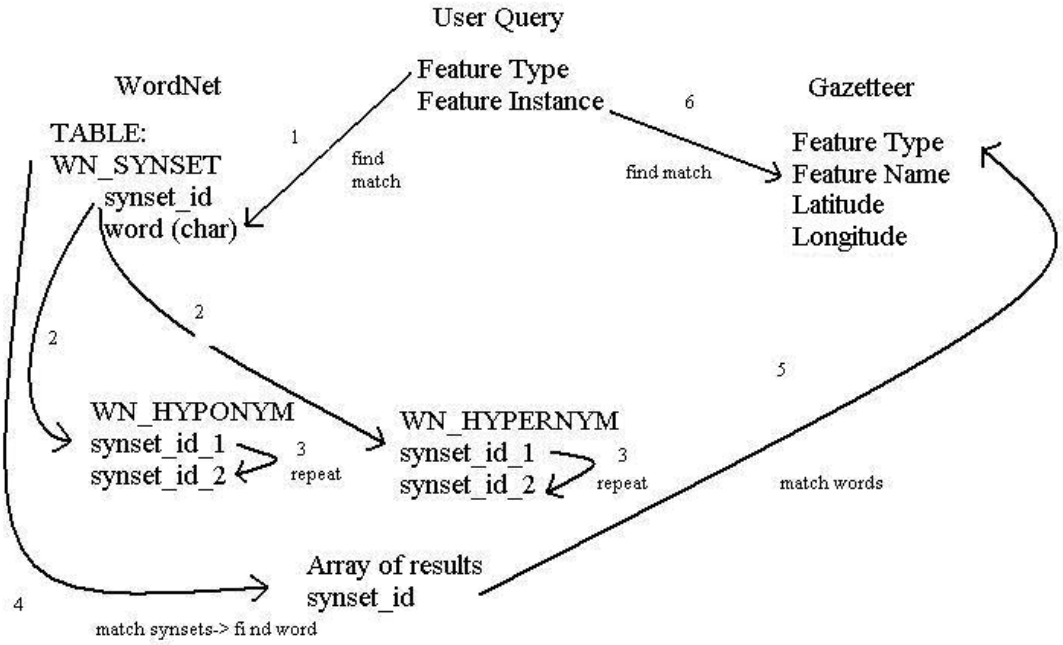


Figure 7.4: Context Diagram

The feature instance from the user query, and feature name from the gazetteer are then matched (6) to determine a base point from which the results of (5) can be compared depending on the choice of relation field from the user query (set by default to *near*) as it is the ranking of semantic distance which is of the most importance.

From Figure 7.4 step (1) when a search word (feature type) has been entered in the user query, the WordNet synset table is searched for all matching words (Figure 7.5).

```
$find =
"select s2.synset_id, s2.word from (wn_synset AS s1, wn_synset AS s2)
where s1.synset_id=s2.synset_id and s1.word LIKE '".$searchWord."'
and s1.ss_type='n'";
$query = mysql_query($find);
```

Figure 7.5: Feature type word match between query and WordNet

The result of this query (Figure 7.5) is the unique key identifying each synonym set. The synset key is then used to find matches in the hyponym and hypernym datasets Figure 7.4 step (2). The WN_HYPERNYM and WN_HYPONYM tables (Figure 7.2) are composed of two-fields, with each field a pointer to a word representing the either the hypernym or hyponym relationship between the two words.

7.3.2 Extraction of Relationships

A typical search (Figure 7.4) matching a query to the WordNet database (Figure 7.4 step (1)) would involve matching the WN_SYNSET table in order to obtain the synset_id for that word. Figure 7.6 indicates a search of the term knoll to the WN_SYNSET table to return the synset_id 108744286.

```
mysql> select * from wn_synset where word like 'knoll';
+-----------+-------+-------+---------+--------------+-----------+
| synset_id | w_num | word  | ss_type | sense_number | tag_count |
+-----------+-------+-------+---------+--------------+-----------+
| 108744286 |     1 | knoll | n       |            1 |         1 |
+-----------+-------+-------+---------+--------------+-----------+
1 row in set (0.30 sec)
```

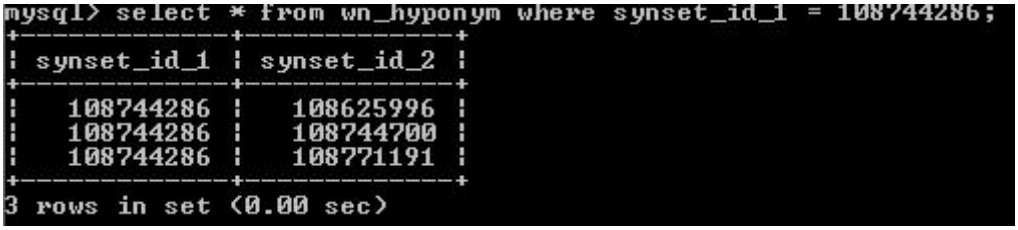
Figure 7.6: Results of a search for the word ‘knoll’ in the WordNet table WN_SYNSET.

The synset_id’s from this search can then be used to determine their corresponding words from the WN_SYNSET table Figure 7.4 step (4). Figure 7.7 indicates the result of this process to include the synonyms of ‘knoll’ which include mound, hillock, hummock, and hammock.

```
mysql> select * from wn_synset where synset_id = '108744286';
+-----------+-------+---------+---------+--------------+-----------+
| synset_id | w_num | word    | ss_type | sense_number | tag_count |
+-----------+-------+---------+---------+--------------+-----------+
| 108744286 |     1 | knoll   | n       |            1 |         1 |
| 108744286 |     2 | mound   | n       |            2 |         1 |
| 108744286 |     3 | hillock | n       |            1 |         2 |
| 108744286 |     4 | hummock | n       |            1 |         0 |
| 108744286 |     5 | hammock | n       |            1 |         0 |
+-----------+-------+---------+---------+--------------+-----------+
5 rows in set (0.00 sec)
```

Figure 7.7: Search results for word match ‘spur’ in WordNet table WN_SYNSET.

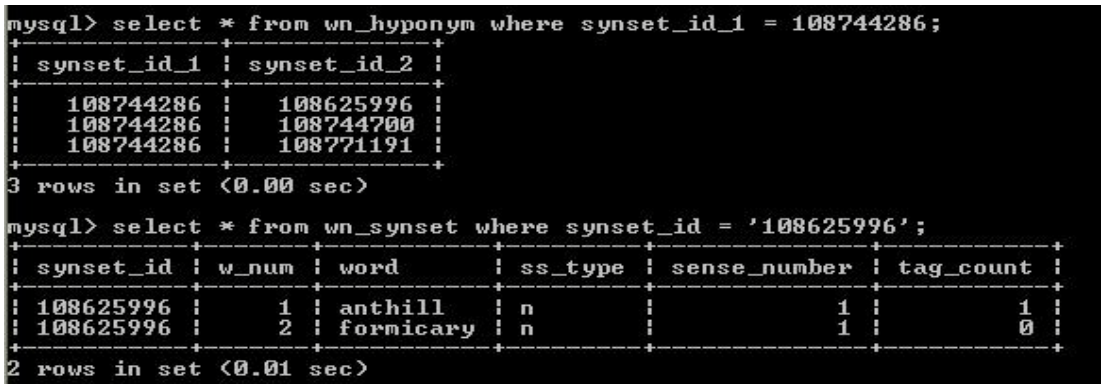
Following a successful match from Figure 7.4 step (1), the WN_HYPONYM and WN_HYPERNYM tables are then searched (Figure 7.4 step (2)) via the retrieved synset_id from (Figure 7.4 step (1)) the result is a list of hyponyms of the initial query term ‘knoll’ (Figure 7.8).



```
mysql> select * from wn_hyponym where synset_id_1 = 108744286;
+-------------+-------------+
| synset_id_1 | synset_id_2 |
+-------------+-------------+
|   108744286 |   108625996 |
|   108744286 |   108744700 |
|   108744286 |   108771191 |
+-------------+-------------+
3 rows in set (0.00 sec)
```

Figure 7.8: Results of a search knoll with synset_id ‘108744286’ on WordNet table WN_HYPONYM.

The results of matching the synset_id of the hyponym (Figure 7.8) to find its corresponding word and synonym (Figure 7.9) reveals that the hyponyms of knoll are anthill and formicary.



```
mysql> select * from wn_hyponym where synset_id_1 = 108744286;
+-------------+-------------+
| synset_id_1 | synset_id_2 |
+-------------+-------------+
|   108744286 |   108625996 |
|   108744286 |   108744700 |
|   108744286 |   108771191 |
+-------------+-------------+
3 rows in set (0.00 sec)

mysql> select * from wn_synset where synset_id = '108625996';
+-----------+-------+-----------+---------+---------------+-----------+
| synset_id | w_num | word      | ss_type | sense_number | tag_count |
+-----------+-------+-----------+---------+---------------+-----------+
| 108625996 |     1 | anthill   | n       |             1 |         1 |
| 108625996 |     2 | formicary | n       |             1 |         0 |
+-----------+-------+-----------+---------+---------------+-----------+
2 rows in set (0.01 sec)
```

Figure 7.9: Search results for word match ‘knoll’ synset_id ‘108635996’ in WordNet table WN_HYPONYM result in the hyponyms anthill and formicary.

This process can be carried out for any word and potentially be repeated until the hypernym or hyponym tree ends, by searching the hypernym or hyponym table with the synset_id (synset_id_2) of the previous search. In order to retrieve results to the full extent of the lexical hierarchy each result returned is searched using the same

function in a recursive manner until the end of the ‘chain’ has been reached (Figure 7.10). This can be applied to either the WN_HYPERNYM table or WN_HYPONYM table by substitution of the variable $table (Figure 7.10 line 1) and is achieved by ‘calling’ the function from within its definition (Figure 7.10 line 11).

```
1.  function search(&$tree, $table, $word){
2.  //RECURSIVE SEARCH:-for each matching synset retrieve its super or sub ordinate
3.  //term, store in an array and repeat procedure on resulting word.
4.  $query =
5.  "select t.synset_id_2, s.word from wn_synset AS s LEFT JOIN $table t ON (s.synset_id
    = t.synset_id_1) where s.synset_id='".$word."' and s.ss_type='n'";
6.  $result = mysql_query($query) or die (mysql_error());
7.  while($row = mysql_fetch_array($result)){
8.   //checks the array to see if it has not already been recalled
9.   if(in_array($row[1], $tree)==false){$tree[] = $row[1]; }
10.  //calls the function from itself so as too retrieve the entire chain of synsets
11.  search($tree, $table, $row[0]);
12.  }//end while
13.}
```

Figure 7.10: Recursive traversal of WordNet WN_HYPERNYM and WN_HYPONYM tables via PHP MySQL search algorithm.

7.3.3 Calculating Semantic Similarity

A ranking is assigned to each returned word, to denote the semantic space it occupies, using the node counting scheme Equation (7.1). This is a simple method to that of the edge counting method employed by Rada (et. al 1989, cited in Nuno 2005, 53) which assumes that the number of links between two concepts in the ontology reflects the semantic distance between them, however includes the nodes of the search term and the returned term as part of the measure.

$$\text{dist}(c_1, c_2) = \frac{1}{\text{edges}(c_1, c_2) + 1} \tag{7.1}$$

The maximum value is 1 which indicates that the two sysnets are in fact the same. As the edge distance between two terms increases the similarity decreases and the

dist(c_1, c_2) tends toward zero. For example in Figure 7.12 the semantic distance between cat and dog using the edge counting scheme is 4. Using the node based approach the distance is the inverse of the number of nodes which is 0.2 (1/5) between and including cat and dog.

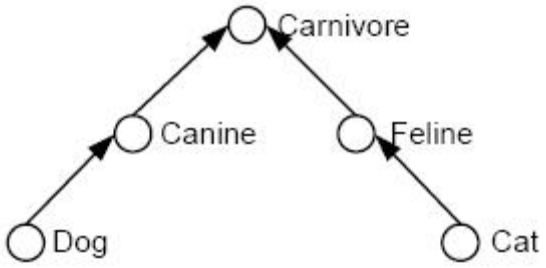


Figure 7.11: The relevant portion of an ontology which can be used to extract the semantic distance (Nuno 2005)

In Figure 7.12 each search result returned is placed in an array with an associated counter value reflecting the semantic distance from the original search term. The inverse of the counter (1/(counter+1)) is then returned as the measure of semantic relatedness.

```
$counter=1;
foreach ($results as $col) {
                foreach ($synsetArray as $row) {
                        if($col == $row){$counter=1;}
                }
                        $searchlist[] = array('word'=> $col, 'rank'=> $counter) ;
                        $counter++;
}
        return $searchlist;
```

Figure 7.12: Implementation of semantic distance metric algorithm via counts of synset connections.

7.4 Evaluation

Figure 7.13 displays in green text the user query ‘wetlands *near* Perth’ in the form of ‘feature type’, ‘relation’ and ‘feature instance’, where as previously stated relation remains as the constant default value of ‘near’. The reference location or feature instance (Figure 7.4 step (6)) (in this instance Perth) is represented as the blue icon marker. This is used to give the search a context (refer Section 7.2.1). The closest matching result (red marker) represents the most semantically similar result which is spatially nearest to the feature instance (blue marker). Subsequent results (ranked 2, 3, and 4) are displayed using the yellow markers.

The results of the search of feature wetlands are the feature *instances* located on the left hand side (Figure 7.13) listed in order of semantic distance from the query, and then spatial distance. The results ranking function is an application of the PHP *multisort array* function (refer Appendix) which orders the two dimensional array with values semantic distance and geographic distance, for each of the search results firstly by semantic distance calculated in Figure 7.12 and the geographic distance calculated using the distance as the base point (blue marker). The results of a query are displayed to the left of the screen (Figure 7.13). The closest match is a ‘swamp’ whose name is N/A (unavailable). The second line indicates the coordinates in latitude and longitude of WGS84 followed by the distance from the reference point (Perth), which is 6139.54 m. In addition to edge distance dist(c_1,c_2) the equivalent similarity measure using the methods of Leacock and Chodorow (Scriver 2006), Resnik(1995a) and Jiang and Conrath (1997) have been summarized using the WordNet Similarity software (Pedersen and Michelizzi 2009) in Table 7.1.

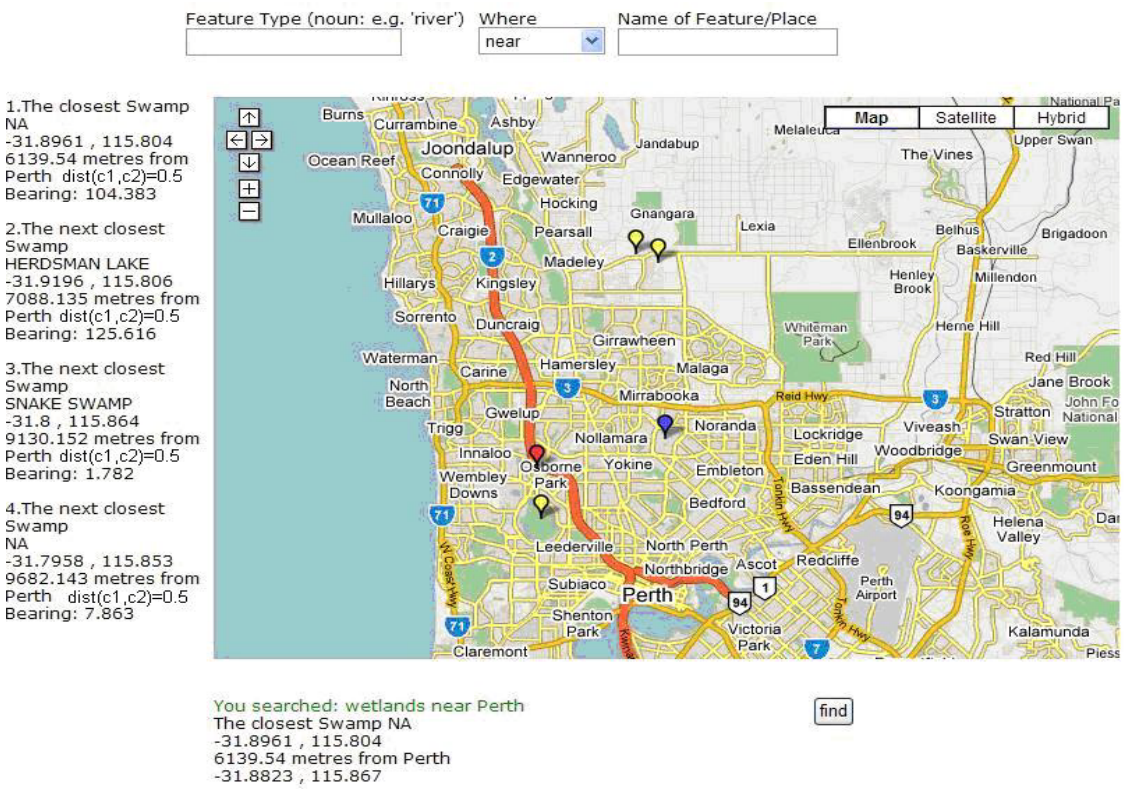


Figure 7.13: Screenshot, results indicating search of ‘wetlands near Perth’.

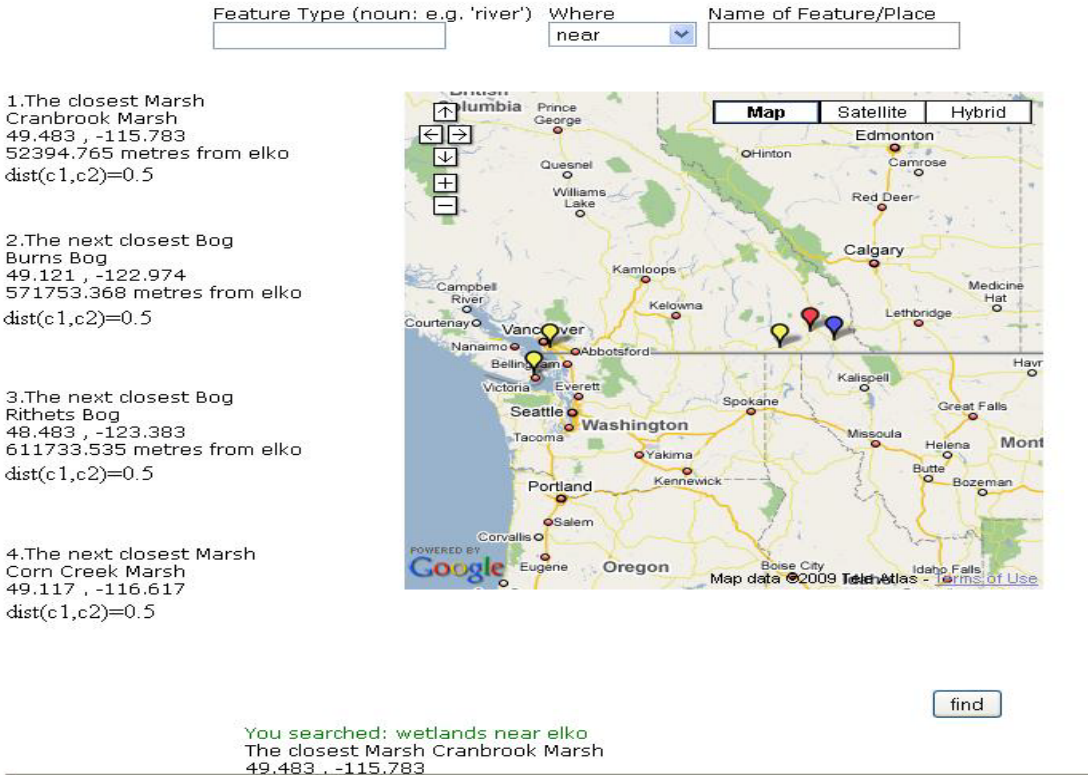


Figure 7.14: Screenshot, results indicating search of ‘wetlands near Elko’.

Figure 7.13 indicates the closest matching results of the query ‘wetlands.’ ‘Swamp’, *low land that is seasonally flooded*; *has more woody plants than a marsh and better drainage than a bog* is a hyponym of ‘wetland’ and represents a generalisation of the concept with an edge distance of 1 and a node based similarity value of 0.5. Figure 7.14 applies the same query using the gazetteer dataset for British Columbia, ‘wetlands near Elko’ which returns the resulting feature types ‘marsh’ and ‘bog’ which are both 1 edge from the term ‘wetlands’ giving a node based distance of 0.5, indicates that the abstraction ‘wetland’ can be successfully used to retrieve a grouping of conceptually similar terms. Figure 7.15 (below) represents the mapping of this conceptual hierarchy the terms in the WordNet taxonomy.

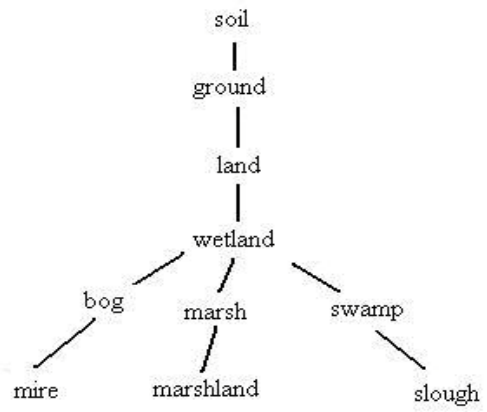


Figure 7.15: The semantic neighbourhood of the geographic term ‘wetland’ in WordNet.

The entire neighbourhood of terms in Figure 7.14 was used to search the database; the best matching results are the immediate hyponyms of wetland, the terms ‘mire’, ‘marshland’ and ‘slough’ also fall within the super concept of wetland but are also sub-concepts of ‘bog’, ‘marsh’ and ‘swamp’ respectively, and are to a degree members of the super concept ‘wetland’. The degree of this similarity, given by this mapping of terms is presented below in Table 7.1.

Search term	Retrieved term	Edge	Node based distance	Leacock & Chodorow	Resnik	Jiang & Conrath
Wetland	land	1	0.5	2.9957	8.1822	0.6106
Mire	wetland	2	0.3333	2.5903	9.8198	0.5139
Mire	marshland	4	0.25	2.3026	9.8198	0.2569
Wetland	bog	1	0.5	2.9957	9.8198	0.7982
Swamp	slough	1	0.5	2.9957	10.3795	0
Mire	bog	1	0.5	2.9957	11.0726	1.4427
Bog	swamp	2	0.3333	2.5903	9.8198	0.5518

Table 7.1: A comparison of similarity measures for geographic features for the concept of ‘wetland’.

From Table 7.1 (above) it is apparent that the measure used by Resnik (1995a) shows a dramatic decrease in similarity once the closest common subsumer or superclass ‘wetland’ has been determined. There is no common subsuming concept linking the terms ‘land’ and ‘wetland’ and this is reflected by a lower measure of similarity (8.1822) as compared to the remaining terms which all share the same superclass ‘wetland’. The pair of ‘mire’ and ‘bog’ has the highest similarity value using the Resnik (1995a) measure as they share a minimum number of links and a common subsuming concept. The information theoretic method of Resnik (1995a) supplements the taxonomy with term occurrences from the Brown corpus. As such each term occurrences of a concepts hyponym is counted as an occurrence of its hyponym in the corpus. As the term ‘mire’ has three hyponyms (not shown in Figure 7.15) ‘quagmire’, ‘quag’ and ‘morass’ whilst the term ‘slough’ is the lowest node in the tree with no hyponyms the pair ‘mire’ and ‘bog’ have a subsequently higher similarity than the pair ‘swamp’ and ‘slough’.

Figure 7.16 displays the results of the query ‘valleys *near* Elko’ using the gazetteer of British Columbia dataset. The resulting best matching feature is ‘gorge’ with a node based similarity value of 0.3. Mapping of the semantic neighbourhood of the

term ‘valley’ (Figure 7.17) in the WordNet taxonomy, reveals a larger structure of greater detail than the previous example of ‘wetland’.

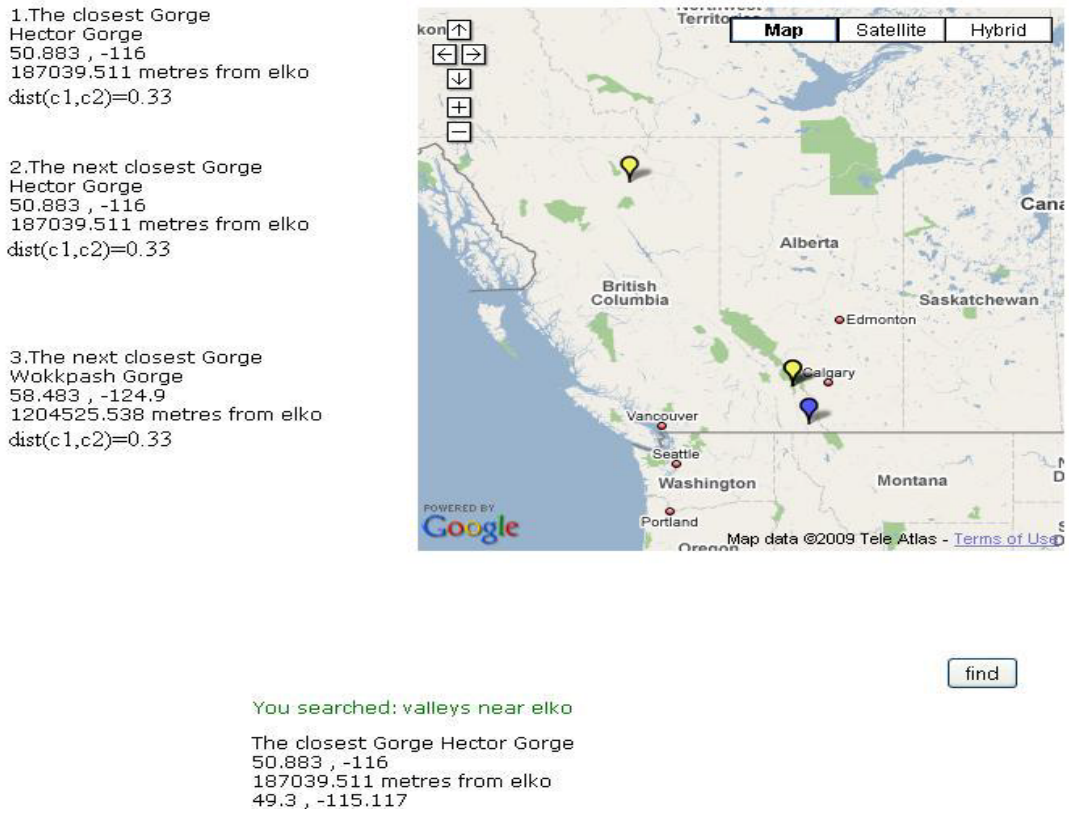


Figure 7.16: Screenshot, results indicating search of ‘valleys near Elko’.

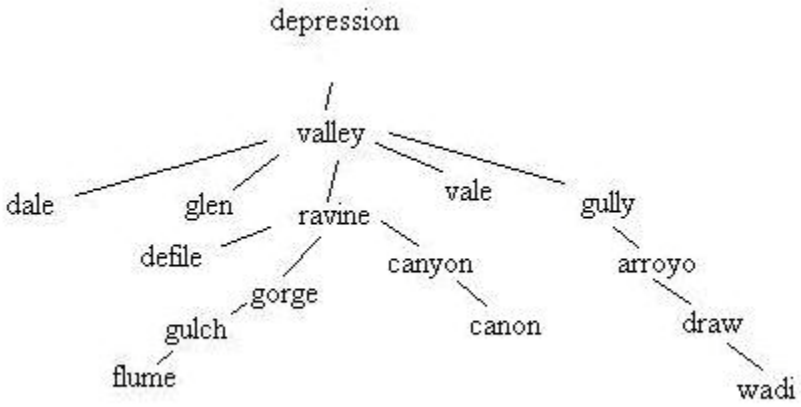


Figure 7.17: A subset of the semantic neighbourhood of the geographic term ‘valley’ in WordNet.

Figure 7.17 indicates that the term valley subsumes a wide range of terms such as ‘ravine’, ‘glen’, ‘gully’, ‘dale’ at a distance of 1 edge and ‘arroyo,’ ‘draw’ and ‘wadi’

at edge distance of 2, 3 and 4 respectively. Therefore, via the generalisation term ‘valley’ *a long depression in the surface of the land that usually contains a river,* it is possible to retrieve any one of the terms from ‘dale’ to ‘flume’ which are sub concepts of ‘valley’, which in turn is a sub concept of the term ‘depression’.

Search term	Retrieved Term	Edge	Node based distance	Leacock & Chodorow	Resnik	Jiang & Conrath
valley	gorge	2	0.3333	2.5903	8.0281	0
ravine	gorge	2	0.5	2.9957	10.1563	0
valley	ravine	4	0.5	2.9957	8.0281	0.4699
depression	valley	1	0.5	2.9957	7.4753	1.809
depression	wadi	5	0.25	2.3026	7.4753	0
arroyo	wadi	2	0.3333	2.5903	9.5685	0
arroyo	brook	11	0.0909	1.291	0.6144	0.0529

Table 7.2: A comparison of similarity measures for geographic features for the concept of ‘depression’.

Analysis of the similarity results of the Leacock and Chodorow measure (Scriver 2006) in Table 7.2 reveals a trend in the comparison of the results from Table 7.1. Whilst the measure is sensitive to changes in edge distance, the scaling method (based on the depth of the taxonomy) is not as sensitive as compared with the Resnik (1995a) method to changes in depth. This confirms the hypotheses that abstract terms higher in the taxonomy have less information content and are thus less similar than more specific terms lower in the taxonomy. Analysis of the word definitions for this mapping (Table 7.3) reveals that the relationship between terms is explicitly defined in the gloss for that term.

Term	Definition/Gloss
valley	a long depression in the surface of the land that usually contains a river
ravine	a deep narrow steep-sided valley (especially one formed by running water)
canyon	a ravine formed by a river in an area with little rainfall
dale	an open river valley (in a hilly area)
glen	a narrow secluded valley (in the mountains)
gully	deep ditch cut by running water (especially after a prolonged downpour)
arroyo	a narrow channel in the ground that is usually dry but becomes a stream after heavy rain
draw	The dry bed of a stream
wadi	gully or streambed in northern Africa and the Middle East that remains dry except during rainy season

Table 7.3: Analysis of the word definitions for the semantic neighbourhood of 'valley.'

It is evident from Table 7.3 that WordNet's taxonomy is designed not to capture common causality, function and behaviour, but rather to show how existing lexemes relate to each other. For example, consider the abstraction that would unite {brook, arroyo}. From WordNet the glosses for brook: 'a natural stream of water smaller than a river (and often a tributary of a river) and arroyo: 'a stream or brook.' These terms have a Leacock and Chodorow similarity of 1.291, a Resnik similarity of 0.6144 and a Jiang and Conrath similarity of 0.0529. Considering that the definition of 'arroyo' implies synonymy with the concept of 'brook' this is not reflected by any of the measures which reveal a greater similarity between 'arroyo' and 'wadi' (Table 7.2). Rather, the terms 'valley', 'gully', 'arroyo' 'draw' and 'wadi' have been linked by the causality of being formations created by heavy rainfall with higher subsuming abstract concepts of 'valley' and 'depression'. Whilst this measure of similarity may be perfectly valid for the perceptions of several groups it also illustrates that the relationships are subjective and may favour one concept over another in this case the concept of 'formation' is favoured over 'water'. In addition the relationship of terms is not explicitly defined in the definition but only in the relation of a term to another. Furthermore, it illustrates the concept of 'information content' as a factor of depth in

the taxonomy. Lower nodes such as ‘wadi’, ‘arroyo’ and ‘draw’ are more similar than ‘valley.’ They are separated by a temporal factor which is more permanent in the feature ‘valley’, and ‘glen’ and temporary in the feature ‘wadi’ and ‘gully.’ These concepts are related by causality, the deformation of the surface of the earth by rainfall, which is not made explicit in the taxonomy. As such it is difficult to quantify ‘what’ information is lost in the generalisation process.

7.5 Conclusions

The immediate value results of the experimentation indicate the potential of the WordNet lexical structure to facilitate knowledge discovery. As a source of additional terms for query expansion, terms that are not within the vocabulary of the user or that are not immediately obvious may be returned. For example, the more domain specific term ‘wetlands’ (Figure 7.12) successfully retrieved the feature ‘swamp’, whilst the more generic (upper level) term ‘water’ can be used to successfully retrieved the feature ‘lake’ or the abstract physical concept of ‘depression’ can be used to retrieve ‘valley’ or ‘ravine’. Using the path length method these terms have a similarity of 0.5, however it is obvious that the term ‘water’ and ‘lake’ are related at a more abstract level than ‘loch’ and ‘lake’ or ‘wadi’ and ‘draw’.

The use of node counting is as such not an accurate indicator of similarity, given the variability in the information content of terms at different depths in the hierarchy. The methods of Jiang and Conrath (1997), Leacock and Chodorow (Scriver 2006) and Resnick (1995a) provide a measure of similarity between terms and determine ‘to what degree one term may be substituted for another’ are also valuable in identifying gaps in the taxonomy. These are more apparent at upper levels of the hierarchy in particular between such concepts as ‘valley’ and ‘depression’ and between ‘land’ and ‘wetland.’

The results presented indicate that WordNet can be used with positive results in applications of geographic query expansion, particularly in the resolution of the Boolean (hit and miss) phenomena characteristic of data retrieval (Section 3). Resolution of the problem of matching a query with the source data in which the

result of a search can either be classified as successful or unsuccessful was improved via the semantic similarity of lexical terms in the WordNet ontology. As illustrated in Figures (7.13), (7.14) and (7.16), in each of these scenarios the feature type retrieved differed from the feature type of the query, indicating that a direct match was not made between query and source data. This indicates that if the terms used in each of example queries were applied in a Boolean retrieval application no results would have been returned. The results of experimentation with query terms indicated in the results obtained in Figures (7.13), (7.14) and (7.16) show that this methodology has implications for knowledge discovery with positive implications for non expert and first time users. Whilst the initial query request was not found in each of these examples, results of a similar nature are returned. This allows the user to understand what data are available in the source dataset and make subsequent queries based on this knowledge.

Analysis of the word groupings indicates that the words pertaining to geographic referents in WordNet are not explicitly defined in the taxonomic structure. Common subsumers such as ‘valley’ and ‘wetland’ suggest the nature of the relationships between the sub concepts they subsume, however this is difficult to ascertain at first inspection, even when incorporating analysis of the word definitions concepts of causality, behaviour and function are not apparent. The concept mappings show the importance of the superclass or closest common subsumer in determining similarity in the taxonomy particularly evident in the method of Resnik (1995a), as terms which are related above or beyond the connection of the superclass are related by more abstract connections. As such it is immediately apparent that a better understanding of the similarity between terms in a structure may be gained via more explicit representation of the abstractions between terms. At present the structure is not representative of the more complex phenomena and mechanisms which relate geographic terms and their word usages as outlined in Section 6.2 questions 2 and 3.

Measures to scale the network based on the depth of a terms in the hierarchy are based on the assumption that the information content of a search term is greater for terms lower in the tree. These differences are not apparent using the node counting scheme. The methods of Leacock and Chodorow (Scriver 2006) and the approach of

Resnik (1995a) more reflect these distinctions using different approaches. An in depth comparison of the approaches of similarity is beyond the scope of this paper however, the use of corpus based approaches inherently has its associated problems. It is very difficult to obtain a statistically valid and reliable corpus that truly reflects word usage; many relatively common words may not appear even in very large corpora. “This problem is usually referred to as the sparse data problem” (Nuno 2005, 9). The results obtained may be systemic of the subjective choice of unique beginners (Section 6.4) which are manifested at the root level of the hierarchy for a conceptualization. As noted by Veal (2005) the main problem is that the relationships that exist between terms are not easy to determine, as the design of WordNet only shows how existing lexemes relate to each other. The concept of information content based on the similarity of terms as a factor which is scaled by the depth of terms in the taxonomy can reveal of how much information is lost but not what meaning is lost. The identification of these ‘gaps’ as described in Section 6.2 question 5 indicate the areas of deficiency in the structure which may benefit from reconceptualization or the incorporation of supplementary concepts. For example in the mapping of the term ‘valley,’ WordNet encodes this grouping under the generalisation abstraction ‘depression,’ however it does make an explicit reference to the other relationships and concepts at work as discussed in Section 5 and 6. For example, the functions which may be afforded by its properties can be used to relate it with other features which share the same properties.

8 CONCLUSIONS AND FUTURE RESEARCH

The objective of information retrieval defined in the context of this paper is to retrieve terms (geographic features) which reflect the user's language. This builds on modern IR theory that usability and design is driven from the user perspective. In this research it has been shown that communities whether they are businesses, government organisations or academic institutions have their own lexicons, to describe the entities which exist in their domains. Building a usable system which accommodates varying domains and conceptualisations of geography has been the primary objective of this research. The research has advocated the use of knowledge structures such as WordNet over corpus based approaches and their inherent disadvantages.

8.1 Summary of Research

Using additional data repositories to supplement user queries via the provision of additional terms has been suggested, and shown to be desirable to achieve probabilistic ranking in database based systems. These are used to overcome the many or no answers problem identified in Section 3 as the primary bottleneck in information retrieval (Objective 1). In information systems design, the aforementioned user based approach is often based on the concept of the metaphor as a base model for interaction in HCI research. The concept on the metaphor revolves about the idea of substitutability of terms and contexts. Ontology and other knowledge structures have been researched as ways of modelling and conceptualizing a domain (Objective 2), and the result of this research has shown that there are both many domains and many ways to conceptualization domains.

The focus of this research has been the ways in which the geographic domain can be conceptualised (Objective 3); as such the contemporary literature points to the abbreviated three universe model (Section 2.6). The abbreviated model summarises the logical and representation universes into ontological vocabulary which is the focus of Section 5.

The result of an ontological study is often a taxonomy (Section 4.6) which is based on specialisation and generalisation relationships. This study implements and evaluates WordNet, which is a dictionary of English language terms with a taxonomic structure as a supplementary source of terms to improve information retrieval. The review of current literature in (Section 5) examines the ways in which people particularly cultural groups communicate about their environment. In particular, the factors which contribute to a word usage are used as the theoretical basis from which the quality of the implementation of the retrieval architecture in Section 7 is analysed.

WordNet's taxonomy is a rich source of terms and it has been shown to be a useful resource as a source of additional terms which can be used for geographic query expansion, and poses a viable alternative to overcome the problems associated with corpus based approaches. The WordNet ontology is however encumbered by two factors which have been discussed in this paper. The first is due to the nature of the structure of the taxonomy which relies on the concept of a single root node and the stipulation that a concept can have one hypernym (Section 4.6). Secondly, the choice of *unique beginners* (Section 6.4) are by the author's own admission subjective. Furthermore, analysis of the mappings of geographic features in Section 7 indicate that the design of WordNet has not been predicated to capture common causality, function and behaviour, but rather shows how existing lexemes relate to each other. There is no one correct way or view when grouping terms (Section 4), as suggested by Mark and Turk (2003b) there are many ways to conceptualise a domain and as such there are a variety of relationships to account for. Words particularly words with geographic referents can be decomposed into several affordance properties, which can be the basis of the relationships in an information retrieval scenario. However these relationships were not present either implicitly in the taxonomy or within term definitions.

Via incorporation of WordNet based similarity metrics, gaps in the taxonomy were identified, which may benefit from the incorporation of more affordance based relationships or a reconceptualisation of current concepts. This paper has purported the view that design of information retrieval should adopt a user focussed approach

and as such WordNet can appear to be a narrow perspective, which runs contrary to the goal of accommodating varied user groups. Wordnet as has been demonstrated can be utilized with success to improve the retrieval (Section 7.4) process, and return data which is similar by a quantifiable value to the query term. Elaboration of the current structure is recommended, to provide greater detail and more diverse relationships between the existing concepts. In addition, modifications to reflect the varied conceptualizations described by the study of language in communities are recommended.

8.2 Methodologies for Enhancement

Typically in GIScience similarity in features can be determined via multidimensional scaling or Tversky's ratio model (Li and Fonseca 2005) which according to Nuno (2005, 11) is more "cognitively plausible." This method is a factor of the number of properties/relations which two categories share, but also exhibit common values for, and also accounts for the number of properties/relations which are present in one but not the other. Analysis of feature similarity often entails grouping features according to attributes they share in common in which a hierarchy manifests itself to which entities lower in the hierarchy possess the features of their parent entities but also additional features which distinguish them.

According to research by (Tversky 1977; Li and Fonseca 2006) there are significant differences in spatial similarity assessment in the GIScience community as compared to that of human cognition and psychology. Li and Fonseca (2006, 2) suggest that psychological spatial similarity assessment emphasizes structural alignment and similarity asymmetry. Where similarity asymmetry means that the similarity of A in relation to B is different from the similarity of B in relation to A. As such, and as has already been explained in Section 5 there are differences between the ways in which different communities class geographic entities as similar. Veal (2005) suggests that by 'tagging' the definitions of a term with an abstract property, *affordances* such as 'water' or 'transportation' for example, then, Tversky's measure of similarity assessment can be applied. Thereby similarity between word forms is not a function of position in the hierarchy but rather based on the number of attributes two terms have in common.

The research of Enfield (2008) and Buhrehult and Levinson (2008) emphasise localised cultural factors contributing to word forms. Dictionary based approaches appear to be the most valuable resource with respect to these utility based approaches, and the relationships created by this approach could be used to supplement the WordNet knowledge base or map lower levels of the taxonomy to alternate concepts. For example the Lao term for swamp (no`o`ng3) could be integrated into the taxonomy via subsumption to the concept of 'wetland', but mapped to an alternate concept such as *source of material* (to gather reeds) or *source of water*. As suggested in Section 5 research emphasising the utility or affordance that a geographic feature provides to the members of a community may provide the best starting point for the integration of geographic terminology between different cultures and languages. Research by Kuhn (2002) (Section 6.3.8) suggests the creation of new relationships by breaking the rules of the taxonomy to allow concepts to have more than one subsuming concept and is a theory cogent with the demand for usability by accommodating multiple conceptualizations.

As discussed in Section 5 ethnophysiography may provide the basis for a data dictionary of geographic terms with an emphasis on the affordances features provide the inhabitants of a culture, implicitly defined in these definitions, being the most pragmatic way towards enrichment of the WordNet ontology which is in keeping with the research question (3) posed in Section 6. Wu and Palmer (1994) and Kuhn (2002) account for the loss of meaning in abstract concepts by creating new mappings by projecting concepts onto controlled common structures. Research by Kavouras, Kokla and Tomai (2005) support the affordance based approach via the extraction of relationships from WordNet glosses predicated by the verb 'for', to determine the purpose of a geographic feature. The use of similarity measures is an indicator of to what degree one term may be substituted for another. The methodology of Wu and Palmer (1994) used in mapping terms between Mandarin and English relies on this metric and is cogent with the goals of obtaining a integrative system based on exploiting the universal salience of concepts in a taxonomy. "Any approach that has a fixed number of target candidate verbs and provides no way to measure the meaning similarity among verbs is not able to handle

new verb usages, i.e., the small portion outside the dictionary coverage" (Wu and Palmer 1994, 134).

New research in the field of computational creativity based on the concept of the metaphor (Veal 2005; Nuno 2005) has been suggested as a method which is comparable to artificial intelligence for the automated generation and recategorization of concepts in which there is more control over the generalization process. In language, a speaker has an unrestricted number of verbs for lexical selection as such the challenge of verb representation is to capture the fluid nature of verb meanings that allows human speakers to contrive new usages in every sentence. Metaphors offer one such mechanism to which the concepts from one domain can be understood in terms of another. Kuhn (2002, 2) defines a metaphor as "a semantic mapping from a source to a target space." As such "conceptual integration in language and other conceptual systems treats the conceptual spaces of the source domains as small theories of concepts that get combined into more complex theories ('blends') which are active at a subconscious level by projection" (Kuhn 2002, 3). An example of this is the metaphor of the desktop. A computer desktop inherits semantic structure from multiple conceptual spaces: physical desktops, office documents, folders, clip boards etc. Words such as 'wetland' are the result of 'blending' with the effect that a concept is created via the combination of the concept of 'water body' and 'land feature'. Veal (2005) suggest this may be achieved via the 'tagging' of noun definitions with pertinent affordances or attributes, mapped to more salient concepts such as *transportation* or *water source*, thereby explicitly defining the relationships between concepts in the structure.

The results of this study highlight the abstract principles which underpin our everyday use of language as a tool for communicating. Confronting these subconscious mechanisms can have immediate benefits for knowledge discovery and understanding phenomena and improving interoperability between user groups. However, because some of these mechanisms are deeply ingrained, articulating these abstractions, or revealing areas of similarity is not always a straightforward task. Whilst implicit relationships have their obvious advantages they can also be subject to misuse, where explicit specification leaves less room for ambiguity, it relies on

more intensive implementation. As has been shown geographic word referents have particular attributes which are based on cultural origins, as such these attributes may present a bias in one cultural group when compared with another. Mapping existing terms based on these properties is therefore next logical step forward for information retrieval in this domain.

9 REFERENCES

Abberley, D., Kirby, D., Renals, S. and Robinson, T. 1999. The THISL broadcast news retrieval system. *In: Proceedings of the ESCA Workshop on Accessing Information in Spoken Audio*, pages 19-24, Cambridge, England.

Agrawal, S., Chaudhuri, S., Das, G., and Gionis, A. 2003. Automated Ranking of Database Query Results. *In: Proceedings of the 2003 Conference on Innovative Data Systems Research*. Asilomar CA. USA

Allocca, L. 2007. MetaCarta Expands Geographic Information Retrieval to Google Earth, ESRI ArcGIS Explorer and NASA World Wind 3D Mapping Software. *Directions Magazine*. http://www.directionsmag.com/press.releases/index.php?duty=Show&id=16595&trv=1 (accessed Jan 12, 2008)

Alspector-Kelly, M.S. 2001. "On Quine on Carnap on Ontology", *Philosophical Studies* Vol. 102, pp. 93-122

Android Technologies Inc. 2007. Artificial Intelligence Discussion Forums http://www.androidtech.com/html/ai-forums.php (accessed May 11, 2007)

Armstrong, D. M. 1997. A World of States of Affairs. Cambridge University Press. United Kingdom.

Bai, J., Song, D., Bruza, P., Nie, Jian-Yun, C. Guihong. 2005. Query Expansion Using Term Relationships in Language Models for Information Retrieval. *In: Proceedings of the CIKM International Conference on Information and Knowledge Management*. Bremen, Germany.

Bakshi, R., Knoblock, C.A., and Thakkar, S. 2004. Exploiting Online Sources to accurately Geocode Addresses, *In: Proceedings of the 12th Annual ACM international workshop on Geographic Information Systems*, Washington D.C., USA.

Balabanovic, M., and Shoham, Y. 1997. Fab: Content-based, Collaborative Recommendation. (Special Section: Recommender Systems). *Communications of the ACM.* Vol 40. No. 3 pp 66

Banerjee, S., and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence,* Acapulco, Mexico, pp 805–810

Berkovski, S. 2006. Carnap and Frege on Ontology. http://www.bilkent.edu.tr/~sandyber/papers.html (accessed May 11, 2007)

Berg, J. 1983. Aristotles Theory of Definition. *ATTI del Convegno Internazionale di Storia dell Logica,* San Gimignano. Bologna, Italy.

Berger, H., Dittenbach, M. and Merkl, D. 2003. Querying Tourism Information Systems in Natural Language. In *Proceedings of the 2nd International Conference on Information Systems Technology and its Applications (ISTA'03),* Lecture Notes in Informatics, Vol. 30.pp 153-164, Kharkiv, Ukraine

Berrios, D.C. 2000. Automated Indexing for Full Text Information Retrieval. *Proceedings of the 2000 Annual Symposium of the America Medical Informatics Association (AMIA)* Los Angeles California.

Bertazzon, S. 2000. (Re)-defining the concept of spatial contiguity: the metaspace of applied analysis. Geographical Domain and Geographical Information Systems Stephan Winter (Ed.) GeoInfo Series 19. *Proceedings of the Euro Conference on Ontology and Epistemology for Spatial Data Standards,* La Londe-lesMaures, France.

Billerbeck, B., Scholer, F., Williams, H.E. and Zobel, J. 2003. Query Expansion using Associated Queries. *Proceedings of the CIKM International Conference on*

Information and Knowledge Management, O. Frieder, J.Hammer, S.Quershi, and L. Seligman (eds), New Orleans, Louisiana, pp. 2-9.

Billerbeck, B., and Zobel, J. 2005. Document Expansion versus Query Expansion for Ad-hoc – Retrieval. *In: Proceedings of the Tenth Australasian Document Computing Symposium,* A Turpin and R. Wilkinson (eds), Sydney, Australia, December 2005, pp.34-41.

Bishr, Y. 1997. Semantic aspects of interoperable GIS, *PhD Thesis,* ITC Publications Series No. 56, Enschede, The Netherlands, 154p

Bishr, Y. 1998. Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science,* Vol. 12, No 4, pp 229--314

Bittner, O. 2000 Ontology, Vagueness, and Indeterminacy, *Conference on the Geographic Domain and Geographic Information Systems - Ontology and Epistemology for Spatial Data Standards,* Agelonde, France

Bittner, T. and Winter, S. 2004. Geo-Semantics and Ontology. *In: Bentley Empowered Conference,* Orlando, Florida..

Bittner, T., Donnelly, M. and Winter, S. 2005. Ontology and semantic interoperability, *Large-scale 3D data integration: Problems and challenges D.* Prosperi and S. Zlatanova (ed.), CRCpress (Tailor & Francis). 139-160

Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. 2006. Adding Dense, Weighted Connections to WordNet. *In: Proceedings of the Thirds International WordNet Conference*. Masaryk University, Brno Slovakia.

Brazer, S. 2007. GeoDecisions Integrates MetaCarta Geographic Search Into IRRIS Logistics Technology. *Yahoo: Financial.*

http://www.directionsmag.com/article.php?article_id=2078&trv=1 (accessed May 11, 2008)

British Columbia Geographical Names. 2008. GeoBC Digital Gazetteer. British Columbia. *Government Home*. http://www.ilmb.gov.bc.ca/bcnames/gaz.html (accessed Oct 11, 2008)

Broder, A. 2006. From Query Based Information Retrieval to Context Driven Information Supply. *Emerging Search Techologies, Yahoo Research*.

Brooks, T.A. 1998. The Semantic Distance Model of Relevance Assessment. *In: Proceedings of the 61st Annual Meeting of ASIS, Information Access in the Global Information Economy*, Pittsburgh, USA. Vol. 35 pp 33-44

Budanitsky, A., and Hirst, G. 2006 Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* Vol. 32, No 1, pp 13-47

Buehler, K. 2003. Data Models and Interoperability. An Open GIS Consortium (OGC) White Paper. *Open GISConsortium*. http://www.opengeospatial.org/pressroom/papers (accessed May 11, 2007)

Burenhult, N. and Levinson, S. 2008. Introduction: Language and Landscape: A cross-linguistic perspective, Burenhult, N. (ed). *Language Sciences*. Vol. 30, no. 2-3. pp. 135-150, Elsevier.

Burners-Lee, T. 2001. Reflections on Web Architecture. *W3C Design Issues*. http://www.w3.org/DesignIssues/CG.html (accessed April 11, 2007)

Buscaldi, D., Rosso, P., Arnal, E.S. 2005. A WordNet-Based Query Expansion Method for Geographical Information Retrieval. *Proceedings of Third International Wordnet Conference (GMC 2005)* Korea (JAPAN)

Buscaldi, D., Rosso, P., and Garcíía, P.P. 2006. Inferring Geographical Ontologies from Multiple Resources for Geographical Information Retrieval IR *In: The 3rd Workshop on Geographic Information Retrieval (GIR 2006)*. Seattle, WA, USA.

Carnap, R. 1937. The Logical Syntax of Language, Routledge and Kegan Paul LTD, London, Great Britain

Carnap, R. 1950. The Problem of Abstract Entities, Empiricism Semantics and Ontology. *Revue Internationale de Philosophie* 4, pp 20-40. University of Chicago Press 1956. http://evans-experientialism.freewebspace.com/carnap.htm (accessed Jan 11, 2009)

Carnap, R. 1958. Introduction to Symbolic Logic and its Applications. Dover Publications, New York, USA

Casati, R., and Varzi, A. 1997. Spatial Entities, *Spatial and Temporal Reasoning* Stock, O. (ed), Dordrecht: Kluwer, pp 73-96

Casati, R., Smith, B., and Varzi, A. 1998. Ontological Tools for Geographic Representation, *Formal Ontology in Information Systems* N. Guarino (ed.), Amsterdam: IOS Press, pp. 77–85

Casati, R., and Varzi, A. 2000. Topological Essentialism. *Philosophical Studies*, Vol. 100, pp. 217-236.

Centre For Geospatial Intelligence. 2004. Automated Scene Description. *Centre for Geospatial Intelligence: Research & Development.* http://geoint.missouri.edu/CGI2/research.aspx (accessed Jan 11, 2008)

Chan, T.O. and Williamson, I.P. 1995. A review of the GIS Planning Methodology for Victoria: Its Relevance to Real Life. *The SEAAS Congress*, Singapore

Chandalia, G., and Srihari, R. 2006. Re-ranking Search Results based on Perturbation of Concept-Association Graphs. *Proceedings of the North East Student Colloquium on Artificial Intelligence*, Ithaca NY, USA.

Chaudhuri, S. Das, G., Hristidis, V., and Weikum, G. 2004. Probabilistic Ranking of Database Query Results. *In: Proceedings of the 30th International Conference on Very Large Databases*. Toronto, Canada.

Chaudhry, O. and Mackaness, W. 2007. Utilising Partonomic Information in the Creation of Hierarchical Geographies. *10th ICA Workshop on Generalisation and Multiple Representation*, Moscow, Russia.

Church, A. 1956. Introduction to Mathematical Logic I. Princeton University Press, Princeton, New Jersey, USA

Cocchiarella, N.B. 2001. Logic and Ontology. *Axiomathes International Journal in Ontology and Cognitive Systems.* Vol. 2 pp 117-150

Cohen, W.W. 2000. Data Similarity Using Similarity Joins and a Word-Based Information Representation Language. *ACM Transactions on Information Systems,* Vol. 18. No. 3. pp 288-321

Copi, I.M. 1979. Symbolic Logic. Prentice Hall,

Croft, W.B. 1995. What do people want from Information Retrieval? *D-LIB The Magazine of the Digital Library Forum (on-line)*. Center for Intelligent Information Retrieval Computer Science Department, University of Massachusetts, Amherst http://www.dlib.org/dlib/november95/11croft.html (accessed Sept 11, 2007)

Deerwester, S., Dumais, S. T., Furnas, G.W., Landauer, T.K., Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science.* Vol. 41. No. 6. pp. 391-407.

De Vries, A.P. and Wilschut, A.N. 2004. On the Integration of IR and Databases. *The Proceedings of the first Workshop on the Integration of Information Retrieval and Databases*. Sheffield, U.K.

Dinesh, S. 2007. Fuzzy Classification of Physiographic Features. Applied Mathematical Sciences, Vol. 1, no. 19, 939-961.

Dittenbach, M., Merkl, D. and Berger, H. 2003. A natural language query interface for tourism information, in A. J. Frew, M. Hitz & P. O'Connor, (eds), *Proceedings of the 10th International Conference on Information Technologies in Tourism (ENTER 2003)*, Springer - Verlag, Helsinki, Finland, pp. 152-162.

Doyle, S., and Daly, M. Enabling Distributed GIS – OpenGIS in the Real World. GISCafe.http://www10.giscafe.com/nbc/articles/view_article.php?articleid=67677&page_no=3 (accessed Sept 11, 2007)

Efthimiadis, E.N. 1996. Query Expansion, Williams, Martha E., (eds). *Annual Review of Information Systems and Technology* (ARIST), v31, pp 121-187, 1996. http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html (accessed Sept 11, 2007)

Egenhofer, M., and Mark, D. 1995. Naïve Geography, *COSIT Frank, A., and Kuhn, W. (ed) Lecture Notes in Computer Science*, Vol. 988, Springer-Verlag, pp. 1-15. Semmering, Austria.

Enfield, N. 2008. Language and Landscape: geographical ontology in a cross linguistic perspective Burenhult, N. and Love, N. (eds). Language Sciences, Vol 30 Issues 2-3, pp. 227-255

Fabrikant, S. I. 2001.Visualising Region and Scale in Information Spaces. *Proceedings of the 20th ICC International Cartographic Conference*, Beijing, China. pp 2522-2529.

Fabrikant, S. I. and Buttenfield, B. P. 2001. Formalizing Semantic Spaces for Information Access. *Annals of the Association of American Geographers,*Vol 91: 263-280.

Fabrikant, S.I., Ruocco, and Middleton, R. 2002. The First Law of Cognitive Geography: Distance and Similarity in Semantic Spaces. *Proceedings of GIScience 2002,* Boulder, CO, Sep. 25-28, 2002: 31-33.

Fabris, N. 2007. Explosive Growth in Real-Time Two-Way Connected Navigation Forecast. *Directions Magazine*. http://www.directionsmag.com/press.releases/?duty=Show&id=20099&trv=1 (accessed Jan 1, 2008)

Fillenbaum, S., and Jones, L. V. 1965. Grammatical Contingencies in Word Association. *Journal of Verbal Learning and Verbal Behavior,* Vol. 4. pp 248-255.

Finin, T., Nicholas, C., and Mayfield, J. 1998. Software Agents for Information Retrieval, *The IEEE Advances in Digital Libraries Conference*. Santa Barbara, CA, USA.

Fisher, P., Wood, J. and Cheng, T. 2005. Fuzziness and ambiguity in multi-scale analysis of landscape morphometry, in Cobb, M., Pettry, F. and Robinson, V. (eds.) Fuzzy Modeling with Spatial Information for Geographic Problems, Berlin: Springer Ch. 10, pp 209-232

Fitting, M. 2001. First-Order Intensional Logic. *Proceedings of the Alfred Tarski Centenary Conference (Niwinski, D., and Wolenski, J. ed)* Warsaw, Poland.

Fonseca, F. 2000. Users, Ontologies and Information Sharing in Urban GIS. In: *ASPRS Annual Conference*, Washington, D.C.

Fonseca, F.T., Egenhofer, M.J., Agouris, P., and Camara, G. 2002a. Using Ontologies for Integrated Geographic Information Systems. In: *Transactions in GIS*. Vol. 6, No. 3, pp 231-257

Fonseca, F.T., Egenhofer, M.J., Davis, C., and Camara, G. 2002b. Semantic Granularity in Ontology-Driven Information Systems. In: *AMA, Annals of Mathematical and Artificial Intelligence – Special Issue on Spatial and Temporal Granularity,* Vol 36, No 1-2, pp 121-151.

Fonseca, F.T., and Li, B. 2006. A Comprehensive Model for Qualitative Spatial Similarity Assessment, In: *Spatial Cognition and Computation* Vol. 6, No 1, pp 31-62

Frietag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., and Wang, Z. 2005. New Experiments in Distributional Representations of Synonymy. *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*. Ann Arbor, Michigan, USA. pp. 25-32

Fuhr, N. 1990. A Probabilistic Framework for Vague Queries and Imprecise Information in Databases. *The Proceedings of the 16th International Conference on Very Large Databases*, Brisbane, Australia.

Gangemi, A., Guarino, N., and Oltramari, A. 2001. Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level. *Proceedings of FOIS International Conference on Formal Ontology in Information Systems*, Ogunquit, Maine, USA.

Garcia, E. 2006. The Classic Vector Space Model. Mi Islita: Online Journal of Information Retrieval Intelligence. http://www.miislita.com/term-vector/term-vector-3.html (accessed March 29, 2008)

Genesereth, M. R. and Fikes, R. E. (1992) Knowledge Interchange Format, Version 3.0 *Reference Manual. Logic Group Report KSL-92-86*. Computer Science Department Stanford University.

Gillam, L., Tariq, M. 2003. Ontology via Terminology. *Proceedings of Workshop on Terminology, Ontology and Knowledge Representation (Termino 2004)*. Lyon, France.

Goodchild, M. F. and Kemp, K.K. 1990. The NCGIA Core Curriculum in GIS. National Center for Geographic Information and Analysis Unit 73, University of California, Santa Barbara CA. http://www.geog.ubc.ca/courses/klink/gis.notes/ncgia/index.html (accessed Dec 3rd, 2008)

Google Maps 2008. Google Maps API. http://code.google.com/apis/maps/ (accessed May 11, 2007)

Gould, M., and Hecht, L. 2001. OGC: A Framework for Geospatial and Statistical Information Integration. *Proceedings of the Joint UNECE/Eurostat Work Session on Methodological Issues Involving the Integration of Statistics and Geography*. Tallinn, Estonia.

Greenwood, J., Hart, G. 2003. Sharing Feature Based Geographic Information – A Data Model Perspective. *Proceedings of the 7th International Conference on Geocomputation*, University of Southampton, U.K.

Grefenstette, G 1992. Use of syntactic context to produce term association lists for text retrieval. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval.* Copenhagen, Denmark pp 89-97

Grenon, P. 2003. Knowledge Management from the Ontological Standpoint. Klaus Freyberg, K., Petsche, H, J., and Klein, B. (eds) *Proceedings of the WM 2003 Workshop on Knowledge Management and Philosophy*, Luzern, Switzerland.

Gruber. T.R. 1993. A translation approach to portable ontologies. *Knowledge Acquisition*, Vol 5. No 2. pp 199-220

Gruber, T. 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies* Vol. 43, Issues 5-6, pp.907-928.

Guarino, N. 1996. Understanding Building and using Ontologies. *In: A Commentary to "Using Explicit Ontologies in KBS Development"*, van Heijst, Schreiber, and Wielinga. International Journal of Human and Computer Studies Vol. 46, pp. 293-310 http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html (accessed Feb 3, 2008)

Guarino, N. 1998. Formal Ontology and Information Systems. *Proceedings of FOIS'98*, Trento, Italy. Amsterdam, IOS Press, pp. 3-15.

Guarino, N. and Welty, C. 2000. A formal Ontology of Properties. *In: Proceedings of 12th Int. Conf. on Knowledge Engineering and Knowledge Management*, Lecture Notes on Computer Science, Springer Verlag

Hansen, P. 1998. Evaluation for IR Interface. Implications for user interface design. *Centre for studies of IT from the Human Computer Interface Institution* , College in Borås, Sweden. http://etjanst.hb.se/bhs/ith//2-98/ph.htm (accessed July 17th 2007)

Harter, S. 1986. Online information retrieval. Concepts, principles, and techniques. Orlando: Academic Press.

Hassam, A., and Maniacco, M. 2005. Orion Technology Inc, and i411, Inc. Form Strategic Partnership. *Directions Magazine.* http://lbs360.directionsmag.com/press/index.php?duty=Show&id=13225 (accessed Jan 3, 2008)

Holt, A. 2001. Understanding Spatial Complexities Utilizing Similes and Metaphors. *Proceedings of the 6th International Conference on GeoComputation*, University of Queensland, Brisbane, Australia 24 - 26 September 2001

Jiang, J.J. and Conrath., D.W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *In: Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, Taiwan, pp 19–33

Jones, C.B, Alani, H., and Tudhope, D. 2003. Geographical terminology servers - closing the semantic divide. Goodchild, M.F., Duckham., M. and Worboys., M.F. (eds) *Foundations of Geographic Information Science*, Taylor and Francis, pp 201-218.

Kavouras, M., and Tomai, E. 2002. Sharpening Vagueness: Identifying Measuring and Portraying its impact on Geographic Categorisations. *The 2nd International Conference on Geographic Information Science*, Boulder, CO, USA,

Kavouras, M., and Tomai, E. 2004. Where the city sits? Revealing Geospatial Semantics in Text Descriptions, *Proceedings of the 7th AGILE conference on Geographic Information Science*, Heraklion, Greece

Kavouras, M., Kokla, M., and Tomai, E. 2005. Comparing categories among geographic ontologies, *Computers and Geosciences*, Vol. 31 pp 145-154

Kardos, J., Moore, A. and Benwell, G. 2003. Visualising Uncertainty in Spatially-Referenced Attribute Data using Hierarchical Spatial Data Structures. *Proceedings of the 7th International Conference on GeoComputation*, University of Southampton, United Kingdom pp 8 – 10

Kemp, K.K. and Vckovsky, A. 1998. Towards an ontology of fields. *In: Proceedings of the 3rd International Conference on GeoComputation*, University of Bristol, United Kingdom

Klien, E. and F. Probst 2005. Requirements for Geospatial Ontology Engineering. *Proceedings of the 8th Conference on Geographic Information Science (AGILE 2005)*. Estoril, Portugal.

Klippel, A. 2000. Representing qualitative spatial knowledge in schematic maps. *Geographical Domain and Geographical Information Systems* Winter, S. (Ed.), Vienna: GeoInfo Series Vol. 19. Institute for Geoinformation, Vienna University of Technology.

Kozima, H. and Ito, A. 1997. Context-sensitive word distance by adaptive scaling of a semantic space. In Mitkov, R. and Nicolov, N. (eds), *Recent Advances in Natural Language Processing: Selected Papers from RANLP'95, Amsterdam Studies in the Theory and History of Linguistic Science: Current Issues in Linguistic Theory*. John Benjamins Publishing Company, Amsterdam/Philadelphia, Vol. 136, Ch. 2, pp 111–124.

Kuhn, W. 1996. Handling Data Spatially: Spatializing User Interfaces. *Proceedings of the 7th International Symposium on Spatial Data Handling, SDH'96,* Advances in GIS Research II, (Kraak, M.-J., & Molenaar, M., eds.), in Delft, The Netherlands (August 12-16, 1996), Published by IGU, Vol. 2, pp: 13B.1 - 13B.23.

Kuhn, W. 2002. Modeling the Semantics of Geographic Categories through Conceptual Integration. *In: Proceedings of the Second International Conference on Geographic Information Science*, Lecture Notes in Computer Science. Vol. 2478, pp 108 - 118

Langefors, B. 1973. Theoretical Analysis of Information Systems. Auerbach. ISBN 0-87769-151-7

Larson, R. R. 1996. Geographic Information Retrieval and Spatial Browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, edited by Linda Smith and Myke Gluck, Urbana-Champaign: University of Illinois. USA. pp 81-124.

Leidner, Jochen, L. 2006. Toponym Resolution: A First Large Scale Comparative Evaluation. *Research Report EDI-INF-RR-0839* (July 2006), School of Informatics, University of Edinburgh. Edinburgh, Scotland, U.K. https://www.inf.ed.ac.uk/publications/online/0839.pdf (accessed May 24, 2007)

Leinfellner, W., Kraemer, E., and Schank, J. (eds.) 1981. Language and Ontology. *Proceedings of the Sixth International Wittgenstein Symposium. 23th to 30th August 1981 Kirchberg am Wechsel (Austria)* - Wien, Hölder-Pichler-Tempsky, pp. 18-20.

Lopez, V., and Motta, E. 2003. Ontology-driven Question Answering in AquaLog. *Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems* (NLDB 2004), Manchester, UK.

Lyons, J. 1977. Semantics. 2 Vols. New York: Cambridge University Press.

Magnini, B., and Speranza, M. 2002. Merging Global and Specialised Linguistic Ontologies. *Workshop on Ontologies and Lexical Knowledge Bases (OntoLex02) 3rd International Language Resources and Evaluation Conference, LREC2002*. Las Palmas, Canary Islands, Spain.

Mark, D.M., Egenhofer, M.J. and Hornsby, K. 1996. Formal Models of Commonsense Geographic Worlds. *Report on the Specialist Meeting of Research Initiative 21 NCGIA Technical Report* 97-2

Mark, D.M., Smith, B. and Tversky, B. 1999. Ontology and Geographic Objects: An Empirical Study of Cognitive Categorization. Freksa, C., and Mark, D. M., (eds), *Spatial Information Theory: A Theoretical Basis for GIS*, Berlin: Springer-Verlag, Lecture Notes in Computer Science No. 1661, pp. 283-298.

Mark, D.M., Skupin, A., and Smith, B. 2001. Features, Objects, and other Things: Ontological Distinctions in the Geographic Domain, in Daniel Montello (ed.), *Spatial Information Theory, Lecture Notes in Computer Science 2205*, Berlin/New York: Springer, pp. 488-502

Mark, D. M., Smith, B., Egenhofer, M. & Stephen Hirtle, S. C. 2004. Ontological foundations for geographic information science, *Research Challenges in Geographic Information Science* R.B. McMaster and L. Usery (eds.). Boca Raton, FL: CRC Press.

Mark, D. M. and Turk, A.F. 2003a. Landscape Categories in Yindjibarndi: Ontology, Environment, and Language. *Spatial Information Theory* – Lecture Notes in Computer Science No 2825, pp. 28-45. Springer-Verlag.

Mark, D. M. and Turk, A.F. 2003b. Talking in COADs: A New Ontological Framework for Discussing Interoperability of Spatial Information Systems. *In: Draft paper for presentation at: Workshop on Spatial and Geographic Ontologies* (prior to COSIT'03)

Mark, D. M. and Turk, A.F. 2003c. Ethnophysiography. *In: Draft paper for presentation at: Workshop on Spatial and Geographic Ontologies on 23rd September, 2003 (prior to COSIT03)*

Mark, D. M. and Turk, A.F. 2004. Ethnophysiography: An Enthnoscience of the Landscape" Centre for Cognitive Science, University at Buffalo, State University of New York. USA.

Mark, D.M., Turk, A.G., and Stea, D. 2007. Progress on Yindjibarndi Ethnophysiography. *In: Winter, S., Duckham, M., Kulik, L. and Kuipers, A.(eds). Spatial Information Theory* – Lecture Notes in Computer Science No 4736, pp. 1-19. Springer-Verlag.

Markowetz, A., Chen, Y., Suel, T., Long, X., and Seeger, B. 2005. Design and Implementation of a Geographic Search Engine. *Proceedings of the Eighth International Workshop on the Web and Databases (WebDB 2005)*, Baltimore, Maryland, USA.

McKee, L. 2001. The Purpose of Geospatial Fusion Services. An Open GIS Consortium White Paper. http://www.opengeospatial.org/pressroom/papers (accessed March 21, 2008)

McKee, L. 2003. The Spatial Web, An Open GIS Consortium (OGC) White Paper. Open GIS Consortium. http://www.opengeospatial.org/pressroom/papers (accessed March 21, 2008)

MetaCarta Inc. 2008. Technology: Geographic Information Retrieval. *MetaCarta, Solutions.* http://www.metacarta.com/solutions/technology/geographic-information-retrieval.html (accessed Jan 3, 2008)

Miller, G. A.1990. WORDNET: An On-Line Lexical Database, *Int. Journal of Lexicography* 3-4:235-312.

Miller, G.A., and Hristea, F. 2006 WordNet Nouns: Classes and Instances. *Computational Linguistics* Vol 32, No 1.

Mishne, G. 2003. Source Code Retrieval using Conceptual Graphs, *Master of Logic Thesis*, Institute for Logic, Language and Computation (ILLC), The University of Amsterdam, The Netherlands.

Montello, D. 1997 *NCGIA Core Curriculum in GIS*, National Center for Geographic Information and Analysis, University of California, Santa Barbara, USA. http://www.ncgia.ucsb.edu/giscc/units/u006/u006.html (accessed Jan 3, 2008)

Moore, G.E. 1965. Cramming more components onto integrated circuits. *The International Journal of Electronics*, Vol 38, No 8.

Myers, B.A. 1995. User Interface Software Tools, *ACM Transactions on Computer-Human Interaction*, Vol 2, No. 1, pp.64-103.

Nahm, U, Y and Mooney, R.J. 2001. Mining Soft-Matching Rules from Textual Data. *The Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01),* Seattle, WA. Pp 979-984

Nambiar, U., and Kambhampati, S. 2004. Mining Approximate Functional Dependencies and Concept Similarities to Answer Imprecise Queries. *The Proceedings of the Seventh International Workshop on the Web and Databases (WebDB 2004),* Paris, France.

National Geospatial-Intelligence Agency. 2006. *Geospatial Intelligence Standards: Enabling a Common Vision*. http://www.fas.org/irp/agency/nga/ (accessed Jan 3, 2008)

Nerlich, G. 1994, The Shape of Space, Second Edition, Cambridge: Cambridge University Press, England.

Nie, J,Y. 2001. A General Logical Approach to Inferential Information Retrieval, *Encyclopedia of Computer Science and Technology,* Ed. A. Kent and J.G. Williams, Vol. 44 pp 203-226

Nunes, J. 1991. Geographic Space as a Set of Concrete Geographical Entities Mark, D. and Frank, A. (eds), *In: Cognitive and Linguistic Aspects of Geographic Space*, Norwell, MA Kluwer Academic, 1991, pp. 9-33.

Nuno, A.L.S. 2005. Computational Models of Similarity in Lexical Ontologies. *Masters Thesis,* University College Dublin

OGC. 2008. Open Geospatial Consortium Inc. Open GIS Standards and Specifications. Open GIS Geographic Mark-up Language (GML). http://www.opengeospatial.org/standards/gml (accessed March 21, 2008)

Øhrstrøm, P., Uckelman, S.L., and Henrik Schärfe, S.L. 2007. Historical and conceptual foundations of diagrammatical ontology. *In: Conceptual Structures:*

Knowledge Architectures for Smart Applications, Simon Polovina, Richard Hill, Uta Priss (eds.). 15th International Conference on Conceptual Structures, ICCS (2007), Sheffield, UK, July 22-27, 2007 - Dordrecht, Springer, pp. 374-386.

OWL. 2004. Web Ontology Language Guide, Michael K. Smith, Chris Welty, and Deborah L. McGuinness, Editors, *W3C Recommendation*, http://www.w3.org/TR/owl-features/index.html (accessed Nov 17, 2008)

Parnas, D.L. 1998. Software Engineering Programs are not Computer Science Programs. Annals of Software Engineering Vol 6. pp 19-37

Pedersen, T. and Michelizzi, J. 2009. WordNet Similarity, University of Minnesota, USA. http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi (accessed Feb 1, 2008)

Peuquet, D. J., Smith, B., and Brogaard-Pederson, B., 1999. Ontology of Fields. *In: Varenius Project Specialist Meeting Report*. Santa Barbara, CA: National Center for Geographic Information and Analysis.

Picard, J. 2000. Probabilistic Argumentation Systems for Information Retrieval. *Ph. D. thesis*, Faculté des sciences, Université de Neuchâtel

Poli, R. 2003.Descriptive, Formal and Formalised Ontologies. *Husserl's Logical Investigations reconsidered Denis Fisette (ed.)*. Dordrecht, Kluwer Academic Publishers. pp. 183-210.

Pressman. R.S. 2001. Software Engineering a Practitioner's Approach, *Fifth Edition. McGraw-Hill*, Singapore.

Probst, F. 2007. Semantic Reference Systems for Observation and Measurement. *PhD Thesis*, University of Munster, Germany.

Qui, Y. and Frei, H. P. 1993 Concept Base Query Expansion, *Proceedings of the Sixteenth Annual International ACM-SIGIR'93 Conference on Research and Development in Information Retrieval*, pp. 160-169.

Quine, W.V.O. 1953. Two dogmas of empiricism. *From a Logical Point of View*: *Nine Logico-Philosophical Essays*. Cambridge, MA: Harvard University Press.

Reed, C. 2004. Integrating Geospatial Standards and Strategies into Business Processes. *An Open GIS Consortium White Paper*. http://www.opengeospatial.org/pressroom/papers (accessed March 23, 2008)

Reid, H. 2006. Center for Geospatial Intelligence: On the Leading Edge. *Directions Magazine*. http://www.directionsmag.com/article.php?article_id=2078&trv=1 (accessed Jan 10, 2008)

Resnik, P. 1995a. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *In: Proceedings of the 14th International Joint Conference on Artificial Intelligence*

Resnik, P. 1995b. Disambiguating Noun Groupings with Respect to WordNet Senses, S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D. Yarowsky (eds.), *In: Natural Language Processing Using Very Large Corpora. Kluwer Academic Publishers*, pp. 77-98.

Resnik, P. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, *In: Journal of Artificial Intelligence Research (JAIR)*, Vol. 11, pp. 95-130

Robertson, S.E. and Jones, K.S. 1994. Simple Proven Approaches to Information Retrieval. *University of Cambridge, Computer Laboratory Technical Report No. 356.* Cambridge U.K. (http://www.cl.cam.ac.uk/TechReports/)

Roddick, J.F., Hornsby, K. and de Vries, D. 2003. A Unifying Semantic Distance Model for Determining the Similarity of Attribute Values, Oudshoorn, M.J. (ed). *Proceedings of the Twenty-Sixth Australasian Computer Science Conference (ACSC2003)*, Adelaide, Australia. CRPIT, 16.. ACS. pp 111-118.

Rodriguez, M.A, Egenhofer, M.J., and Rugg, R.D. 1999. Assessing Semantic Similarities Among Geospatial Feature Class Definitions, *Interoperating Geographic Information Systems, Second International Conference, Interop '99* Vckovski, A., Brassel, K. and Schek, H.-J (eds.), Zurich, Switzerland, *Lecture Notes in Computer Science*, Vol. 1580, Springer-Verlag, pp. 189-202

Rodriguez, M.A, and Egenhofer, M.J. 2004. Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *The International Journal of Geographical Information Science*, Vol 18, No 3. Taylor and Francis Ltd, pp. 229-256 (28)

Rosch, E. 1978. Principles of Categorisation. In: *Cognition and Categorization* Rosch. E. and Lloyd, B.B. (eds.). Hillsdale, NJ: Erlbaum.

Salton, G. 1983. Introduction to Modern Information Retrieval. *McGraw-Hill.*

Safran, C. 2005. A Concept-Based Information Retrieval Approach for User-oriented Knowledge Transfer, Personalized and Contextual Search in the Context of Digital Libraries. *Masters Thesis*, Graz University of Technology, Graz, Austria.

Schneider, L. 2002. Formalised Elementary Formal Ontology. Institute of Cognitive Scinece and Technology, *Italian National Research Council. Laboratory for Applied Ontology ISIB-CNR Technical Report* No 3.

Schwering, A. 2004. 'Semantic Neighbourhood for Spatial Relations' *The 3rd International Conference on Geographic Information Science*, GIScience 04. University of Maryland, Baltimore, USA

Scriver, A.D. 2006. Semantic Distance in WordNet: A Simplified and Improved Measure of Semantic Relatedness, Masters Thesis, University of Waterloo, Ontario Canada.https://www.uwspace.uwaterloo.ca/bitstream/10012/1016/1/adscrive2006.pdf (accessed Jan 10, 2009)

SDTS 2008. Spatial Data Transfer Standard. United States Geological Survey. http://mcmcweb.er.usgs.gov/sdts/ (accessed March 21, 2008)

Smart, P. D., Abdelmoty, A. I., and Jones, C. B. 2004. An Evaluation of Geo-Ontology Representation Languages for Supporting Web Retrieval of Geographic Information. *Proceedings of the GIS Research UK 12th Annual Conference*, Norwich, UK, pp. 175-178. http://www.geo-spirit.org/publications/psmart-gisruk04-final.pdf (accessed May 24, 2007)

Smeaton, A.F. 1995. Natural Language Processing and Information Retrieval. In: *The Second European Summer School in Information Retrieval ESSIR'95*. Glasgow, Scotland.

Smith, B. 2001. Ontology and Information Systems, ontology.buffalo.edu/ontology(PIC).pdf (accessed May 10, 2008)

Smith, B. 2003. Ontology, *Blackwell Guide to the Philosophy of Computing and Information, L. Floridi (ed.),* Oxford: Blackwell, 2003, pp 155–166.

Smith, B. 1995. On Drawing Lines on a Map. Frank, A. U., Kuhn, W., and Mark, D. M. (eds.), *Spatial Information Theory. In: Proceedings of COSIT '95,* Berlin/Heidelberg/Vienna/New York/London/Tokyo: Springer Verlag, 1995, pp. 475–484. http://ontology.buffalo.edu/smith/articles/drawing.html (accessed Jan 10, 2008)

Smith, B., 1996. Mereotopology: A Theory of Parts and Boundaries, *Data and Knowledge Engineering*, 20, 287-303.

Smith, B., and Mark, D.M. 1998. Ontology and Geographic Kinds. *Proceedings, International Symposium on Spatial Data Handling (SDH'98),* Vancouver, Canada, 12-15 July, 1998, pp. 308-320.

Smith, B., and Mark, D.M. 2000. Geographic Categories. *The International Journal of Geographical Information Systems.* Vol. 15, No 7. pp 591-612.

Smith, B. and David M. Mark, 2001, Geographic Categories: An Ontological Investigation, *International Journal of Geographical Information Science*, 15 (7), 591-612.

Smith, B., and Mark, D.M. 2003. Do Mountains Exist? Towards an Ontology of Landforms. *Environment and Planning B,* 30(3), pp 411-427.

Smith, B., and Varzi, A.C. 1999. The Niche, *Noûs,* Vol 33, pp. 214-238.

Smith, B., and Varzi, A.C. 2000. Fiat and Bona Fide Boundaries. *Journal of Philosophy and Phenomenological Research*. Blaxkwell Publishing. Vol. 60, pp 401-420

Stenmark, D. 2003. Query Expansion Using an Intranet-Based Semantic Net. *Proceedings of IRIS-26 Information Systems Research Seminar in Scandinavia*. Porvoo, Finland.

Strzalkowski, T. 1999. Natural Language Information Retrieval. Text Speech and Language Technology Series, Ide, N and Veronis, J. (Eds), Vol 7, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Stubinz, J. and Whighli, S. 2007. Information Retrieval System Design for Very High Effectiveness. *The Second Review of April Fool's day Transactions (RAFT'2007),* Montbonnot, Isère, France.

Sowa, J.F. 1999. Knowledge Representation: Logical, Philosophical, and Computational Foundations, *Brooks Cole Publishing Co*., Pacific Grove, CA, USA.

Sowa, J. F. 2000. Knowledge Representation. Logical, Philosophical, and Computational Foundations, *Brooks Cole Publishing Company*, Pacific Grove, CA, USA. http://www.jfsowa.com/krbook/ (accessed Feb 3, 2008)

Sundheim, B. 2006. Gazetteer Linkage to WordNet. Proceedings of GWC'06 The Third International WordNet Conference. South Jeju Island Korea.

The Australian Government. 2008. Geoscience Australia, The Commonwealth of Australia. http://www.ga.gov.au/ (accessed May 10, 2008)

Thellefsen, M. 2004. Concepts and terminology reflected from a LIS perspective. – How do we reflect meanings of concepts? *The 12th Nordic Conference on Information and Documentation*. Aalborg, Denmark.

Tiainen, E., and Carlson, E. 2006. Spatial Semantics for Geoinformatics. *The Nordic GIS Conference*. Helsinki, Finland.

Terra, E. 2004. Lexical Affinities and Language Applications. *Ph.D thesis*. School of Computer Science, University of Waterloo, Canada

Tversky A. 1977. *Features of similarity*, Psychological review, Vol. 84, pp. 327-352.

Varzi, A.C. 1998. Basic Problems of Mereotopology. *Formal Ontology in Information Systems*, Amsterdam IOS Press, pp 29-38

Varzi, A.C. 2003. Mereology, *Stanford Encyclopedia of Philosophy Zalta., E. N. (ed.)*, Stanford: CSLI. http://plato.stanford.edu/entries/mereology/ (accessed April 5, 2008

Varzi, A.C., Smith, B. 1997. The Formal Ontology of Boundaries. *The Electronic Journal of Analytical Philosophy*. Issue 5 Spring 1997 Methods of Ontology.

Vestavik, O. 2004. Geographic Information Retrieval: An Overview. *Internal Doctoral Conference, IDI, Norwegian University of Science and Technology* http://www.idi.ntnu.no/~oyvindve/ (accessed Sept 18, 2008

Winter, S. 2000. Ontology: Buzzword or Paradigm Shift in GI Science? *EuroConference on Ontology and Epistemology for Spatial Data Standards,* La-Londe-Les Maures, France

Welie, M., and Veer. G. 1999. Ontologies and Methods in Interdisciplinary Design. *Computer Science Education: Challenges for the New Millenium* ,pp. 143-158, Casa Cartii de Stiinta, Cluj

Wen, J., Lao, N., and Ma, W. 2004. Probabilistic Model for Contextual Retrieval. *Proceeding of SIGIR'04 Special Interest Group in Information Retrieval.* Sheffield, South Yorkshire. U.K.

WordNet Lexical Database. 2007. Cognitive Science Laboratory, Princeton University http://wordnet.princeton.edu/ (accessed April 11, 2007)

World Gazetteer. 2007. Population of cities and towns of the world. http://world-gazetteer.com (accessed May 11, 2007)

Wu, Z., and Palmer, M. 1994. Verb Semantics and Lexical Selection. *In: The proceedings of the 32nd Conference of the Association for Computational Linguistics,* pp 133-138. New Mexico State University, Las Cruces, New Mexico, USA

Wyssusek, B., Schwartz, M., and Kremberg, B. 2001. A Philosophical Foundation of Conceptual Knowledge – A Sociopragmatic Approach. *Department of Computer Science, Technical University* Berlin, Germany.

W3C. 2008. The World Wide Web Consortium. http://www.w3.org/ (accessed March 21, 2008)

XML Topic Maps. 2001. TopicMaps.org Specification Pepper, S. and Moore, G. (eds). http://www.topicmaps.org/xtm/ (accessed March 30th 2008)

10 BIBLIOGRAPHY

Alani, H., Jones, B., and Tudhope D. 2001. Voronoi-based region approximation for geological information retrieval with gazetteers. International Journal of Geographic Information Science 15 (4): 287-306

Anderson, C.A. 1998. Alonzo Church's Contribution to Philosophy and Intensional Logic. *The Bulletin of Symbolic Logic* .Vol, 4 No, 2.

Bodenreider, O., and Burgin, A. 2002. Characterising the definitions of anatomical concepts in WordNet and specialised sources. *Proceedings of the First Global WordNet Conference.*

Buscaldi, D., Rosso, P., Arnal, E.S. 2005. A WordNet-Based Query Expansion Method for Geographical Information Retrieval. *Proceedings of Third International Wordnet Conference (GMC 2005)* Korea (JAPAN)

Buscaldi, D., Rosso, P., and Garcíía, P.P. 2006. Inferring Geographical Ontologies from Multiple Resources for Geographical Information Retrieval IR *In: The 3rd Workshop on Geographic Information Retrieval (GIR 2006).* Seattle, WA, USA.

Camara, G., Monteiro, A.M.V., Paiva, J.A., Gomes, J., and Velho, L. 2000. Towards a Unified Framework for Spatial Data Models. *Journal of the Brazilian Computer Society*. Vol. 7, No. 7

Chaudhry, O. and Mackaness, W. 2006. Modelling Geographic Phenomena at Multiple Levels of Detail. *AutoCarto*, Vancouver, Canada.

Chaves, M.S., Silva, M.J., and Martins, B. 2005. A Geographic Knowledge Based for Semantic Web Applications.

Christen, P., Churches, T., and Zhu, J.X. 2002. Probabilistic Name and Address Cleaning and Standardisation, *Proceedings of the Australasian Data Mining Workshop*, Canberra, Australia.

Cockcroft, S. 1997. A Taxonomy of Spatial Data Integrity Constraints. *The Information Science Discussion Paper Series*. No 97, Vol. 05.

Dramowicz, E. 2004. Three Standard Geocoding Methods, Directions Magazine, Directions Media. http://www.directionsmag.com/article.php?article_id=670&trv=1(accessed December 11, 2006)

Fabrikant, S. I. 2000. The Geography of Semantic Information Spaces. *GIScience 2000*, Savannah, GA Oct. 28-31, 2000.

Fabrikant, S. I. and Buttenfield, B. P. 1999. Exploring Semantic Spaces: Geographic Metaphors for Information Access. In: *The Annals of the Association of American Geographers*.

Fonseca, F., and M. Egenhofer. 1999. Ontology-Driven Geographic Information Systems C. B. Medeiros (Ed.), *7th ACM Symposium on Advances in Geographic Information Systems*, Kansas City, MO, pp. 14-19

Fu, F., Jones, C.B., and Abdelmoty, A.I. 2005. Building a Geographical Ontology for Intelligent Spatial Search on the Web. *Proceedings of IASTED International Conference on Databases and Applications (DBA2005)*

Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. 1987. The Vocabulary Problem in Human-System Communication: an Analysis and a Solution. *Communications of the Association for Computing Machinery, Vol. 30 No 11* pp 964-971

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. 2001. Understanding top-level ontological distinctions. *Proceedings of IJCAI The 17th International Joint Conference on Artificial Intelligence: workshop on Ontologies and Information Sharing*, Seattle Washington, USA.

Gillam, L., Tariq, M. and Ahmad, K. 2005. Terminology and the Construction of Ontology**.** *Terminology*. (In Press).

Goldberg, D.W., Wilson, J.P. and Knoblock, C.A. 2006. From text to Geographic Coordinates: The Current State of Geocoding, *Journal of the Urban and Regional Information Systems Association*. http://www.urisa.org/goldberg(accessed December 11, 2006)

Hadacz, L., and Horák, A. 2000. Knowledge Representation and Reasoning with Transparent Intensional Logic. *Proceedings of the Fourth Joint Conference on Knowledge-Based Software Engineering JCKBSE'2000.* Amsterdam, Netherlands. pp 74-80

Husik, I. 1952. The Categories of Aristotle. *Philosophical essays, ancient, mediaeval, and modern* - Edited by Milton C. Nahm and Leo Strauss, Oxford, Blackwell, pp. 96-112.

Hutchinson, M. and Veenendaal, B. 2005. Towards a Framework for Intelligent Geocoding. Proceedings of the SSC 2005 Spatial Intelligence, Innovation and Praxis: The national biennial Conference of the Spatial Sciences Institute, Melbourne, AU, September, 2005.

Hutchinson, M. and Veenendaal, B. 2005. Towards Using Intelligence To Move From Geocoding To Geolocating. *Proceedings of the 7th Annual URISA GIS in Addressing Conference*, Austin, TX, USA, August, 2005.

Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., and Weibel, R. 2002. Spatial Information Retrieval and Geographical Ontology's: An

Overview of the SPIRIT project. SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, ACM Press, pp.387 - 388.

Karasová, V., Jukka, M. K. and Kirsi, V. 2005. Application of Spatial Association Rules for Improvement of a Risk Model for Fire and Rescue Services. *Proceedings of the 10th Scandinavian Research Conference on Geographical Information Science (ScanGIS 2005)*, Stockholm, Sweden, 13-15 June 2005

Kavouras, M., Kokla, M., and Tomai, E. 2005 Comparing categories among geographic ontologies. *Computers and Geosciences* Vol 31, pp 145-154

Keet, C.M. 2006. Introduction to part-whole relations: mereology, conceptual modelling and mathematical aspects. *Knowledge Representation Meets Databases Research Centre Technical Report.* Faculty of Computer Science, Free University of Bozen-Bolzano, Bolzano, Italy.

Kuhn, W. 1993. Metaphors Create Theories for Users. In: *Spatial Information Theory*. Frank, A.U., and Campari, I. (eds). Lecture Notes in Computer Science, Vol. 716, Springer, pp: 366-376

Landauer, T. K., Foltz, P. W., and Laham, D. 1998. Introduction to Latent Semantic Analysis.*Discourse Processes, No* 25, 259-284.

Larson, R. R. 1996. Geographic Information Retrieval and Spatial Browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, edited by Linda Smith and Myke Gluck, Urbana-Champaign : University of Illinois. USA. pp 81-124.

Leidner, Jochen, L. 2005. Preliminary Experiments with Geo-Filtering Predicates for Geographic IR. *CLEF 2005 Working Notes. Cross-Language Evaluation Forum Workshop (CLEF 2005)*, Vienna, Austria. http://www.iccs.informatics.ed.ac.uk/~s0239229/documents/leidner-2005-geoclef.pdf (accessed May 24, 2007)

Lembo, A.J., Bonneau, A., and O'Rourke, T. 2004. Advanced Web-Based GIS Management Technology. Research Progress and Accomplishments: A selection of technical achievements of the multi-disciplinary centre for earthquake engineering research. National Science Foundation EEC-9701471

Lenci, A (2001) Building an Ontology for the Lexicon: Semantic Types and Word Meaning, Jensen, P. A. Skadhauge, P.R. (eds). In: *Ontology-Based Interpretation of Noun Phrases*, Department of Business Communication and Information Science, University of Southern Denmark – Kolding: 103-120.

Lui, X., and Croft, B. (2004) Statistical Language Modelling for Information Retrieval. The Annual Review of Information Science and Technology, Vol. 39, pp. 3-31.

Mackaness, W., and Chaudhry, O. 2006. Creating and Using Partonomic Structures to Support Consistency in Geographical Databases. *Inspire*, Italy

Maki, W., McKinley, L.N. and Thompson, A.G. 2004. Semantic Distance Norms Computed from an Electronic Dictionary (WordNet). *Behaviour Research Methods, Instruments and Computers* Vol 36 No. 3, pp 421-431

Martins, B., Silva, M.J. and Andrade, L. 2005. Indexing and Ranking in Geo-IR Systems. *The Proceedings of the 2005 Workshop On Geographic Information Retrieval*, Bremen, Germany

Mennis, J., and Liu, J.W. 2003. Spatial Association Rules in Spatio-Temporal Data. *Proceedings of the 7th International Conference on GeoComputation*, University of Southampton, United Kingdom 8 - 10 September 2003

Moyaert, J.R. 2004 Geocoding Challenges and the Internet, Directions Magazine, Directions Media. http://www.directionsmag.com/editorials.php?article_id=504 (accessed December 11, 2006)

Muresan, G. 2006. An Investigation of Query Expansion Terms, *Proceedings of the Annual Meeting of the American Society for Information Science and technology (ASIST 2006)*, Austin, Texas.

Openshaw, S. 1999. Geographical Data Mining Key Design Issues. *Proceedings of the 4th International Conference on GeoComputation* Mary Washington College Fredericksburg, Virginia, USA 25 - 28 July 1999

Orilia, F. 1999.Foundations of Intensional Logic and Natural Language Semantics. *ESSLLI 99 The Eleventh European Summer School in Logic, Language and Information.* Utrecht, The Netherlands

Parker, N. 2005. Point Level Geocoding in Group 1 Project, Directions Magazine, Directions Media. http://www.directionsmag.com/article.php?article_id=933 (accessed December 11, 2006)

Parker, N. 2004. A look at NAC Geographic, Directions Magazine, Directions Media. http://www.directionsmag.com/article.php?article_id=654&trv=1 (accessed December 11, 2006)

Paull, D. 2003 *A Geocoded National Address File for Australia: The G-NAF What, Why, Who and When,* Public Sector Mapping Agencies, Canberra, ACT, Australia. http://www.psma.com.au/about-g-naf (accessed December 11, 2006)

Pazienza, M., Pennacchiotti, M., and Stellato, A. 2006. Ontological Support to Knowledge Management in a Hydrogeological Information System. *The Seventh International Conference on Data, Text and Web Mining and their Business Applications and Management Information Engineering*, Prague, Czech Republic.

Public Sector Mapping Agencies Australia. 2006. *Geocoded National Address File Product Description*. http://www.psma.com.au/g-naf-technical-info (accessed December 11, 2006)

Public Sector Mapping Agencies Australia. 2006. *Geocoded National Address File Product Process Flow Diagram*. http://www.psma.com.au/g-naf-technical-info (accessed December 11, 2006)

Probst, F., Espeter, M. 2006. Spatial Dimensionality as Classification Criterion for Qualities. *FOIS'06, International Conference on Formal Ontology in Information Systems*, Balitmore, USA.

Rijsbergen, C.J. 1979. Information Retrieval, 2nd edition. *Butterworth-Heinemann.*

Sabou, M., Lopez, V., Motta, E., and Uren, V. 2006. Ontology Selection: Ontology Evaluation on the Real Semantic Web. *Proceedings of the EON'2006 Workshop, "Evaluation of Ontologies on the Web", held in conjunction with WWW'2006*

Schiavone, B. 2006. *Terra Address Helper, Developers User Guide*, Groffen, J. (ed). LISAsoft Pty Ltd, Adelaide South Australia.

Sears, B. 2004. Geocoding Challenges: Why Accuracy Matters, Directions Magazine, Directions Media. http://www.directionsmag.com/article.php?article_id=558&trv=1 (accessed December 11, 2006)

Smith, B., and Bittner, T. 2001. A Taxonomy of Granular Partitions. Spatial Information Theory, COSIT 2001, Morro Bay, CA, USA.

Smith, B., and Brogaard, B. 2000. Quantum Mereotopology. *Spatial and Temporal Granularity. Papers from the AAAI Workshop (AAAI Technical Report WS-00-08)*, Menlo Park: AAAI Press, pp 25-31.

Shyllon, E. Hutchinson, M., and Veenendaal, B. 2006. Developing an Address Knowledge Base for intelligent Geocoding. WALIS Forum, Perth, Western Australia.

Tanasescu, V., and Domingue, J. 2006. Toward User Oriented Semantic Geographical Information Systems. *Proceedings of the 2nd AKT Doctoral Symposium*, Aberdeen Univeristy, UK,

Tomai, E., and Spanaki., M. 2005. From ontology design to ontology implementation: A web tool for building geographic ontologies, *Proceedings of the 8th AGILE Conference on Geographic Information Science,* Estoril, Portugal, May 26, 28, 2005.

Vallez, M. and Pedraza-Jimenez, R. 2007. *Natural Language Processing in Textual Information Retrieval and Related Topics*. Hipertext.net, No. 5. http://www.hipertext.net (accessed May 10, 2007).

Veenendaal, B., Hutchinson, M. 2005. Improved Emergency Response through Intelligent Geocoding, WALIS Forum, Perth, Western Australia.

Vogel, C. 2001. Dynamic Semantics for Metaphor. *Metaphor and Symbol.* Vol 16, No. 1&2, pp 59-74

Waters, T., and Evans, A. 2003. Tools for web-based GIS mapping of 'fuzzy' vernacular geography. *Proceedings of the 7th International Conference on GeoComputation*, University of Southampton, United Kingdom 8 - 10 September 2003

Wei, X., and Croft, W.B. 2007. Investigating Retrieval Performance with Manually-Built Topic Models. *RIAO 8th Conference Large-Scale Semantic Access to Content (Text, Image, Video, Sound)* Pittsburgh P.A. USA.

Welie, M., and Veer. G. 1999. Ontologies and Methods in Interdisciplinary Design. *Computer Science Education: Challenges for the New Millenium* ,pp. 143-158, Casa Cartii de Stiinta, Cluj

Welty, C., and Guarino, N. 2001. Supporting ontological analysis of taxonomic relationships. *Data and Knowledge Engineering*, Vol 39. pp 51-74

Welty, C., and Guarino, N. 2000. A Formal Ontology of Properties. *Proceedings of 12th International Conference on Knowledge Engineering and Knowledge Management*, Lecture Notes on Computer Science, Juan-les-Pins, French Riviera France.

Wermelinger, M. 1995. Conceptual Graphs and First-Order Logic. *Conceptual Structures: Applications, Implementation and Theory*, Lecture Notes in Computer Science. Vol. 954. Springer, Berlin/Heidelberg, Germany

Wermelinger, M., and Lopez, J.G. 1994. Basic Conceptual Structures. *Proceedings of the Second International Conference on Conceptual Structures*. Lecture Notes in Artificial Intelligence, Springer-Verlag, College Park, MD, USA.Vol. 835 pp 144-159

Wilson, J.P., Lam, C.S., and Holmes-Wong, D.A. 2004. A New Method for the specification of Geographic Footprints in Digital Gazetteers. *Cartography and Geographic Information Science*, Vol.31, No. 4, pp. 195-207

Witt, M., and Turau, V. (2006) Delivery Semantics for Geographic Routing, *Marrón, P.J. (ed.)*. In: *5. GI/ITG KuVS Fachgespräch "Drahtlose Sensornetze"*. Technical Report 2006/07. University of Stuttgart, Germany.

Every reasonable effort has been made to acknowledge the owners of the copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

11 APPENDIX A

1. Aircraft Facilities,
2. Built up areas,
3. Caves,
4. Flats,
5. Islands,
6. Lakes,
7. Locations: hills, mountains etc.,
8. Marine Hazards,
9. Mine Areas,
10. Mine Points,
11. Place Names,
12. Populated Places,
13. Prohibited Areas,
14. Recreational Areas,
15. Reserves,
16. Reservoirs,
17. Road Crossing Points,
18. Watercourse Areas,
19. Waterfall Points,
20. Waterholes.

12 APPENDIX B

1. Mountain
2. Locality
3. District Municipality
4. Lake
5. Indian Reserve
6. Community
7. Creek
8. Swamp
9. Reef
10. Peak
11. Railway Point
12. Inlet
13. River
14. Hill
15. Point
16. Pass
17. City
18. Ridge
19. Island
20. Glacier
21. Bank
22. Rock
23. Peninsula
24. Sound
25. Islet
26. Passage
27. Cove
28. Islands
29. Shoal
30. Rocks
31. Range

32. Bay
33. Harbour
34. Provincial Marine Park
35. Provincial Park
36. Plateau
37. Lakes
38. Head
39. Channel
40. Banks
41. Slough
42. Former Locality
43. Valley
44. Cave
45. Hotsprings / Hot Springs
46. Post Office
47. Narrows
48. Abandoned Locality
49. Reach
50. Land District
51. Hotspring / Hot Spring
52. Canyon
53. Lagoon
54. Icefield
55. Dome
56. Brook
57. Landing
58. Marsh
59. Village
60. Falls
61. Bridge
62. Tunnel
63. Rapids
64. Reefs

65. Pond
66. Arm
67. Coulee
68. Cone
69. Bluff
70. Basin
71. Ranges
72. Bight
73. Islets
74. Flats
75. Anchorage
76. Spit
77. Riffle
78. Recreational Community
79. Peaks
80. Hills
81. Provincial Recreation Area
82. Tower
83. Notch
84. Waterfall
85. Canadian Forces Station
86. Knob
87. Trough
88. Butte
89. Indian Government District : Land Un
90. Provincial Heritage Property
91. Group
92. Gulch
93. Flat
94. Reservoir
95. Beach
96. Protected Area
97. Canal

98. Spire
99. Prairie
100. Pinnacle
101. Whirlpool
102. Fumarole
103. Meadow
104. N,v,
105. Shoals
106. Bar
107. Ledge
108. Bog
109. Cliff
110. Stream
111. Cascade
112. Province
113. Lookout
114. Archipelago
115. Strait
116. Entrance
117. Isthmus
118. World Heritage Site
119. Regional District
120. Mountains
121. Urban Community
122. Col
123. Crags
124. Forest
125. Patch
126. Former Indian Village
127. National Historic Site
128. Caves
129. Tidal Rapids
130. Crater

131. Mount
132. Gap
133. Town
134. Cliffs
135. Crag
136. Dam
137. Ponds
138. Recreation Facility
139. Spur
140. Elbow
141. Domes
142. Escarpment
143. Bend
144. Ledges
145. Spires
146. Volcanoes
147. Summit
148. Gorge
149. Bluffs
150. Site
151. Cirque
152. National Park
153. Knoll
154. National Park Reserve
155. Springs
156. Plains
157. Slide
158. Airfield
159. Region
160. Bench
161. Plain
162. Picnic Area
163. Towers

164. Ditch
165. Regional Park
166. Provincial Historic Park
167. Historical Route
168. Spring
169. Highland
170. Indian Village
171. Ocean
172. Glaciers
173. Ice Cap / Icecap
174. Ravine
175. Beaches
176. Port
177. Portage
178. Trench
179. Indian Government District
180. Cape
181. Fishing Site
182. Bays
183. Airport
184. Eddy
185. Trail
186. Resort Municipality

13 APPENDIX C

```
foreach ($expandedSearchArray as $key => &$keyRow) {

$compArray =
distVincenty($latitude, $longitude, $keyRow['lat'],$keyRow['lng']);

$keyRow['distance'] = $compArray[0];
$keyRow['bearing'] = abs($azimuth - abs($compArray[1]));

if(!isset($azimuth)){
        if($minDistance > $keyRow['distance']){
        $minDistance = $keyRow['distance'];
        $theLat  = $keyRow['lat'];
        $theLong = $keyRow['lng'];
        $theName = $keyRow['name'];
        $theType = $keyRow['type'];
        $wordRank = $keyRow['wordRank'];
        $theBearing = $keyRow['bearing'];
        }
}else{
        if($minBearing > $keyRow['bearing']){
        $minDistance = $keyRow['distance'];
        $minBearing = $keyRow['bearing'];
        $theLat  = $keyRow['lat'];
        $theLong = $keyRow['lng'];
        $theName = $keyRow['name'];
        $theType = $keyRow['type'];
        $wordRank = $keyRow['wordRank'];
        $theBearing = $keyRow['bearing'];
        }
}
}
```

```
if(!isset($azimuth)){
        $featureArray = remove_dups($expandedSearchArray, 'distance');
        foreach ($featureArray as $key => $row) {
                $word[$key]  = $row['wordRank'];
                $distance[$key]  = $row['distance'];
        }
array_multisort($word, SORT_DESC, $distance, SORT_ASC, $featureArray);

}else{
        $featureArray = remove_dups($expandedSearchArray, 'bearing');
        foreach ($featureArray as $key => $row) {
            $bearing[$key]  = $row['bearing'];
        }
array_multisort($word, SORT_DESC, $distance, SORT_ASC, $featureArray);

}
```