Spatial Information Retrieval with Place Names

von Jörg-Thomas Vögele

Dissertation

zur Erlangung des Grades eines Doktors der
Ingenieurwissenschaften
– Dr. Ing. –

Vorgelegt im Fachbereich 3 (Mathematik und Informatik)
der Universität Bremen
im März 2004

Datum des Promotionskolloquiums: 28. Mai 2004

1. Gutachter: Prof. Dr. C. Schlieder, Universität Bamberg
2. Gutachter: Prof. Dr. C. Freksa, Universität Bremen

Acknowledgements

This work is the result of research I pursued during my stay at the *TZI Center for Computing Technologies, Universität Bremen*. I would like to thank the head of my department, Prof. Dr. Otthein Herzog, for giving me the opportunity and the necessary material support to conduct this research. I wish to thank all members of the *Artificial Intelligence Research Group*, and in particular my colleagues Rainer Spittel, Sebastian Hübner, Dr. Ubbo Visser, Björn Gottfried and Dr. Heiner Stuckenschmidt, for providing me with a friendly and intellectually stimulating work environment.

My special thanks go to Prof. Dr. Christoph Schlieder (Universität Bamberg) for his motivating support. He, as the intellectual mentor of my thesis, helped me to find my scientific path and engaged me in numerous fruitful and supportive scientific discussions.

I would also like to express my gratitude to Prof. Dr. Christian Freksa (Universität Bremen), not only for evaluating my thesis, but also for providing me with many interesting new ideas through his scientific work and his research group.

Last but not least, I want to thank my family, and in particular my wife Inger Seemann, for their patience and understanding, especially during the final months that led to the completion of my thesis.

Jörg-Thomas Vögele
Bremen, July 2004

Abstract

This work outlines an approach to handle the spatial dimension of information retrieval, i.e. queries of the type *concept @ location*. It applies a low-level notion of *spatial relevance* that is closely linked to the location, or *geo-reference*, of an information object. Methods for the semi-qualitative representation of spatial regions, and for the evaluation of the relative spatial relevance of information objects that are *indirectly geo-referenced* by *place names*, are presented.

The approach supports a representation of crisp and vague geographic regions that is adequate for most information retrieval tasks. A graph-based abstraction of the underlying discrete referencing system of *polygonal tessellations* reduces data volumes and supports efficient reasoning algorithms. By combining a semi-qualitative approximation of Euclidean distance with a relative measure for the hierarchical part-of relations of regions into one simple metric, the partonomic semantics of hierarchical *place name structures* can be leveraged for information retrieval processes.

The method was evaluated against a large-scale operational meta-information system (*Umweltdatenkatalog - UDK*). Implementations include the *BUSTER* information broker middleware and an *OGC* compatible catalog-service within a spatial data infrastructure. Other potential application areas include the creation of machine-readable indices of digital maps, the development of spatial reasoning modules for highly distributed ad-hoc (*P2P*) networks and *Semantic Web* applications, as well as the integration into new light-weight and location-based applications.

Contents

List of Figures

Part I

Introduction and Related Work

Chapter 1

Introduction and Motivation

1.1 Introduction

Naive geographic reasoning is one of the most common and most basic forms of human intelligence [Egenhofer and Mark, 1995]. Many decisions in our daily lives depend on knowledge about geographic space, geographic locations, and the spatial relation between geographic objects. Often, they involve the comparison of two objects based on spatial parameters:

- If we want to go out for dinner, we not only have to find the right *type* of restaurant, but also one that is conveniently *located*. After a tiring working day, we may prefer the pizzeria around the corner to our favorite restaurant at the other end of town.
- If we plan to buy a house, one of the most important parameters to consider is the *location* of the property. Location in this context refers to parameters like the socio-economic structure of the *neighborhood*, whether public transport is *accessible*, or if there are shopping facilities *nearby*.
- If we have to decide between two interesting job offers, the *distance* to the new workplace is a relevant criterion. However, what ultimately counts is the actual commute time. This can be proportional to Euclidean distance, but it often depends on factors like which means of transportation are available, or if there are good (road) connections.

A common task in the domain of information discovery and information retrieval is to select from large, often distributed and heterogeneous collections (e.g., the Internet) those information objects (e.g., web pages) that promise to have the highest relevance with respect to an information request. Given the importance of space and location there is a need for search algorithms that take into account the spatial as well as the thematic dimension of an information request: A search algorithm should not only be able to find the right type of information object, but rather the right *type* at the right *place*. In general terms, a comprehensive search algorithm should support requests of the type *concept @ location* [Schlieder et al., 2001].

Within the last decade, considerable effort has been invested in the development of methods that address the first, i.e., the conceptual part of such a query. Many of these methods are based on qualitative representations of conceptual knowledge and support complex reasoning about the conceptual, or *semantic* relevance of information objects. Some of the more advanced approaches in this area use representations based on formal ontologies to encode implicit conceptual knowledge. Because these representations support reasoning with decidable sub-sets of first-order logic (e.g., description logics), software tools based on such approaches have been applied in practical applications and information retrieval systems (for an overview see [Wache et al., 2001]).

On the other hand, comparatively little has been done so far to develop methods that address the spatial dimension of an information request[1]. The bulk of systems used to handle spatial representations and spatial queries pertains to the domain of *Geographic Information Systems (GIS)*, *spatial databases*, and *gazetteers*. In these systems, geographic space is represented as a 2- or 3-dimensional coordinatized space. Objects and locations within this space are represented on the basis of exact coordinates. Spatial relations between such objects are computed with the help of algorithms taken from computational geometry.

Coordinatized spaces, exact representations of objects and locations, and quantitative methods to evaluate spatial relations are effective, but do have a number of shortcomings: For one, many geographic regions and geographic objects are either indeterminate (i.e., the location of their exact boundary is unknown), or they are intrinsically vague (i.e., they do not have an exact boundary at all)[Cohn and Gotts, 1996b]. Exact representations are not well suited to cope with vague spatial objects. They often enforce a level of precision that does not reflect the conditions found in real-world applications.

Secondly, common-sense reasoning in geographic space is qualitative by nature. Many cognitive studies indicate that humans prefer qualitative over quantitative methods when it comes to modelling and reasoning about space, in particular geographic space (e.g., [Freksa and Roehrig, 1993, Egenhofer and Mark, 1995]). This has been recognized by the GIS community as a major impediment to a user-friendly and intuitive interaction with spatial information systems [Couclelis, 1992, Egenhofer and Kuhn, 1998].

Thirdly, quantitative representations of objects in geographic space tend to be complex and bulky. This is true in particular when the representations try to encode spatial objects at a high level of detail, which is needed to perform meaningful spatial computations. The complexity of the representations introduces a number of problems related to data quality (e.g., [Goodchild, 1993, Worboys, 1998]), data interoperability, and the computational costs associated with the processing of the data.

Overall, the limitations mentioned above reduce the value of quantitative representations and quantitative reasoning methods for information retrieval applications. This is true in particular for applications that are targeted at the

[1] An exception is the work of Alani who, for the application domain of cultural heritage and archeological artefacts, developed an information retrieval approach that makes use of what he calls *"an integrated thematic and spatial ontology"*[Alani, 2001]. While the general idea of the spatial part of Alani's approach is very similar to the one described in this work, he uses quantitative representations of spatial objects and a metric based on Euclidean distance to compute spatial relevance.

retrieval *"indirectly"* geo-referenced information objects, i.e., objects for which a reference to geographic space is given by a reference to named geographic objects. Indirectly geo-referenced information objects can be found in large collections of unstructured and semi-structured data like the WWW and the *"Semantic-Web"*, but also in digital libraries and other collections of data that are annotated with (spatial) metadata.

We argue that for information retrieval applications in general, and for the applications mentioned above in particular, there is a need for simple, yet expressive methods to resolve indirect geo-references, and to reason about the spatial relevance of indirectly geo-referenced information objects. This implies the need for adequate representations of geographic objects, and for methods that support spatial reasoning on basis of these representations.

1.2 General Approach

The goal of this work is to find a simple but expressive representation for geographic objects and to develop effective algorithms for spatial relevance reasoning based on this representation. The resulting method should be able to support practical information retrieval applications in the evaluation of the spatial dimension of information requests of the type *concept @ location*, i.e., in the evaluation of the *spatial relevance* of information objects. Our approach is based on the assumption that the spatial relevance of two objects in geographic space is a function of the spatial relations between these objects. When thinking about which spatial relations may be important in this context, metric (e.g., *distance*) and topologic (e.g., *intersection*) relations come to mind immediately. Other spatial relations, like mereologic relations (*partonomy*) or ordinal relations (*direction*) are less obvious at first sight, but may be important nevertheless. In section 4.1 of this work, we will try to define spatial relevance and the parameters, or spatial relations, it depends on. We will examine these parameters and select the ones that prove to be the most important ones.

We already mentioned that our approach has a strong focus on the retrieval of so-called *"indirectly geo-referenced"* information objects. Indirectly geo-referenced information objects are not linked to geographic space by (geographic) coordinates, but through a reference to other (named) geographic objects. The natural language identifiers that are given to such geographic reference objects are usually called *place names*. We are interested in indirectly geo-referenced information because for one, there are indications that a large fraction of the geo-referenced data and information objects available in large information collections are in fact indirectly geo-referenced[2]. Secondly, place names are often used in common-sense spatial reasoning to model geographic space and to identify objects in geographic space. Place names are therefore important components of intuitive and user-friendly interfaces to spatial data.

Standard applications to organize place names are place name lists, or *gazetteers*. Today, digital gazetteers with up to several million place name entries are available (e.g., the gazetteer of the *Alexandria Digital Library* [ADL, 2004] or the *Getty Thesaurus of Geographic Names* [TGN, 2002]). They play an important role for the management of place names as well as for providing a common

[2]This assumption is supported by the fact that place names play an important role even in metadata standards that were specifically developed for geospatial data.

standardized vocabulary of place names in information retrieval applications.

In a gazetteer, place names are related to geographic space through their *spatial footprint* which represents an approximation of the 2-dimensional extent of the named object in geographic space. Which types of spatial footprint representations are used has a significant impact on the quality and complexity of the spatial reasoning supported by a gazetteer. Most footprint representations are coordinate-based and exact representations of geographic objects, at different levels of dimensionality and complexity. However, although there are systems that support complex polygonal footprints, most state-of-the-art systems still use mainly simple point footprints to represent place names.

This reflects a general dilemma faced by most gazetteer applications: While highly simplified spatial footprints are easy to obtain and to manage, they limit the expressiveness of the respective spatial reasoning algorithms. On the other hand, more complex representations (e.g., in form of polygonal representations) are often not available and introduce a number of problems generally associated with geospatial *GIS* data (e.g., problems related to data quality, to costly data acquisition, and to complex and ineffective computations). In addition, gazetteers (like all other systems that rely on quantitative spatial representation and reasoning) suffer from their inability to adequately address the qualitative nature of human spatial reasoning, and the human preference for vaguely stated and semantically ambiguous spatial queries.

In search for alternative methods for spatial representation and spatial reasoning, we look at the scientific field of *Qualitative Spatial Reasoning (QSR)*. In *QSR*, numerous methods for the qualitative representation of space, and the qualitative reasoning about spatial relations, have been developed within the last two decades. As we will show in this work, all spatial relations that are important within the context of spatial information retrieval and spatial relevance reasoning can be addressed by qualitative methods. However, many of these methods either do not work very well in practical applications or address only very specific subsets of spatial relations. Because we are interested in applying these methods in practical applications, and because the spatial relevance of geographic objects depends on a number of different spatial relations, we have to find appropriate qualitative or semi-qualitative methods and combine them into an integrated approach.

In this work we will outline an approach for the representation of geographic objects based on a semi-qualitative representation in discrete space. This representation uses qualitative spatial footprints obtained through a projection of regional geographic objects onto polygonal standard reference tessellations of administrative subdivisions or postal code zones. Using graph-based representations of these tessellations we are able to achieve a considerable simplification and reduction of data volumes. At the same time, the representation is designed to retain enough information necessary to support a simple but effective metric to compute the relative spatial relevance of place names. Because in this approach no coordinates are needed for the representation of geographic objects, we can model crisp as well as vague regions in an intuitive user-friendly way[3].

[3]Because the approach combines a light-weight and interoperable representation scheme with a simple but effective reasoning algorithm it is well suited to be integrated in applications build for distributed and heterogeneous environments. Examples are the Semantic Web, service-based network infrastructures, peer-to-peer networks, as well as mobile and location-based services.

This work is organized as follows: In Chapter 2, we set the stage by defining key-concepts like geographic space, geographic objects, and geo-referenced information. This is followed by a review of typical general-purpose and spatial metadata standards and the methods used in these standard to geo-reference information objects.

In Chapter 3, we look at place names and their importance as natural language identifiers of geographic objects. We discuss different types of place names and problems related to the often observed vagueness of place names. We introduce gazetteers, their basic components, and their role as tools to manage place names and to support spatial queries in information retrieval systems. We focus in particular of the importance of spatial footprint representations within the context of gazetteers.

In Chapter 4, we discuss the general concept of spatial relevance and try to identify the spatial relations that are important in this context. We continue with a brief review and comparison of both quantitative and qualitative methods to evaluate spatial relations. The main purpose of this section is to provide an overview of the most important qualitative methods in this area. We will also look at sources and types of spatial vagueness and at some of the available qualitative methods to handle spatial vagueness.

Chapter 5 introduces a framework to represent spatial regions in discrete space. We will describe spatial tessellations, and in particular "natural" polygonal tessellations, as discrete spaces that can be used for the representation of spatial regions. We will show how multi-resolution qualitative spatial reference models can be built using graph-based abstractions of standard polygonal reference tessellations.

In Chapter 6, we show how named geographic objects with both a crisp or a vague regional extent can be approximated on the basis of the qualitative spatial reference models introduced in Chapter 5. We will outline the organization of place names into place name structures and how these structures can be used as user-friendly and intuitive tools to model geographic space.

In Chapter 7, a simple metric for the computation of the relative spatial relevance between units of a qualitative spatial reference model will be described. This metric will be applied as well to compute the spatial relevance of approximated (place name) regions in place name structures.

In Chapter 8, we will present an evaluation of our approach using a real-world data collection. We will also describe the implementation of the reasoning tool in BUSTER, an information broker middleware ([Visser et al., 2002, Vögele et al., 2003a]), and outline an application of our approach for the development of machine-readable indices of geospatial data.

Finally, we conclude in Chapter 9 with a summary of our work and an outlook to future research topics and applications.

Chapter 2

Geographic Objects and Spatial Data

2.1 Geographic Space

2.1.1 Spaces and Scales

The intention of this work is to support *spatial queries* in information retrieval applications. To achieve this goal, we want to develop a representation scheme and appropriate reasoning mechanisms that allow for a ranking of information objects according to their *spatial relevance*. To do so, we first have to define the semantics of the terms *"spatial"* and *"space"*.

Depending on the given application context, the semantics of *"space"* can vary considerably (for an overview, see [Freundschuh and Egenhofer, 1997]): There is *table-top space*, which is an example for a typical *small-scale space* and relates to the space at which many laboratory experiments are conducted. On the other hand, there are *large-scale spaces* which an observer cannot overlook from a single view but only by moving through the space. Geographic space certainly qualifies as a large-scale space, but so does the space inside a building.

In the context of information retrieval, *"space"* refers to the set of locational references which individual information objects are linked to. In most cases, these locations can be found within the natural environment of their human creators, i.e., they refer to a space at geographic dimensions. In the context of information retrieval we therefore use the terms *"space"* and *"geographic space"* as synonyms.

Although we think to have an intuitive knowledge of what he term *"geographic space"* relates to, a precise definition proves to be rather difficult. First and foremost, the decision whether a space can be classified as *"geographic"* seems to be a matter of scale: While it is obvious that very small entities on an microscopic level and very large things on a cosmic, macroscopic level can be excluded, it is not easy to decide which of the scales in between belong to a *geographic scale*.

Most authors agree with Egenhofer [Egenhofer and Mark, 1995] who defines geographic scale as *"beyond the scale of the human body"*. To make this rather

vague definition more precise, we can say that on one hand, the geographic scale must be definitely larger than the scale of the table-top-space mentioned above. On the other hand, the *"geo"*-graphic scale is limited by the dimensions of the earth. Following Smith [Smith and Mark, 1998], we can call this intermediate scale *"mesoscopic"*.

While this terminology uses spatial dimensions as the only classification criterion, other authors like Montello [Montello, 1993] developed a differentiated scale system based on the size of the human body in combination with human perception and movement through space. He distinguishes between *figural space* which pertains to sizes smaller than the human body and which can be perceived without a change of location by the observer (e.g. table-top-space, image space), *vista space* which is similar to figural space but extends to scales larger than the human body, *environmental space* that cannot be perceived without the observer moving from one location to another, and finally *geographic space* which cannot be properly understood by moving around in space, but only through indirect perception with the help of abstract representations, e.g. maps.

It is Smith's mesoscopic scale, or Montello's geographic space that is the domain of interest for this work. In particular, we are interested in the scales of towns, countries and continents, i.e. the scales that are typically visualized with the help of topographic and other maps.

2.1.2 Objects in Geographic Space

It is not easy to find a common-sense definition of constitutes a *geographic object*. Through empirical studies, Smith [Smith and Mark, 1998] found that the associations humans have with nouns such as *geographic feature*, *geographic object*, and *something that can be portrayed on a map* may vary considerably. While most contestants associated natural phenomena like mountains, rivers, and lakes with *geographic feature*, abstract objects with man-made boundaries, like administrative subdivisions, were seen as *something that can be portrayed on a map*. *Geographic objects*, on the other hand, were associated in most cases with physical objects that have some relation to geography, like a compass or a globe.

Nevertheless, we use the term *geographic object* in this work to denote a union of *geographic feature* and *something that can be portrayed on a map*, as defined by Smith. Following this definition, three general types of geographical objects can be distinguished:

- Physical objects such as forests, rivers, and bridges. These objects have in common that they are of a mesoscopic, geographic scale and do have properties that can be studied with the methods of physics. These objects can be both of natural origin (e.g. forests), or created by man (e.g. bridges).

- Geographic objects that are physical, but whose identity is at least partially defined through human cognition and action. Examples are objects like a bay, or a promontory.

- Geopolitical objects like nations or neighborhoods. Their identity is completely defined through human cognition and action.

All three types of geographic objects have in common that they can be abstracted as 2-dimensional *regions*, i.e. their identity can be determined by the specification of the location of their boundaries in the 2-D Cartesian Plane. Here, Smith distinguishes between *bona fide* and *fiat* boundaries [Smith, 1995], [Smith and Mark, 1998]. The former are boundaries that refer to real discontinuities in the physical world. Examples are the shoreline of a lake, or the edge of a parking lot. The latter are projected onto geographic space and are, to a certain degree, independent of physical boundaries. Typical fiat boundaries are postal code areas, or the administrative subdivisions of a nation.

The last example illustrates the inherent problem of this definition, which is to make a clear distinction between the different boundary types: Because many administrative boundaries follow natural geographic features (e.g., the boundary between Germany and France follows the Rhine river), *"artificial"* fiat boundaries may be coincident with the boundaries of *"real"* physical features. And on the other hand, the bona fide boundaries may at least partly be determined by anthropogenic cognitive processes. A classic example is the boundary of a mountain, which depends very much on the (arbitrary) definition of where the valley ends and where the mountain begins.

2.1.3 Representations of Geographic Space

Generally, geographic space may be conceptualized either as a set of locations with properties (referred to as *absolute space*), or as a set of objects with spatial properties (referred to as *relative space*) [Worboys, 1995]. Depending on which view is chosen, this dichotomy has implications on how geographic space is represented: Absolute space is modelled using continuous fields that are mapped onto discrete partitions of space to be able to handle them in computerized systems. Relative space is modelled using geographically referenced objects. In the domain of digital geospatial data, both the field and the object approach are used and generally referred to as *raster* and *vector* data formats.

Depending on the view chosen, we either have to deal with terrestrial locations over which geographic phenomena take place, or with geographic entities that are referenced to locations on the earth and taking part in geographic processes. Whether the field or the object approach should be used, i.e. whether we model geographic space as a *"jig-saw puzzle of polygons, or a club-sandwich of data layers"* [Couclelis, 1992], depends on the purpose of the model: If the model is used for an analysis of (natural) spatial processes, like rainfall patterns or landforms, the field view may be used. If the focus is more on the management of geographic objects in an administrative context (e.g. to outline planning zones and information about land parcels), the object view seems more appropriate [Burrough and McDonnell, 1998]. In information retrieval applications, we are generally more interested in geographic objects than in processes. Accordingly, there is a bias in this field towards the object view of geographic space and representations using vector data formats.

Both the object- and the field-based data models are typically displayed and visualized with the help of maps. Maps have a long tradition as tools for the visualization of geographic space, and they are still the most common tools for this purpose. The history of printed maps and the associated science of cartography dates back to pre-roman times. However, throughout the centuries, there has been a significant cartographic evolution [Hirtle, 2000]. While early

maps were often drawn from a planar perspective and included 3-dimensional images, they started only recently to adopt the areal perspective typical to modern maps. Modern maps are 2-dimensional projections of the 3-dimensional world onto a cartesian plane[1].

Within the last decades, digital maps have begun to replace the traditional, paper-based maps. Digital maps are managed with the help of computer-based *Geographic Information Systems (GIS)*. These systems use raster or vector based data models to represent geographic space and objects in geographic space. The data are stored digitally in data files or so-called *geodatabases*. In many cases, digital *geospatial* data are the basis for the online display and manipulation of digital maps, as well as for the creation of printed maps.

One advantage of digital geospatial data is that they can be manipulated by algorithms from computational geometry. These can be used for a multitude of tasks, for example to infer topological spatial relations between geographic objects. A number of different *geometries* are available to represent 2-dimensional geographic space. A geometry is defined as *"a group of transformations of space under which their propositions remain true"* [Worboys, 1995]. Some of the most common geometries used in the context of geographic space are the Euclidean space, set-based geometries, topological geometries, network spaces, and metric spaces.

A common framework to model geo-spatial phenomena in GIS is the *Euclidean Space*. Mathematically, the *Euclidean n-space* $\mathbb{R}^n$, sometimes called *Cartesian space* or simply *n-space*, is the space of all n-tuples of real numbers, $(x_1, x_2, ..., x_n)$. In spatial applications, the probably most frequently used Euclidean space is the *2-dimensional Euclidean plane* $\mathbb{R}^2$, which is also often called the *Cartesian plane* (Figure 2.1). This coordinatized space is defined through a point O, called the *origin*, and a pair of orthogonal *axes* that intersect at the origin. Each point in the plane is associated with a unique pair of real numbers. These x and y coordinates measure the distance from the origin in the direction of each axis. An alternative way to represent points in $\mathbb{R}$ is to write them as vectors that are defined by their direction and magnitude as measured from the origin.

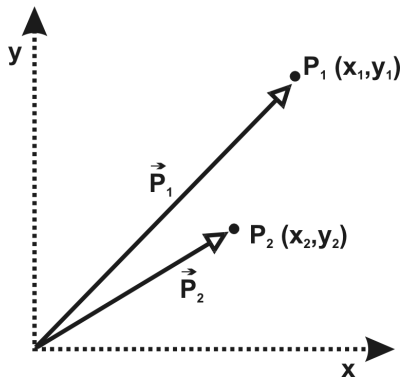

Figure 2.1: The Cartesian plane

[1]Driven by the development of new visualization devices, there is a new trend towards representations that use 3-D objects.

The Cartesian plane offers a number of possibilities to represent geographic objects. The simplest representation for an object is a pair of x and y coordinates. A more complex representation is given by a *line* object. Using vector notation, we can define a *line* through two points P_1 and P_2 in $\mathbb{R}^2$ as the point set $\{\lambda P_1 + (1-\lambda)P_2 | \lambda \in \mathbb{R}^2\}$. A *half line* that radiates from P_2, passing through P_1 is defined as the point set $\{\lambda P_1 + (1-\lambda)P_2 | \lambda \geq 0\}$. A *line segment* between P_1 and P_2 is defined as the point set $\{\lambda P_1 + (1-\lambda)P_2 | \lambda \in [0,1]\}$. A set of line segments can be linked together to form a *polyline*. A polyline consists of a set of *vertices*, and the *edges* between them. The area enclosed by a closed polyline is called a *polygon*. Polygons are frequently used for the representation of regional geographic objects. Together with point and line objects, they form the most common data types in GIS.

Computational geometry offers a number of simple yet effective algorithms to compute spatial relations between objects that are represented as such data types (e.g., [de Berg et al., 2000]). For point objects, only metric (distances) and ordinal (bearings) relations can be inferred. Simple topological relations (intersection) are added in the case of line objects. In addition to metric and ordinal relations, polygonal representations support the computation of complex topologic and mereotopologic relations, such as connection, intersection, overlap, and partonomy.

When we use Cartesian coordinates to reference geographic objects, our frame of reference is the underlying *geographic coordinate system*. For historical and practical reasons, there exists a large number of different geographic coordinate systems. To discuss them in detail is out of the scope of this work. However, it is important to know that each system is optimized for a specific geographic region, or for a specific task. Optimization for geographic regions is needed because, due to the spherical character of the globe, it is impossible to project the surface of the sphere onto a 2-dimensional plane without introducing errors. The projection will either be distorted and yield erroneous angles, or the distances between points will be wrong.

Which system is used as a standard in a specific country depends on historical and political reasons as well as on the useability of a particular coordinate system for a specific region. Many nations (among them Australia, Belgium, Germany, Great Britain, Finland, Ireland, Italy, The Netherlands, New Zealand, and Sweden) have defined *national grid systems* based on coordinate systems that cover their territory. In most cases, the national coordinate system (or set of coordinate systems) of a given country is set up in such a way that, for the specific sector of the globe that is covered by the country, it offers true angles (this is important for navigation) while keeping the distance errors at a minimum. In Germany, the *Gauss-Krüger* coordinate system is the standard geographic coordinate system since 1927. In the UK, the *The British National Grid (BNG)*, which is based on the *National Grid System of England* and administered by the *British Ordnance Survey*, is used. Many geospatial data in the United States are based on *Universal Transverse Mercator (UTM)* coordinates.

A global coordinate system is the system of geographic latitude and longitude coordinates. Latitude/longitude coordinates are a direct projection onto the spherical globe (or, because the Earth is not a perfect sphere, a reference ellipsoid), and therefore do not provide true angles. For this reason, they are not suitable for many applications. In addition, although geographic coordinates provide a global coverage, there are differences between systems depending on

the choice of reference ellipsoid. For example, the exact location of a geographic object in central Europe may differ by up to 100 m, depending on whether the geographic coordinates refer to the Bessel or a Clarke ellipsoid.

With respect to data interoperability, the use of different coordinate systems means that coordinate transformations are necessary to integrate geospatial data that cover the same geographical region. This may seem not very problematic from a technical point of view. In fact, many GIS tools already provide facilities for automatic coordinate transformations, and future spatial data infrastructures (SDIs) will incorporate specialized coordinate-transformation services [OGC, 2001c].

However, even the most comfortable tools and services have to be correctly parameterized. The parameterization requires either suitable metadata, or GIS-related knowledge on the side of the users, neither of which is available in many cases. For this reason we argue that the complexity of coordinate-based geographic representations can be an obstacle to the development of intuitive and user-friendly spatial models based on interoperable geodata.

2.1.4 Geo-Referenced Information

In the domain of information retrieval, some authors refer to atomic data units as *information objects* (e.g., [van der Weide, 2001]), others use the term *information item* (e.g., [Arms, 2000]). In this work, we use the term *information object*. Depending on the application context, information objects can be web pages, images, books, documents, database records, digital maps, and other structured or unstructured data. Information objects are organized in collections. Examples of such collections are libraries and large archives. But also individual documents like telephone books or technical manuals can be referred to as a collection. The probably largest collection at the moment is the distributed collection offered by the World Wide Web. This collection holds a large number of distributed, heterogeneous, and mostly unstructured information objects.

Information objects that are related to a location in geographic space are called *geo-referenced*. In principal, there are two ways to geo-reference an information object [Goodchild, 1999]:

Direct geo-referencing: An information object may be referenced *directly* to a location in geographic space with the help of exact, *quantitative* metadata. Some authors (e.g., [Goodchild, 1999]) limit these metadata to geographic coordinates. However, references on the basis of other quantitative reference systems (e.g., street addresses) may be used for direct geo-referencing as well.

Indirect geo-referencing: An information object may be geo-referenced *indirectly* by associating it with a geographic object. Named geographic objects are typically abstracted as *place names*. Place names are natural language expressions used to identify and disambiguate geographic objects. For applications where the geographic location of a place name can be inferred by the (human) user, place names do not have to be geo-referenced themselves. However, for automated applications (which are the focus of this work), a link between a place name and geographic space

has to be established. Below, we will discuss a number of options to establish quantitative and qualitative links between place names and geographic space.

In practical applications, both direct and indirect methods are used to geo-reference information objects. In general, the majority of information objects that can be described as *"geospatial"* data are directly geo-referenced. These include a wide range of digital cartographic products, survey data, satellite images, aerial photographs, data from ground-based and atmospheric monitoring stations, and other digital maps and data.

On the other hand, the majority of un-structured and semi-structured information objects stored in (distributed) digital collections like the World Wide Web and digital libraries are indirectly geo-referenced. In addition, most of the non-digital information objects in collections like library catalogs or bibliographic indices are indirectly geo-referenced as well.

While per definition all geospatial data are geo-referenced, it is difficult to make a judgement about what percentage of information objects in a distributed collection like the World Wide Web do have a (indirectly or directly established) relation to geographic space. In general, we can distinguish between information objects for which *spatial metadata* are available, and information objects that lack such metadata. Information objects that are part of structured collections typically fall under the first category. Here, an information object is given an explicit, direct or indirect geo-reference, often based on a well-defined coordinate system or vocabulary of place names. [2]

The geographic reference of the second category of information objects, i.e. information objects that lack spatial metadata, is much more difficult to evaluate. Nevertheless, this type of information objects cannot be neglected because it forms the bulk of information available through the currently largest distributed collection of information objects, the World Wide Web.

Depending on the definition of the term *geo-referenced*, and the method used to extract this geo-reference from an information object, the estimates of the percentage of information objects in the World Wide Web that are geo-referenced vary considerably. In experiments conducted at the *IBM Almaden Research Center*, McCurley [McCurley, 2001] found that only about 10% of the analyzed web content had a relation to geographic space. However, he based his analysis on the existence of *recognizable* US ZIP code and address information in a web page and acknowledged that, if other indicators for a geographic reference are used, this percentage may be much higher.

If we classify web pages as *geo-referenced* as soon as the name of a geographic object (i.e., a place name) is recognizable on the page, the percentage of geo-referenced web-pages is considerably higher. To get a rough estimate of the percentage of web-pages that are geo-referenced we conducted a simple experiment: We used *Google* search engine [GOOGLE, 2004] to retrieve web-pages that were relevant with respect to different keywords. We then manually analyzed the first 10 hits in the respective result sets for the occurrence of place names. The result of this little experiment was, that indeed the percentage of pages that contained one or more place names could reach 80% and more. However, we found that there is a correlation between the thematic concept we were

[2] In section 2.2, we will describe spatial metadata and reference systems in more detail.

looking for and the frequency of place name in the respective web pages: The keyword *"water"*, for instance, produced a result set with approximately 80% of geo-referenced pages. The keyword *"computer"*, on the other hand, retrieved had much less pages with a place name (about 60%).

2.2 Spatial Metadata

A number of different definitions for metadata exist: Some authors refer to metadata as *"data about data that describe the data on the semantic, structural, statistical and physical level"* [Lenz, 1994]. Others simply state that metadata provide *"information which makes data useful"* [Bretherton and Singley, 1994]. In any case, metadata have a long tradition in the management of information collections. It were librarians who first developed elaborate indexing systems to manage collections of books and other documents. In a typical library catalog, each information object is represented by an index card holding relevant information about the object (e.g., title, author, publisher, a short summary of the content).

With respect to an index card in a card catalog providing information about a book, it seems to be rather straightforward to decide which one is the metadata (the card) and which one is the data (the book). But how about the *table of contents* of the book? It obviously is part of the book, and as such part of the data. But it also provides information about the contents of the book, so it could be referred to as metadata. As we can see, it is often not easy to draw a sharp line between data and metadata.

One way to solve this confusion is to distinguish two classes of metadata:

- document-related metadata, and
- content-related metadata.

Document-related metadata comprise most of the information that can be found on the index cards of typical library catalogs. They provide *"technical"* information about a document, like the name of the author, the date and location of publication, and the name of the publishing company. Content-related metadata, on the other hand, describe the concepts and topics that are addressed by an information object. In most library catalogs, content-related metadata are limited to a set of keywords or a short abstract of a document.

The boundary between data and their content-related metadata is not clear-cut, but rather a matter of *abstraction* and *generalization* [Vckovski, 1998]. Content-related metadata are simply data at a higher level of abstraction and, depending on the intended use (e.g., production and quality control, data exchange, or data directories), they are a more or less abstracted version of the actual data. Using the example of a book, metadata at increasing levels of abstraction refer to the whole text (i.e., no abstraction at all), an abstract of the book, the book's table of contents, and finally a highly abstracted version given through relevant keywords in a library catalog record.

In the context of this work we use the term *"spatial metadata"* to denote content-related metadata that describe the relation between the content of an information object and geographic space. For example, to properly reference a report about the environmental damage caused by the 2002 floods of the

Elbe river near *Dresden*, the content-related spatial metadata should contain a reference to the city of Dresden and the Elbe river, irrespective of the fact that the report was published by the German federal environmental agency, the *Umweltbundesamt (UBA)* in Berlin. [3]

Which metadata schema is appropriate to describe an information object depends very much on the type of the object, and the intentions of the creator of the information collection. Because each information collection has a very specific content, purpose, and potential user community, different metadata schemata are required. To enable the exchange of (meta)data and information objects between different information collections possible, the respective metadata schemata have to be made comparable. For this purpose, *metadata standards* were established.

The bulk of metadata standards are designed for information objects that do not necessarily qualify as geospatial data. However, there are a number of emerging standards specifically designed to be used for geospatial and other geo-referenced information objects. Both types of standards do provide options to geo-reference information objects.

In the following we analyze selected examples of general purpose metadata standards, as well as metadata standards that were specifically designed to reference geospatial data. The intention of this analysis is to determine which are the key elements offered by both types of standards to (directly and/or indirectly) geo-reference information objects.

2.2.1 "General Purpose" Metadata Standards

Probably the first metadata to be used on a large scale were the index cards introduced by librarians to reference books and other documents. With time, a number of different indexing systems evolved, most of them specifically designed for the library that used them. Over time, some initially library-specific standards developed that were accepted by a larger community of libraries. This development was driven by the use of digital library catalogs, and the increasing need to be able to interconnect these catalogs. A prominent example for an indexing system that evolved into a national metadata standard is the *MARC Standard* developed the *U.S. Library of Congress*, one of the largest libraries in the world.

The MARC Standard

To organize its vast data holdings, the Library of Congress began in the 1960s to develop *MARC*, a storage and classification system that provides a human-readable and machine-searchable index of catalogued documents [Furrie, 2000]. In its current version (MARC21), the system is available in most US libraries through national networks and plays an important role in the management of libraries in the US.

A MARC catalog record contains basically the same information as it's non-digital counterpart. Like the traditional catalog card, a typical MARC record includes at least

[3] However, a reference to the *UBA* should be part of the document-related metadata of the report.

- the names of the author, joint author, editor, or illustrator,
- the document title,
- information about edition, publication, and series,
- a physical description of the document,
- a summary of the document-content,
- classification and call numbers (i.e. LLCN, ISBN), and
- information about the *"topical subject"*.

.

Although the Library owns almost 4.5 million maps, atlases, and reference works, of which at least the maps acquired since 1969 have been cataloged and made available through MARC, the MARC data model offers only limited options to geo-reference information objects. References to geographic locations may appear in several of the MARC tags, for example to describe the geographic context of a topical subject of a book, or to specify the location of a conference in the description of conference proceedings. But only information objects that are indexed using a specific subset of the MARC tags (called the *USMARC Geographic Subject Subdivision*) can be given an explicit geographic reference. Using this subset, a MARC entry describing the subject of a real estate map covering *Tippah County* in Mississippi would look somewhat like

```
650 \#0\$aReal property\$zMississippi
    \$zTippah County\$vMaps
```

Here, the classifier *$z* refers to the *"geographic subdivision"* of tag *650*, which represents *"subject added entry - topical term"*. The Library of Congress Policy and Support Office maintains list of place names and county codes to be used in MARC records [LOC, 2002]. More complex representations, for example to outline the extent of river basins, congressional districts, or study areas, are not supported in MARC [Nebert and Fullton, 1995].

The Dublin Core

A general problem of all indexing systems is the effort involved in the creation and maintenance of the metadata records. For information objects like books and documents that are organized in libraries, a comprehensive metadata description (e.g. using the MARC standard) is typically provided by an authorized organization (e.g., the Library of Congress) and distributed with the information object. Outside the world of libraries, however, the creation and distribution of metadata is less standardized. Because many organizations do have only limited resources available for documenting their information collections, the creation of comprehensive and complex metadata descriptions is often a problem. This can result in incomplete or outdated metadata that ultimately defeat the whole purpose of metadata documentation.

To address this problem, standards for simplified *discovery metadata* have been developed. The purpose of discovery metadata is to enable information seekers to find what they need as fast as possible, while at the same time enabling

information producers and intermediaries to sustain the effort of keeping up-to-date metadata records.

An important initiative in this direction is the *Dublin Core Metadata Initiative (DCMI)*. The result of this initiative is the so-called *Dublin Core (DC)* [DCMI, 2000b]. Although it is usually referred to as a *"standard"*, the Dublin Core is not a binding specification in the sense of an ISO standard, but rather a consensus-based and open international recommendation. Like the MARC standard, the origins of the Dublin Core Initiative and the Dublin Core are rooted in the library and publishing communities. Initially designed to provide a *"core"* set of metadata for data discovery purposes in libraries, the DC has been adopted by many other disciplines and developed into something like a *"general-purpose"* metadata standard. The Dublin Core has been endorsed at a European level by the *Comité European de Normalization (CEN)*, through a Workshop Agreement, i.e. a voluntary mechanism involving a wide range of partners from different sectors [CEN, 2004].

The DC does not replace domain-, or sector-specific standards such as the *ISO 19115* standard described below. It can be viewed as complementary to such standards and helps to discover information resources across disciplinary and sectoral domains. The design of Dublin Core had to balance the needs for simplicity in describing digital resources with the need for precise retrieval. It also had to offer a compromise between thorough, but expensive manual annotation, and error prone but cheap automatic indication. As a consequence, the Dublin Core standard comprises only fifteen elements, the semantics of which were established through consensus by an international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields.

As a simple metadata schema that can be easily be understood and handled, the DC is not necessarily suited to express complex relationships or concepts. However, although it was initially designed for documents and document-like objects, the Dublin Core developers recognized the importance of providing a mechanism for extending the DC element set for additional resource discovery needs. They added mechanisms that allow to extend the standard with different, domain-specific metadata schemata. This model of a fixed set of core metadata and domain specific extensions enables different user communities to use the DC elements for core descriptive information, while it supports domain specific additions as needed. It is the extensibility of the DC that is one of the main reasons for the success of the standard [Hillmann, 2001].

The Dublin Core features two classes of terms, *elements* (nouns) and *qualifiers* (adjectives), which can be arranged into simple statements. The resources themselves are the implied subjects in this language. Each element in the Dublin Core element set is optional and may be repeated. Each element also has a limited set of qualifiers, i.e. attributes that may be used to further refine (not extend) the meaning of the element. The Dublin Core Metadata Initiative has defined standard options to *"qualify"* elements with various types of qualifiers. A set of recommended qualifiers conforming to DCMI *"best practice guidelines"* is available, with a formal registry in process.

The Dublin Core standard was not specifically designed for geospatial or georeferenced data. Accordingly, among the 15 basic elements of the DC, only the elements *coverage* and *spatial coverage* refer to the geographic reference of a data object. *Coverage* describes the extent or *scope* of the content of the resource.

The semantics of the *coverage* element are not exclusively spatial. It may refer to coverage in the sense of a spatial extent, but also to a temporal coverage (i.e., a point or period in time), or a jurisdiction (i.e., a named administrative entity). The element *spatial coverage*, on the other hand, has an explicit spatial connotation. It can be used to geo-reference an information object in terms of a place name as well as by geographic coordinates.

In order to provide fillers with a well-defined semantics for the *spatial coverage* element, the DCMI recommends to use well-defined spatial representations, or to select a value from a controlled vocabulary of place names. The DCMI guidelines list the following specifications and controlled vocabularies for this purpose:

DCMI Point: The DCMI Point specification identifies a point in space using its geographic latitude/longitude coordinates [Cox, 2000b].

DCMI Box: The DCMI Box specification identifies a region in space through a bounding box specified in geographic latitude/longitude coordinates [Cox, 2000a].

ISO 3166: The ISO 3166 codes provide a controlled vocabulary of country references [ISO, 1997].

TGN: The Getty Thesaurus of Geographic Names (TGN) is an extensive controlled vocabulary of (international) place names [TGN, 2002].

2.2.2 Metadata Standards for Geospatial Data

"General-purpose" metadata standards like *MARC* or the *Dublin Core* can be used to manage geo-referenced data to a certain extent. Although they were designed primarily with books an other *"non-spatial"* data in mind, they do offer options to geo-reference information objects. These geo-references can be expressed both indirectly through place names, and directly by using geographic coordinates to specify simple spatial representations like points or bounding boxes.

Metadata that are used for the management of *geospatial* data, however, have to provide additional information. They should be able to answer the following questions [Tschangho, 1999]:

Availability: Which data sets exists for a given thematic *topic* and a given *geographic region*?

Fitness for use: Does the data set meet a specific *need*?

Access: How can the data set be *accessed*?

Transfer: How can the data set be *processed* and used?

To answer these questions for geospatial data (e.g., cartographic and remote sensing products), detailed information about applicable coordinate systems, map projections, and other parameters that are important in this context have to be provided in addition to the usual *"standard"* metadata elements. Obviously, information about the *spatial coverage* of a geospatial data set should be an integral part of each metadata description.

Efforts are under way within the environmental and geo-science community to develop metadata standards that fulfill these requirements. The standardization process is driven by a number of national and international organization and the absence of a generally accepted international standard has led to the development of several parallel standardization projects[4].

In the United States, the U.S. *Federal Geographic Data Committee (FGDC)* developed the *Content Standard for Digital Geospatial Metadata (CSDGM)* as part of the *National Spatial Data Infrastructure* [FGDC, 1994]. The CSDGM is a comprehensive metadata standard for geospatial data. It is mandatory for all federal agencies in the United States.

In Europe, the *Comité European de Normalization (CEN)*, and within CEN, the *Technical Committee 287 (CEN/TC 287)*, took a lead in the development of a European standard for geographic information. The efforts of this committee resulted in *ENV 12657*, a voluntary pre-standard released in 1998 [CEN, 1998]. CEN/TC 287 is now *"dormant"*, and ENV 12657 was integrated in the development of the *ISO 19100* series of standards. Under the general topic *"ICS 35.240.70 - IT Applications In Science"* [ISO, 2004], these standards cover most aspects of spatial data management, including reference models for geospatial data (*ISO 19101*), a standard for conformance and testing (*ISO 19105*), spatial and temporal schemata for geographic information (*ISO 19107* and *ISO 19108*), standards for spatial referencing by coordinates and spatial referencing by geographic identifiers (*ISO 19111* and *ISO 19112*), quality principles and quality evaluation procedures (*ISO 19113* and *ISO 19114*), functional standards for geospatial data (*ISO/TR 19120*), a standard for imagery and gridded data (*ISO/TR 19121*), and finally *ISO 19115*, a standard for spatial metadata [ISO, 2003c].

Although some of it's components it were released only recently, the *ISO 19100* standard series is already used worldwide as a basis for a number of regional standards and applications. In Australia and New Zealand, for example, the emerging *Australian Spatial Data Infrastructure (ASDI)* developed by *ANZLIC* [ANZLIC, 2004] follows in many aspects the *ISO 19115* standard. Likewise, the *Food and Agriculture Organization of the United Nations (FAO)* [FAO, 2004] uses a sub-set of *ISO 19115* for its spatial metadata.

In the following, we will take a closer look at the *CSDGM* and *ISO 19115* to see what options they offer to geo-reference information objects.

The FGDC Content Standard

In the United States, the Federal Government has been working on the creation of a *National Spatial Data Infrastructure (NSDI)* since the early 1980s. The intent of the *NSDI* is to enable and improve the exchange of spatial data between a number of federal government agencies such as the *United States Geological Survery (USGS)*, the *Defense Mapping Agency (DMA)*, and the *National Ocean Service* [Günther, 1998]. To coordinate the development of the *NSDI*, a working group was formed that includes representatives of all major government agencies involved in the handling and management of geospatial data. In 1994, this working group (the *Federal Geographic Data Committee (FGDC)*) published a

[4]However, there seems to be agreement among the experts on the need for more cooperation. Efforts are under way to integrate the different standards into one general metadata standard for geospatial data.

draft for a *Content Standard for Digital Geospatial Metadata (CSDGM)*, also referred to as the *FGDC Content Standard*, or simply the *Content Standard* [FGDC, 1994]. An improved version of this standard was endorsed in 1998 with the addition of guidelines for the development of profiles and user-defined metadata entities and elements.

The CSDGM specifies the structure and content of some 220 metadata elements. It structures a metadata set into seven groups: identification information, data quality information, spatial data organization information, spatial reference information, entity and attribute information, distribution information, and metadata reference information. Only the first and the last group are mandatory.

The geo-reference of an information object is specified in the first group, the *"identification information"*. Here, a spatial reference can be implemented using either a *bounding box* or a more detailed *polygonal description*.

In the same group, the *"geographic locations characterized by the data set"* can be specified by a place name using a *place keyword* that may be taken from a *place keyword thesaurus*. The place keyword is attributed with type information, a reference to the keyword thesaurus, and a short name.

Note that group 4, which is called *"spatial reference information"*, refers exclusively to the description of the spatial properties of the data set, such as which projections and coordinate systems were used.

The *ISO 19100* Series and *ISO 19115*

The *ISO 19100* series is a multi-part international standard for geographic information intended to *"provide a foundation for the development of technological implementations in the field of geospatial data management"* [ISO, 2004]. The series was developed by the *International Organization for Standardization (ISO)*, a *"world-wide federation of national standardization bodies which is responsible for promoting the development of standards to facilitate the international exchange of goods and services"* [ISO, 2004]. Within the ISO, *Technical Committee 211 (ISO/TC 211)* played a key role in the development of the *ISO 19100* series of standards.

Among the standards of the *ISO 19100* series *ISO 19115 Geographic Information – Metadata* is the one that provides detailed specifications for spatial metadata. In February 2001, *ISO 19115* was approved for publication as a *Draft International Standard (DIS)*. In August 2003, the final version of the standard was published [ISO, 2003c].

To develop this standard, *ISO/TC 211* built on the experience of the *FGDC*, the *CEN/TC 287* (which had developed a pre-standard on GI metadata in 1997), and standardization activities in Australia, New Zealand, and Canada. All together, 33 countries and 12 observer organizations were involved in the development of the standard. Consequently, the new standard already has a wide international acceptance. It is expected that many organizations will modify their own geospatial metadata standards to become interoperable with *ISO 19115*. Version 3 of the FGDC Content Standard, for example, will become a profile of *ISO 19115*. Similarly, it is expected that CEN will endorse *ISO 19115*, thus making this standard a point of reference for all European organizations. The *OpenGIS Consortium (OGC)* [OGC, 2004], an international consortium of businesses, governments and universities that plays an important role in the

development of specifications for interoperable geodata, works in close cooperation with ISO/TC 211. The *OGC* itself does not develop metadata standards, but the respective *OGC Abstract Specifications* [OGC, 2001a] explicitly refer to *ISO 19115* as the standard of choice.

In the field of geographic and geospatial information, *ISO 19115* has therefore the potential to become *"the"* internationally accepted metadata standard that integrates most existing national and international metadata standards and specifications.

ISO 19115 provides specifications for the description of digital geographic datasets using a comprehensive set of 420 metadata elements. These elements are designed to provide information to cover the four use-categories mentioned at the beginning of this section, namely regarding the availability of an information object, its useability for a specific task, a description of how to access the information object, and specifications for how to use it.

To provide this information, the standard specifies metadata elements with which the identification, extent, quality, spatial and temporal schema, spatial reference and distribution of digital geospatial data can be described. The standard can be used to catalogue geospatial data at different levels of granularity. In *ISO 19115*, a referenced information object can be a collection of datasets, an individual dataset, an individual geographic feature, or a specific property of a geographic feature.

There are a number of metadata elements and logic element groups in *ISO 19115* that are specific to geospatial data. These include elements that describe the general *type of spatial reference method* used (i.e. vector or raster), the *scale* of a data set, information about *map projections* and *reference systems*, information about *data quality* and *lineage*, as well as detailed information about *spatial object types*. Through a references to *browse file graphics* small static thumbnail images can be displayed with the metadata to give a quick impression of the contents of the data set.

With respect to the specification of a *geographic reference* for an information object, *ISO 19115* offers two options:

- The metadata elements *geograpicDescription* , *geographicIdentifier* and *administrativeArea* can be used to describe the geo-reference of an information object in terms of natural language identifiers. The element *geographicIdentifier* can be filled with *place names*. However, *ISO 19115* does not prescribe a specific controlled vocabulary from which these place names should be taken. The element *administrativeArea* refers specifically to administrative subdivisions.

- The elements *polygon* and *geographicBoundingBox* (with the four subelements *northBoundLatitude*, *southBoundLatitude*, *eastBoundLongitude*, and *westBoundLongitude*) allow to approximate the areal extent of an information object reference in terms of a bounding polygon, or a bounding box. The coordinates that can be entered here have to be geographic latitude/longitude coordinates expressed in decimal degrees.

.

Spatial References in Metadata Standards

In summary we can say that with respect to the geo-referencing of information objects, metadata standards explicitly designed to represent *"geospatial"* data (e.g., the FDGC-CSDGM and *ISO 19115*) do provide essentially the same options as *"general-purpose"* standards like the *Dublin Core* or *MARC*: An information object may be directly geo-referenced with *geographic coordinates*, using more or less simplified spatial objects (points, bounding boxes, polygons). Or it may be indirectly geo-referenced by *geographic identifiers*, i.e. place names.

This shows that the ability to resolve indirect geo-references given by place names is not only important for information retrieval from collections that do not provide metadata descriptions (e.g., the World Wide Web as well as many databases and un-structured document repositories). It also plays a role for collections that organize information objects with the help of metadata annotations (e.g., (digital) libraries and data catalogs).

The consistent use of place names in a wide range of information collections supports our assumption that using geographic objects represented by place names to geo-reference information objects follows the human intuition and cognitive modelling of geographic space. In this context it is interesting to note that for practical applications, even some collections of geospatial data prefer the use of place names over the use of coordinates for geo-referencing. The *United Nations Food and Agriculture Organization (FAO)*, for example, uses a data catalog based on a sub-set of *ISO 19115* in which place names are the only allowed option to geo-reference information objects [FAO, 2004].

In the following section, we will examine the concept of place names more thoroughly. We will define what place names are, how they are managed, and how they can be used in the context of spatial information retrieval.

Chapter 3

Geo-Referencing with Place Names

3.1 Geographic Objects and Place Names

3.1.1 Named Geographic Objects

The metadata standards described in section 2.2 offer options to indirectly geo-reference information objects through references to geographic objects denoted by *place names*. This is true for both *"general-purpose"* standards like the *Dublin Core*, as well as for standards specifically targeted at the management of geospatial data (e.g. *ISO 19115*).

Place names are used for geo-referencing because they are perceived as an intuitive and *"natural"* method to describe locations and objects in geographic space. Outside the relatively small community of cartographers, geographers, planners and other (geo)-scientists, geographic coordinates are often seen as being too technical and too complicated. In fact, geo-referencing with coordinates does only make sense in conjunction with the usage of (digital or non-digital) maps and the respective (technical) user interfaces. The task to determine the geo-reference of an information object in terms of exact geographic coordinates is not a trivial task. Geographic Positioning Systems (GPS) are helpful only when the location of the person annotating the metadata and the geographical reference of the data sets coincide.

But even when a GPS can be used, one of the major shortcomings of exact coordinate positions is that they do not provide any information about the spatial context of a location: A coordinate tuple $(3490600, 5885450)$ (in *Gauss-Krüger* coordinates) does by itself not convey any information about which spatial object(s) it refers to. If we were to use a GIS to zoom in on these coordinates, we may find that they refer to a location in *Europe*, or (more specifically) in *Germany*, or (more specifically) in the city of *Bremen*, or (more specifically) in the *Horn-Lehe* district, or (more specifically) in the *TZI* building. Which of these geographic objects, and which level of spatial granularity (i.e., scale) is actually relevant does not become obvious through the coordinate itself but has to be explicitly stated (i.e., *"the Horn-Lehe district with centroid coordinates of* $(3490600, 5885450)$*"*). Place names, on the other hand, refer directly to ge-

ographic objects and are therefore equipped with intrinsic semantics. From a statement like *"the document refers to the Horn-Lehe district"* we can not only deduce the spatial region the information object refers to, but also the level of spatial granularity at which this reference (and the information object) is relevant.

In general, because it is often easier to annotate a data set with a meaningful place name than with a set of geographic coordinates, place names play an important role for geocoding in many areas of information management. Apart from digital and non-digital libraries and distributed information systems, place names are used as geographic references in virtually all information services, from newspapers to news clips on television. As a consequence, effective systems for spatial information retrieval have to be able to process geographic references through place names in addition to geographic coordinates.

3.1.2 Standardized and Non-Standardized Place Names

Place names, or *geographic names*, are natural language identifiers for geographic objects. Because the cognitive perception of and navigation through our terrestrial environment and its features is closely linked to the perception of geographic objects, the names of these objects play a crucial role for the purpose of *"perception, recognition, distinction and communication"* about our geographical environment [UNGEGN, 2001]. But place names are more than simply natural language identifiers for objects and locations in geographic space. The set of place names of an entire country provide a window into the history and characteristics of this country [Helleland, 2002]. They reflect the migrations of peoples, their religious and cultural traditions, local languages, conquests and ancient fortifications, as well as the topography and industrial development of a place. Therefore, apart from their address function, place names do have a cognitive, emotive, an an ideological dimension as well.

With the help of place names, geographic features can be disambiguated and cognitive models of geographic space can be developed. If we want to describe the natural regions of a country, for example, we typically do this by using place names (e.g., *"the Harz mountains"*, *"the Lüneburger Heide"*, *"Northern-Germany"*). Likewise, if we want to describe the location of one geographic object with respect to another geographic object, we often use a combination of place names and qualitative spatial relations like *"Bremen is **in** Northern-Germany"*, or *"The town of Quedlinburg **is close to** the Harz Mountains"*.

All place names are equal in that they refer to objects or locations in geographic space. However, they differ with respect to the user communities they are used by. Based on empirical studies in Norway, Hellelund [Helleland, 2002] sets up a simple a three-tier structural hierarchy. He distinguishes between

1. place-names that the people on a farm or in other micro communities have in common,
2. place-names which are common to a whole rural district or a town, and
3. place-names which are common to the whole nation.

Hellelunds *micro-communities* are not confined to a rural environment, but can be found in complex urban societies as well. The inhabitants of a specific neighborhood, ethnic groups, or even the employees of a large firm can be

seen as such micro-communities that tend to develop their own, specific set of place names. Transferred to the realm of virtual communities like professional domains or user communities on the Internet, a potentially large number of micro-communities using their own vocabulary of place names exist.

Based on their different characteristics and user groups, we can distinguish between different types of place names (Figure 3.1). The major categories are *standardized* and *non-standardized* place names.

Standardized Place Names

The importance of geographic names has long been recognized by national and international organizations, and the *"consistent use of accurate place names"* is seen as *"an essential element of effective communication worldwide and supports socio-economic development, conservation and national infrastructure"* [UNGEGN, 2001]. Because of their importance for the verbal and written communication about geographic space, and because they are also part of the cultural heritage of a country [Helleland, 2002], efforts are under way in most countries to gather, preserve and standardize the national heritage of place names. These efforts are led by the national cartographic or mapping agencies, or other institutions that were set up for the purpose. In Germany, the *Ständiger Ausschuss für geographische Namen (StAGN)* (Permanent Committee on Geographic Names) is responsible for the standardization of geographical names within the German linguistic area [BKG, 2002]. The *StAGN* is an independent scientific body without sovereign functions. Affiliated to the *StAGN* are scientists and practitioners from Germany, Austria and Switzerland and other German-speaking regions.

Like similar organizations in other counties, the *StAGN* cooperates with the *United Nations Group of Experts on Geographical Names (UNGEGN).* The *UNGEGN* is part of an initiative towards the standardization of geographical names that was launched by the United Nations after World War II. Through *UNGEGN*, the United Nations organize conferences on the standardization of geographical names that are held every five years. So far, there have been eight international UN conferences on this topic. A number of working groups within *UNGEGN* address specific topics pertaining to questions of toponymy (toponymic terminology, country names, romanization systems etc.), as well as the management and organization of place names in data files and (digital) gazetteers.

Different countries are at different stages of their national place name standardization programmes. While some countries do already have comprehensive listing of their national place names, others are still trying to gather the whole wealth of place names available. In the UK, for example, the British *Ordnance Survey* with it's more than 200 year old history is has been in charge of a comprehensive list of all place names in the United Kingdom for many decades. The Kingdom of Quatar, on the other hand, is still in a stage where the names of geographic objects and locations have to be gathered in field campaigns.

The standardization of place names, and the maintenance of official place name lists requires a considerable technical and administrative effort. Consequently only place names that show a certain persistence over time, and that are used by a sizeable group of a population (or an important ethnic minority) are typically included in standardized place name lists. In terms of Hellelund's

categorization mentioned above, standardized place names belong mainly to the category of place names that are common to a whole nation. Typical examples for standardized place names are named administrative subdivisions, or the names of long-living geographic features like mountains, rivers and lakes.

Non-Standardized Place Names

We call place names that are not included in the official standardization efforts *non-standardized*, or *ad-hoc place names.* Following Hellelund, non-standardized place names typically belong to the categories of place names that are used on a regional or town level, or by even smaller micro-communities. In a more general classification approach, we can use the two factors *place name lifetime* and *size of place name user community* for an approximate categorization of the different types of place names (Figure 3.1). Here, *place name lifetime* denotes the approximate time a place name stays in use. This can range from hundreds of years to only months. *Size of user community* refers to the approximate number of people that use a place name. This number may range from billions of people (global standardized place names) to 1 person (personalized place names). Within the category of non-standardized place names, we distinguish two main sub-categories, namely *local colloquial place names* and *ad-hoc place names*:

Local colloquial place names : They are used and understood only by local or regional populations. In large cities, for example, specific neighborhoods are often given names that are only understood within the city and its suburbs. Examples are the *"Viertel"* in Bremen or the *"Bermuda Triangle"* in Vienna (both denoting some city-blocks with an above-average density of pubs). Even if through press coverage or advertisement by tourist agencies such local colloquial place names become known outside the respective city (like for example the famous *"Haight-Ashbury"* district in San Francisco), it is very unlikely that they will make their way into national or international official place name lists.

Ad-Hoc place names : Many organizations "invent" their own place names, mainly to geographically manage and structure their data and information. Such names are used to denote project areas, regional offices, or product marketing areas. Examples are place names like *"the Desert States"* or *"the Far West"* used by the Smithsonean guides to historic and natural America, respectively (see Figure 6.8 in chapter 6).

Ad-hoc place names typically used only within a specific organization, and for a specific purpose. They are valid only for a limited time period, often on a project-to-project basis. Ad-hoc place names that are created by individual persons, for example to conceptualize a specific neighborhood, can be referred to as *personalized* place names (e.g., *"my neighborhood"*).

Figure 3.1 depicts examples for standardized and non-standardized place names, and their approximate ranges of lifetime. Historic place names like *"the British Empire"* have the longest lifetime because they have been used for a long period of time and are still used in a historical context. Ad-hoc place names have the shortest lifetime. They are created and used in the context of a specific

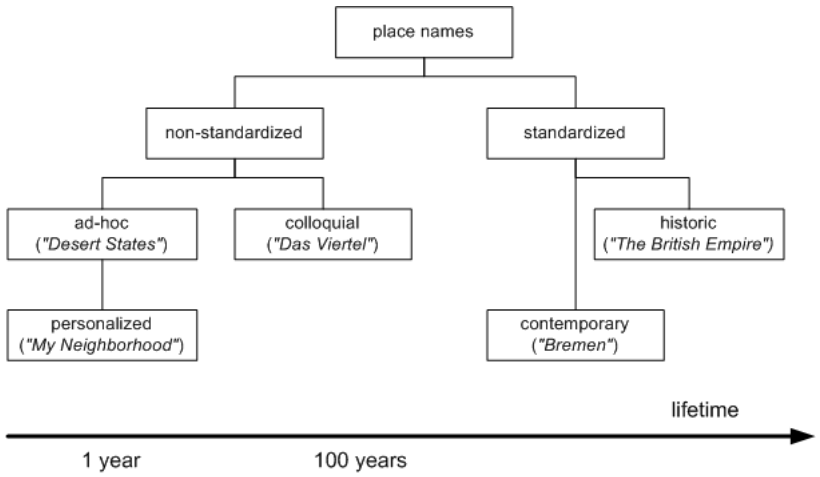

Figure 3.1: Different types of place names

task (e.g., to name a geographic area that is part of a urban planning project) and often disappear after the completion of that task.

Until recently, the research on place names was very much focussed on establishing the identity of geographic objects and providing standardized name identifiers for them. Consequently, most of the tools and standards used to administer and manage place names are geared towards the requirements of standardized place names only. However, some researchers are beginning to discuss the need for the development of more flexible and extensible tools. In addition to standardized place names, these tools have to be able to handle non-standardized place names as well [Goodchild, 1999]. These are seen as a basis for distributed geolibraries. We predict an even greater need for such tools caused by the development of location-based services (LBS) and other *"spatially-aware"* applications. Many of these applications are likely to use standardized as well as non-standardized place names, including ad-hoc and personalized place names.

3.1.3 Vague Place Names

Like many terms used in natural languages, place names are often not very well defined. This may involve the spelling of the proper place *"name"*, but also the *type* of place and it's *location* in geographic space.

According to the *Webster Dictionary of the English Language* [Webster, 2001], things that are *"not clearly, precisely or definitely expressed or stated, not precisely determined or known; uncertain"* can be called *"vague"*. In the context of place names, we refer to a the *taxonomic*, *conceptual*, and *locational vagueness* of a place name[1]:

Taxonomic Vagueness

The naming of geographic objects is often not consistent. Although a place name refers to a concrete instance of a natural or man-made geographic object, the

[1] A more thorough and general discussion of vagueness and uncertainty can be found in section 4.3.

same object may be named differently by different groups of a population. Often the naming reflects the historical evolution, the topography, and the ethnic composition of a specific geographic region. A specific mountain range, for example, may be named differently by the local populations native to two valleys located on one or the other side of the mountain.

In countries with populations of large ethnic variety, the use of *variant names* is particularly prevalent. Here, the same geographic object may have different names in different ethnic languages. Variant names are frequently used on a global level as well: When we refer to places in foreign countries, we often adjust the spelling and pronunciation of the place names to our own language, i.e. we use *exonyms*. For example, most publications that are written in English use the English term *"Vienna"* instead of the German *"Wien"* to refer to the capital of Austria.

Another source of taxonomic vagueness of place names is the use of the same name for different geographic objects (i.e. *homonyms*). This particularly the case in countries where place names have been assigned only relative recently, like in the United States or in Canada. Here, we find thousands of *"Clear Lakes"*, *"Bear Creeks"* and other rather generic place names. Also, because the European immigrants tried to keep the ties to their old origins, many European place names where duplicated in the new world. A well-known example is *"Paris"*, which not only exists in France but also among others, in Texas.

Conceptual Vagueness

The standardization efforts described above address mainly the taxonomic vagueness of place names. Much effort is spent to determine generally agreed-upon names and synonyms with a standardized spelling. Only recently there has been a discussion about *type hierarchies* and ontologies for geographic objects (e.g., [Kuhn, 2001]). The geographic objects referred to by place names can be classified according to their *type*, i.e., each place name refers to an instance of a specific class of geographic objects. Apart from the fact that the term *"geographic object"* is rather vague by itself, there exists a large number of different types of geographic objects.

From a practical point of view, the *Alexandria Digital Library (ADL) project* [Hill, 2003a] and several workshops in the United States have tried to define standards for the types of geographic objects that should be referred to by place names in gazetteers. As a result of these efforts, the *ADL Gazetteer Content Standard* identifies more than 1100 *feature types* (210 preferred terms and 946 related non-preferred terms) that are organized into a thesaurus-like hierarchical structure (see also section 3.2) .

On a more theoretical level, a discussion about *"geographic kinds"* in general and *"geo-ontologies"* in particular is going on in the scientific community. In information science, a popular definition by Gruber [Gruber, 1993] (see section 4.1) refers to an ontology as a *"specification of a conceptualization"*. Accordingly, a geo-ontology can be viewed as a specification of a conceptualization of geographic space, i.e. a description of types of natural and man-made phenomena that can be found on a mesoscopic geographic scale. Smith [Smith and Mark, 1998] refers to such descriptions as *"ontologies of geographic kinds"*. The *"geographic kinds"* range from physical features like lakes, mountains, or cities, to abstract concepts like states and regions. In any case,

they denote geographic concepts (example: *"a mountain"*) rather than instances (example: *"the Mont Blanc"*).

The question arises whether this amounts to the general (and well studied) problem of how to deal with conceptual vagueness in ontology design. Or if geographic concepts do have a special, intrinsically geographic quality, that calls for a special treatment of vagueness in geo-ontologies.

Locational Vagueness

Each geographic object that is referred to by a place name occupies a specific area in Geographic space. We call the 2-dimensional representation of a geographic object it's *regional extension*. The projection of the regional extension of a geographic object onto the 2-dimensional plane is equal to the *spatial footprint* of the respective place name. Although at first sight it may seem trivial to describe the shape and exact extent of a geographic object, it often isn't. Consider for example a coastline: If there is no tidal influence, the coastline defined as the interface between land and water can be determined more or less exactly. However, the land-water interface may show considerable temporal variations in regions with a strong tidal influence.

This example demonstrates the close relation between the scale, or granularity, of the representation of a geographic object and the spatial vagueness of the respective place name: On a global scale, the variations caused by tidal influence can be neglected, and a sharp line can be used to represent the coast line as an exact boundary between water and land. On a human scale, however, tidal influences will cause the land-water interface to vary, making the the boundary blurred. Finally, on a crab-scale, even regular wave-action will constantly change the land-water interface. In this work, we mainly discuss phenomena on a geographic scale (see section 2.1). Therefore, a geographic phenomenon like a coastline may be approximated as a crisp boundary, separating one region (for example an island) from another (the ocean).

However, there are place names for which the definition of a reasonably exact spatial footprint is not straight-forward. In fact, for many place names, only a *vague* regional extension can be defined. Following Cohn [Cohn and Gotts, 1996b], we can identify two different types of vague of regions in Geographic space:

Regions that do have crisp but unknown boundaries: For many regions in Geographic space, the knowledge necessary to define a crisp boundary is not available. A typical example for this *vagueness through ignorance* is the regional extension of an oilfield: The boundary between the area that belongs to the oilfield (i.e., the area where the pore space of the sediment is filled with oil) and the area that is outside of the oilfield (i.e., the area where the pore space of the sediment is not filled with oil) can only be inferred from limited sample data gathered through well-drillings. For two adjacent wells A and B, where A is within the field and B is outside the field, we can only say that the boundary of the field has to be somewhere between A and B.

Another reason for the difficulty to define the crisp boundaries of regions is linked to *temporal variations*, i.e. the fact that the extent of a region may change with time. Examples in the physical world are tidal zones,

flood plains, and rivers. Here, a region can be seen as a succession of short-lived regions with crisp boundaries.

Regions that do not have crisp boundaries: If the geographic object represents an artefact which is the result of a classification of some physical (field) properties, it may exhibit an intrinsic vagueness due to *field variations*: A mountain, for example, can be viewed as a specific pattern in an elevation field. It is easy to define the mountain top as the maximum elevation value. Where, however, does the mountain begin, and where does the valley start?

Another source of intrinsic vagueness is the *lack of consent* about the extent of a region. Different parts of a population may have different opinions about what area a place name refers to. A citizen of the North German town of *Emden*, for example, may have a very different opinion about the extent of the region called *"Ostfriesland"* than someone living on one of the islands off the coast. Often, such interpretational conflicts reflect long standing cultural and social differences between different ethnic groups. For that reason they are very different to resolve.

In general, the locational vagueness of place names often make it difficult to define appropriate spatial footprints in terms of crisp polygonal boundaries. To circumvent this problem, many state-of-the-practice applications resort to more or less crude approximations of the regional extent of a place name. In the following section, we will present these approximations in more detail and discuss their importance for the quality of spatial reasoning in the context of information retrieval.

3.2 Gazetteers - Tools to Manage Place Names

3.2.1 Functions of a Gazetteer

Standardized place names are organized and managed in place name lists, or *gazetteers*. A common example of a simple non-digital gazetteer is the index of place names that can be found among the appendices of any well-done atlas. A record in this index typically consists of the place name itself, information about the geographic object it represents, information about the country or region it belongs to, and a reference to the page and map-quadrant where it can be found in the atlas. Some indices include geographic latitude/longitude coordinates that typically refer to the centroids of the geographical objects denoted by the respective place names.

In general, the main functions of a gazetteer are :

- To provide a well-defined and standardized typonomic definition of the name of a geographic object.
- To provide information about the type and thematic classification of the object.
- To link the place name to a geographic location, i.e. to specify a *spatial footprint* for the place name.

Based on this functionality, a gazetteer can be used to answer questions like *"Where is the place with the name X?"* and *"What features are at location X?"*. Traditionally, gazetteers were published as part of topographic maps or as stand alone documents. Today, many gazetteers are available digitally. Such *digital gazetteers* offer a subset of the functionality usually attributed to geographic information systems (GIS), i.e. they can be seen as specialized GISs. But while organizations like the *UNGEGN* use gazetteers primarily to manage official and standardized typonomic definitions of place names, their application in the domain of information retrieval is more diverse. Here, gazetteers are used as *"geospatial dictionaries of geographic names"*[Hill, 2000]. Accordingly, their functionality is extended to:

1. Define a vocabulary of spatial indices, i.e., to select from all possible natural language terms those substantives that can be used as indices for spatial references, and as search terms for spatial queries.

2. Disambiguate place names in cases where several geographic objects may share the same name (e.g. there are at least 3 cities with the name *Bremen* worldwide).

3. Support spatial queries and reasoning about the spatial relevance (see chapter 7) of place names with respect to a spatial query.

4. Integrate data vertically by establishing a relation between a set of attribute data (including a place name) and a location in geographic space. This allows the fast and simultaneous access to a potentially large number of properties of a specific location.

5. Support the handling of large data sets by using relatively simple and lean data formats (as compared to standard GIS data).

As indicated above, digital gazetteers are currently used mainly in two application areas: Organizations concerned with the collection and standardization of place names use digital gazetteers for documentation and management purposes. In the digital library community, gazetteers are used for information indexing and retrieval tasks. Typical examples for the first application area are the GN250 and GN1000 gazetteers published by the German Bundesamt für Kartographie und Geodäsie (BKG) in close cooperation with StAGN [Beinstein and Sievers, 2002]. Two examples for gazetteers used within the digital library community are the *Getty Thesaurus of Geographic Names (TGN)* [TGN, 2002] and the gazetteer integrated in the *Alexandria Digital Library* [Hill and Zheng, 1999, Hill, 2000].

During the last decade, the importance of gazetteers for spatial information retrieval has been recognized within other communities as well. In the field of environmental information management for example, digital gazetteers have been integrated into a number of information management systems. In Germany, two of the most advanced applications in this area are the *German Environmental Information Network (GEIN)* [Bilo and Streuff, 2000, Angrick et al., 2002], and the German-Austrian *Umweltdatenkatalog (UDK)* [Legat et al., 1999, Lessing et al., 1995].

One of the major application areas of gazetteers (which we already mentioned above) is indicated by the name of the gazetteer integrated in GEIN.

This gazetteer is called the *Geographical Thesaurus of Environment (GTE)*, and analogous to thesauri that provide a taxonomy of terms, it provides a common vocabulary for place names. This common vocabulary can be used to annotate spatial metadata as well as to specify spatial queries. And just like the availability of *controlled vocabularies* is a pre-requisite for interoperable metadata in the terminologic field [Stuckenschmidt et al., 2000], gazetteers are the basis for interoperable spatial metadata. Controlled vocabularies have to be built on the basis of generally accepted standards. Consequently, a number of efforts are under way to develop national and international *gazetteer standards*.

3.2.2 Gazetteer Standards

Several parallel and partially interrelated initiatives are currently developing standards for gazetteers and gazetteer services:

***OGC Gazetteer Services*:** The OpenGIS Consortium (OGC) is developing specifications for interoperable gazetteer services as part of service-oriented *Spatial Data Infrastructures (SDIs)*. In March 2001, the *OGC* published a draft candidate implementation specification of a *Gazetteer Service Specification* [Atkinson, 2001]. In September 2002, the *Gazetteer Service Profile* of the *Web Feature Service Implementation Specification* was released [Atkinson and Fitzke, 2002]. The interface proposed in this paper extends the *OGC Web Feature Server specification (v1.0.0)* [OGC, 2002b] with the functionality needed to provide gazetteer services.

In a Gazetteer Service, the queryable attributes are specified as properties that describe geographic objects. These attributes include, but are not limited to, the type of the object, the name of the object, the public authority that is responsible for the object, and the identification code of the object. A Gazetteer Service may apply to a geographic region, such as a country, or some other specialized grouping of geographic objects. The Gazetteer Service is able to receive requests that define object attributes (e.g., a place name). The returned data will include a representation of the object (e.g., it's spatial footprint) in one or more geometries expressed in an OGC Spatial Reference System.

The *ISO 19112* Standard: In an effort closely related to the ongoing developments within the OGC, the *International Standardization Organization (ISO)* developed *ISO 19112 - Geographic Information – Spatial Referencing by Geographic Identifiers* as part of the *ISO 19100* series of standards (see section 2.2.2). This standard was released as a final international standard in October 2003 [ISO, 2003b].

ISO 19112 provides a conceptual schema and a general model for spatial references by geographic identifiers. Following this standard, geographic objects can be represented by place names only, i.e. spatial referencing by coordinates and coordinate-based spatial footprints is not addressed[2]. However, a mechanism for recording complementary coordinate references is outlined. In general, the standard defines the components of a spatial reference system and the basic components of a gazetteer.

[2]Coordinate based geo-referencing is dealt with by the *ISO 19113* standard, which is also a member of the *ISO 19100* family of standards.

According to ISO/TC-211, the standard intends to help users understand spatial references in datasets. Its application area is seen mainly in digital geographic data, but it can be extended to be used with other forms of geographic data such as maps, charts and textual documents.

The *ADL Gazetteer Content Standard*: Within the digital library community and in context of the *Alexandria Digital Library Project (ADL)*, Linda Hill and others at UC Santa Barbara developed the *ADL Gazetteer Content Standard (ADL-GCS)*. Version 3.1 of the ADL-GCS was released in September 2003 [Hill, 2003b]. This effort is based on a close collaboration of geo-scientists and proponents of the digital library community. In a series of international workshops (among others 1999 in Washington, and 2002 in Portland), the content standard and the underlying concepts were discussed by an interdisciplinary group of scientists and practitioners.

As a result of these discussions, the ADL content standard identifies three core components of a gazetteer: *name*, *type*, and *location*. A clear distinction is made between *spatial relations* and *functional relations*:

Spatial relations between ADL place names are given implicitly, i.e. they can be computed on the basis of coordinate based spatial footprints. The types of spatial footprints supported by the current ADL standard are points, bounding boxes,simplified polygons and complex polygons (see below).

The complexity of spatial relations that can be computed depends on the complexity of the spatial footprints involved. For one-dimensional point representations (which still make up the majority of all footprint representations in the ADL gazetteer), only metric distances and angles between points can be computed. Two-dimensional representations like bounding boxes and polygons support also reasoning about topological relations (e.g., *meet*, *overlap*).

Functional relations have to be specified explicitly using a number of hierachically organized relations. The relation hierarchy consists of a heterogeneous set of administrative relations, relations pertaining to naming conventions, and relations expressing physical-geographic concepts. The semantics of the relations are defined in the annex of the ADL content standard.

With the help of the functional relations, the position of a place name within a

- (abstract) administrative partonomy, and/or a
- (physical) topographic partonomy

can be defined. In addition to the partonomic relations, functional topologic relations (*PhysicallyConnectedTo* and *FlowsInto*) exist. These are used mainly for rivers and other pseudo 2-dimensional features.

By providing a set of semantically well-defined functional relations, the ADL content standard tries to set a standard for the representation of place names in gazetteers. The strong bias towards administrative

partonomies reflects the historic (and still prevalent) use of gazetteers as standardized lists of officially organized place names.

However, this focus on administrative and topographic partonomies expressed by explicitly defined functional relations is also one of the major weak-points of gazetteers like the ADL: It limits the spatial conceptualizations that can be represented by the gazetteer to one specific model, namely that of place names that are part of an (official) administrative and/or topographic hierarchy.

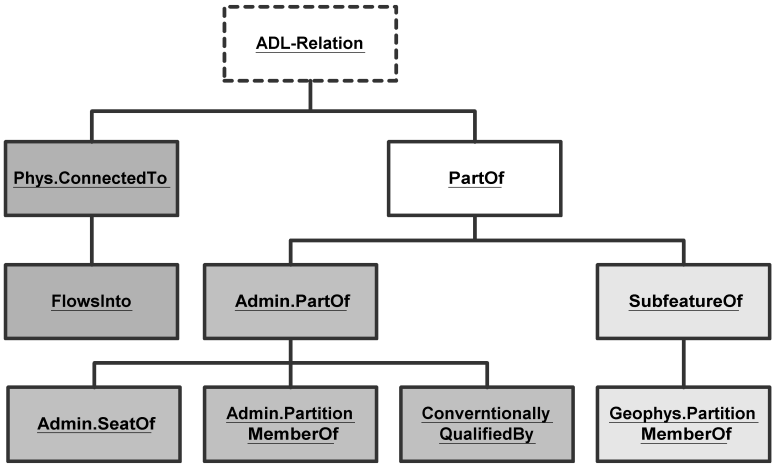

Figure 3.2: Functional relations in the ADL Content Standard

All three standards provide definitions of the basic components and data models needed to represent geographic objects in interoperable gazetteers and gazetteer services. Although the specifications show some differences, at least two of the standards, namely the OGC gazetteer standard and the ADL-GCS agree on the importance of *spatial footprints* as a core part of a gazetteer besides name and type information.

3.2.3 Spatial Footprint Representations

Apart from providing a common vocabulary for spatial metadata and spatial queries, one of the most important functions of gazetteers and gazetteer services in spatial information retrieval is to geo-reference place names, i.e. to establish a relation between a place name and a geographic location. The geographic location or region a place name is referred to is called the *spatial footprint* of a place name. By evaluating spatial relations between a geographic location specified in a spatial query and the spatial footprint of a place name, we are able to reason about the spatial relevance of the place name with respect to the query location.

The ability of a gazetteer to support reasoning about the spatial relevance of place names depends very much on which type of spatial footprint representations are used. A spatial footprint is a more or less crude abstraction from reality, i.e. from the real shape and extent of the geographic object denoted by a place name. The first major abstraction is the reduction of a 3-dimensional

geographic object to it's 2-dimensional projection onto geographic space[3]. How much further a spatial footprint has to be abstracted to be used in a gazetteer depends on a number of factors, including the representation scheme supported by the gazetteer and the availability of appropriate data. Nevertheless, the type and quality of the footprint representation has a strong impact on the complexity of spatial reasoning supported by a gazetteer. In general, we distinguish four different types of footprint representations at different levels of abstraction:

Point Footprints

The simplest footprint representation is one-dimensional, i.e. a *point footprint* expressed as a tuple of cartesian coordinates. For regional geographic objects, this point typically refers to the centroid of the object. Linear objects like rivers are often given a point representation that refers to the location where the river enters another river, or the sea. These examples already show the serious shortcomings of point footprints: Neither *"natural"* 3-dimensional geographic objects (lakes, rivers, mountains etc.), nor *"man-made"* 2-dimensional regions (administrative districts, boundaries of natural regions etc.) can be adequately represented as one-dimensional points. As a result, a large amount of spatial information is lost. The spatial reasoning capabilities of a gazetteer using point footprints are very limited.

As an example, we will look at Figure 3.3 which shows a map of the administrative districts of Bremen, Germany. The representation of the districts by their center points would reduce the areal extent of each district to zero. Such a simplifying representation could yield an answer to a simple query like *"Where is the district of* ***Mahndorf*** *?"*, or *"Which data objects are annotated with the place name* ***Mahndorf****?"*. It could also be used to find other districts within a certain buffer zone around the center point of the *Mahndorf* district. However, only those districts will be retrieved whose centroids are located within the buffer zone. Consequently, the result of the query will depend heavily on the specification of the buffer zone: If the buffer zone is too small, many of the larger neighboring districts will be overlooked. If the buffer is too large, districts that are not really neighbors of *Mahndorf* will be included as well.

Bounding Boxes

Although they are still rather simple geometric objects, a more realistic representation of a geographic object is given by it's *bounding box*. To span a bounding box, only two coordinate tuples are needed, typically one for the upper left and one for the lower right corner. This makes the bounding box a compact and easy to use spatial representation. Like point representations, bounding boxes for geographical objects can be computed if their spatial extent is known and available in digital (polygonal) format. But also the manual specification of a bounding box is not overly problematic. Compared to point footprints, bounding boxes support more complex spatial operations, like inclusion and intersection.

However, bounding boxes have some serious drawbacks: For one, depending on the geometry of the geographic object, bounding boxes tend to be much

[3]Note that we consider only 2-D spatial footprints in this work.

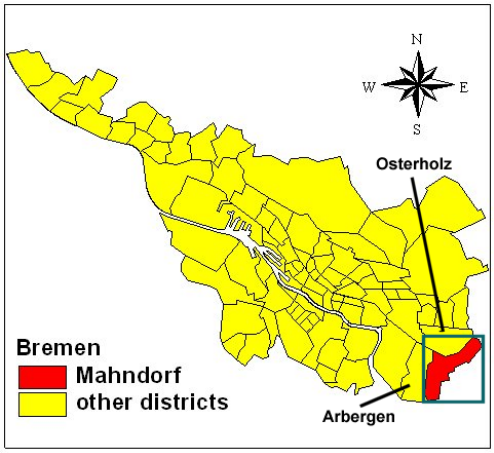

Figure 3.3: Districts of the city of Bremen

larger than the actual object. The bounding box of a specific district, for example, would cover a considerable part of the bounding box of another district as well. Secondly, the topological information represented by bounding boxes is rather imprecise: it cannot be concluded without doubt that the overlapping of two bounding boxes means the overlapping of the respective geographic objects. Applied to the map shown in Figure 3.3, the bounding box of *Mahndorf* would intersect with the districts of *Arbergen* and *Osterholz*, while these districts intersect in reality.

Polygonal Representations

Probably the best solution with respect to the accuracy and potential complexity of a spatial query is the use of *polygonal spatial footprint representations*. Based on exact polygonal footprints, full-scale GIS functionality can be applied to select all geographic objects within a given region, to determine neighboring polygons, or to perform other complex spatial queries.

Nevertheless, exact polygon representations have a number of disadvantages that make their application in gazetteers problematic:

Availability : For many geographic objects exact polygon representations are not available or difficult to obtain. For many place names denoting historical regions, natural regions, or colloquial names for geographic regions, exact polygonal representations cannot be constructed because their exact regional extent is not known .

On the other hand, there are geographic objects that could be outlined properly, but for various reasons such as high digitizing costs, or an unresolved controversy about the exact boundaries between objects, the creation of exact polygon footprints is not feasible (see section 3.1.3).

Complexity : Intuitively one may think that the more detailed a polygon representation is, the better. However, there may be true only to a certain extent. Sometimes the gain of more detail, for example by adding

more and more vertices to a footprint-polygon, is outweighed by a loss in efficiency due to the increasing amounts of data that have to be handled.

Computational costs : As indicated above, the processing and management of detailed polygon data involves complex software and high computation costs. Despite the availability of more and more efficient hardware and software, performance problems can be an issue for large online gazetteers, especially if they are accessed by many users simultaneously.

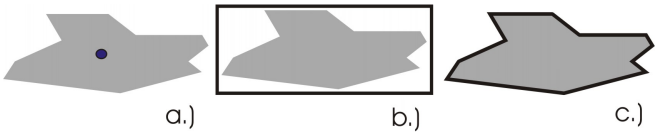

Figure 3.4: Common spatial footprint representations:(a) point,(b) bounding box,(c) polygon

Footprints Based on Spatial Indices

An alternative approach to the spatial footprint representations described above is the use of *spatial indices*. Spatial indices are used in spatial databases to manage spatial data in an effective way, and to increase the performance of spatial queries. The reference grids that are the basis for such spatial indices can be regular, irregular, or of variable density, like for instance *regional quadtrees* [Worboys, 1998].

But also outside the domain of spatial databases, spatial indices based on uniform reference grids have a long tradition as means to geo-reference geographic objects. For example, most of the place name lists integrated in atlases and city maps use spatial indices to geo-reference place names, often in conjunction with coordinate-based point footprints. Here, a place name is geo-referenced through the page(s) on which the map(s) depicting the respective geographic object can be found. The exact location of the geographic object on the map is given through identifiers of sections of a reference grid that overlays each map. In most cases, the reference grid is rectangular, but other geometries are possible as well.

In general, the spatial indices of a region are obtained through a binary mapping of the region onto a reference grid. If the region intersects a unit of the reference grid, this unit becomes part of the region's spatial footprint. A region is thus approximated as a set of reference units that belong to a specific reference grid. This representation does not require the specification of exact coordinates, nor the explicit statement of spatial relations. Spatial relations between regions in support of queries like: *"Give me all regions that intersect with region a"* can be computed directly, as long as all spatial indices are based on the same reference grid. The absolute references of the regions to geographic space can be derived from the spatial embedding of the reference grid.

Although spatial indices provide an effective tool to geo-reference regions and place names, they are used only by a few state-of-the-art digital gazetteers

as spatial footprint representations. One example is the *Geographical Thesaurus of Environment (GTE)* of the *German Environmental Information Network (GEIN)* [Bilo and Streuff, 2000]. The GEIN gazetteer uses spatial indices based on a uniform grid that subdivides the area of Germany into squares of $3x3$ km size. The region associated with a place name is projected onto the reference grid, and each place name is given a number of spatial indices based on which grid cells are covered.

Compared to coordinate-based approximations, index-based spatial footprint representations have the advantage that, although they offer less flexibility than complex polygonal representations, they are more expressive than the simpler coordinate-based representations (e.g., point- or bounding-box footprints). In that sense, index-based spatial footprints offer an practical spatial representation at an intermediate level of complexity. In addition, to lower computational costs and allow for faster queries, implicit topologic relations between index-based spatial footprints (e.g., inclusion and intersection) can be computed a-priori and stored in the gazetteer database.

But although index-based spatial footprint representations are quite useful, they also do have a number of shortcomings:

Fixed resolution: How accurately a regional geographic object can be approximated in terms of spatial indices depends very much on the resolution of the reference grid relative to the size of the objects to be represented. Good results, i.e. relative accurate abstractions can only be achieved when the size of the grid cells is small compared to the size of the objects. However, geographic objects may vary considerably in size. Uniform reference grids like the one used by GTE are therefore optimized only for objects within a specific scale range.

Solutions to this problem may be found by the application of irregular reference grids, regular grids with a variable resolution (like the quadtrees mentioned above), or a combination of multiple reference grids at different resolutions. However, the this would jeopardize one of the major advantages of digital gazetteers over regular GIS systems, namely the use of a relatively simple and light weight data structure.

Data interoperability: In general, the aspect of gazetteer interoperability is an argument against the use of index-based spatial footprints: Only footprints that refer to the same reference grid can be directly compared. For footprints based on other grids, transformations have to be applied that are non trivial if the grid cells do not match (i.e., if the reference grids have a different resolution and/or a different origin). Coordinate-transformations are also necessary to compare spatial footprints based on different coordinate systems. But while a specific coordinate system may represent a standard in one or several countries, there are no standardized reference grids so far. Each gazetteer application uses its own, application-specific reference grid. This makes transformations between reference grids difficult and cumbersome.

In summary, all four types of spatial footprint representations described above do have specific drawbacks that reduce their practical applicability in digital gazetteers. Which representation should be chosen depends on the specific requirements of the gazetteer application. Hill [Hill, 2000] proposes to use

the attribute *"satisficing"* as a criterion for the selection of footprint representations for specific purposes. A *"satisficing"* spatial footprint is one that, for a given problem, finds the best compromise between the (technically) optimal and the cheapest (in terms of money and computational costs) solution. This may be the reason why, despite their limitations, simple coordinate-based point representations are still the most frequently used spatial footprints in digital gazetteers.

Chapter 4

Qualitative Representation and Reasoning

4.1 Spatial Information Retrieval

4.1.1 Information Discovery and Information Retrieval

In recent years, *information retrieval (IR)* and *information discovery* has gained increasing attention from both industrial and academic researchers (e.g., [Levine, 1995, Baeza-Yates and Ribero-Neto, 1999, van der Weide, 2001]). For one, this can be attributed to the growing number of documents and information objects that are produced and become available in digital format. Another factor is that the distribution of this information does not rely any more only on proven and well-established distribution channels like newsletters, reviewed journals, (physical) libraries, but that more and more information sources become available through distributed and heterogeneous networks like the Internet and the World Wide Web. One of the advantages of this development is that vast amounts of data and information become available to every PC connected to the Internet. From a users perspective, however, it has becomes increasingly difficult to find the right *"needle"* (i.e. the information object of interest) in this huge *"haystack"* of digital data.

Numerous IR techniques have been developed to tackle the information overload problem. Most of these techniques concentrate on mathematical models and algorithms for information retrieval. A number of IR models such as the *Boolean model*, the *vector-space model*, the *probabilistic model* and their variants are already well established in the community [Baeza-Yates and Ribero-Neto, 1999]. Some of these methods left the realm of academic research and were successfully integrated in popular search engines like Google [GOOGLE, 2004].

With recent efforts to establish a *"Semantic Web"* (e.g., [Fensel, 1999, Berners-Lee et al., 2001b, Fensel et al., 2003]), the development and application of so-called *"intelligent"* methods for information retrieval has received much attention. These methods use thesauri and formal *ontologies* to define concept spaces that can be used both to provide metadata descriptions for information objects, as well as to specify queries for information retrieval based

on such metadata [Guarino, 1998, Bishr and Kuhn, 1999]. The term *"ontology"* has been used in many different ways by different scientific communities [Guarino and Giaretta, 1995, Uschold and Grueniger, 1996]. In computer science, ontologies are applied to overcome the problem of implicit and hidden knowledge by making the conceptualization of a domain explicit. Many ontologies used in computer science follow a popular definition by Gruber [Gruber, 1993]:

DEFINITION 4.1
An ontology is an explicit specification of a conceptualization.

There are a number of ways in which an ontology may explicate a conceptualization and the corresponding context knowledge. The possibilities range from a purely informal natural language description of a term corresponding to a glossary to strictly formal approaches with the expressive power of full first order predicate logic or even beyond (e.g. Ontolingua [Gruber, 1991]). An overview of the use of ontologies for information integration is given in [Wache et al., 2001] and [Visser et al., 2001]).

Formal ontologies are generally used to encode the semantics of concepts. In information retrieval, logic reasoning based on such representations allows to make assumptions about the relevance of individual concepts with respect to concepts specified in an information request. The relevance of an information object with respect to the information request is then determined as a function of the relevance of the concepts it is associated with (e.g., through metadata). These concepts determine what can be called the *"thematic"* or the *"conceptual coverage"* of an information object.

In addition to (one or multiple) specific concepts, many information objects can be associated with specific locations in geographic space. The *"spatial coverage"* of an information object can be expressed using *place names* (see section 3.1). For example, the spatial coverage of a news report describing the terrorist attacks of $9-11$ may be described through the place names *"New York"* and *"World Trade Center"*[1].

The *"temporal coverage"* of this report is obviously be given by a specific date, namely *September 11, 2001.* However, in colloquial English, this specific day is often referred to simply as *9/11* which indicates that to resolve the temporal coverage of an information object in an information request, methods for the representation of natural language temporal expressions have to be found.

The conceptual, but also the spatial and the temporal coverage of an information object play an important role for information retrieval. Meaningful information retrieval algorithms should therefore be able to address all three of them in a comprehensive way. We call information requests that are based on the joint analysis of the spatial and conceptual coverage of an information object *"spatio-conceptual"* queries, or queries of the type *concept @ location* [Schlieder et al., 2001]. If the temporal dimension is added, we can speak of information requests of the type *concept @ location in time* [Vögele et al., 2003a]. To be able to process such queries, an information retrieval agent has to be able

[1] However, depending on the contextual focus of the report, other place names like *"Afghanistan"*, *"Saudi Arabia"* or *"Hamburg"* could also be relevant. This shows that it is often difficult to determine the spatial coverage of an information object if no metadata are provided. In this work, we therefore focus on information objects that are annotated with the respective spatial metadata.

to evaluate the *conceptual relevance*, the *temporal relevance*, and the *spatial relevance* of an information object.

As we have seen above, the evaluation of conceptual relevance is a topic that has been addressed by a number of researchers. Temporal and spatial relevance, on the other hand, seem to have evaded the interest of researchers in the IR community so far. Only a few researchers have addressed information retrieval that integrate the conceptual and the spatial, or even the conceptual, spatial, and temporal search dimensions.

Alani [Alani et al., 2001], for instance, developed the OASIS system which implements an approach for spatio-conceputal information retrieval in domain of archeology and cultural heritage. He combines an ontology on archeological artefacts with an ontology of places that uses place names taken from two standard gazetteers (i.e., the *Getty Thesaurus of Geographic Names (TGN)* [TGN, 2002] and *Bartholomew's* digital data [Harper-Collins, 2001]). He uses the integrated ontology to derive measures for the thematic and the spatial relevance of archeological artefacts associated to specific places, and to rank results of a database search accordingly.

For the computation of spatial relevance, the OASIS system combines Euclidean distance with a distance measure that accounts for the hierarchical distance of places with respect to administrative units. The metric to compute this distance was derived from similar metrics used to evaluate concept similarities in semantic nets. This indicates that spatial relevance is more than just closeness, or a lack of distance. Our first step to find a practical solution to the problem of how to evaluate the spatial relevance of information objects is therefore to take a closer look at the semantics of the term *spatial relevance* itself.

An example for an information broker that addresses all three search dimensions is the *BUSTER* system [Visser et al., 2002, Vögele et al., 2003a], which is described in more detail in section 8.2.1.

4.1.2 Parameters of Spatial Relevance

In this work, the term *spatial relevance* is used mainly in the context of information retrieval and information discovery. An information object is said to be *spatially relevant* with respect to a (spatial) query if the spatial coverage of the information object is spatially relevant to the spatial coverage of the query. Under the assumption that the spatial coverage of both the information object and the spatial query is given by their geo-references (i.e., location(s) in geographic space), the spatial relevance of an information object with respect to a spatial query is a direct function of the spatial relevance of geographic locations:

Definition 4.2
The spatial relevance of an information object with respect to a spatial query is a function of the spatial relevance of the geographic location(s) the information object is associated with and the geographic locations(s) associated with the query.

Following this definition, a method to evaluate the spatial relevance of georeferenced information objects has to be based on a comparison of *geographic*

locations[2].

This notion of spatial relevance is based on a pragmatic view of the relation between objects and geographic space. Analogous to the formalized ontologies and decidable logics used to reason about the conceptual semantics of objects, we try to develop a practical approach to evaluate what can be called the *"spatial semantics"* of geographic objects. This approach is based on the assumption that by evaluating a number of simple spatial relations, practical measures to compare the spatial semantics and spatial relevance of objects can be found.

Accordingly, if we say that a location a is *"spatially relevant"* to a location b, we have to think about which parameters this judgement may depend on. For one, as indicated by the fact that we intuitively speak of two locations that are *"relevant to each other"* we can say that spatial relevance is a *binary relation* between geographic locations. This means that spatial relevance is not an intrinsic property of location, but can only be defined between a pair of specific geographic locations.

Definition 4.3
The binary relation $\sigma(a, b)$ denotes the spatial relevance σ of a location a relative to a location b.

Secondly, in an information retrieval context, we want to use spatial relevance to sort locations relative to a query location. We therefore have to define an *ordering* for spatial relevance: It does not suffice to say that a location b is spatially relevant to a location a. We need to develop a metric that allows us to decide whether a location c is spatially more relevant to a than a location b.

Definition 4.4
For three geographic locations a, b, and c, the partial ordering $\sigma(a, b) > \sigma(a, c)$ signifies that location b is spatially more relevant to a than location c.

The ordering of spatial relevance is a function of a number of spatial relations. Among the most important spatial relations in this context are distance, topology, orientation, and partonomy. At this point, we will give a brief overview of these relations. A more detailed discussion follows in section 4.2.

Distance

According to an often cited statement by Tobler [Tobler, 1970], in geographic space *"everything is related to everything else, but near things are more related than distant things"*. This indicates that the *distance* between two locations is an important parameter to determine their spatial relevance and corresponds well with our common-sense knowledge about spatial relevance: In most cases, locations *close* to a location of interest are considered to be more relevant than locations that are *far* away. However, the semantics of natural language terms like *close* and *far* are highly context-dependent. One context, for example, is the available means of transportation and the speed with which we can move

[2]In the context of this work, the terms *"geographic object"* and *"geographic location"* are used as synonyms. This is feasible because every (physical or non-physical) geographic object has a spatial extension that, mapped onto the 2-D plane, constitutes its spatial footprint. Through its spatial footprint, each geographic object refers to one or several locations in geographic space.

through geographic space: For a pedestrian, *close* has a significantly different meaning than for the driver of a fast car or a passenger in an airplane.

To make things even more complicated, the semantics of an expression like *a is near b* may also depend on intrinsic properties of the locations and the respective spatial objects. Studies suggest that humans build *influence areas (IA)* around spatial objects they perceive in their environment [Gahegan, 1995]. The influence area of a spatial object is proportional to the salience of the object in the environment, and it allows humans to evaluate metric measures and to qualify positions and distances between objects.

Kettani and Moulin [Kettani and Moulin, 1999] illustrate the notion of an influence area using the following example: If two mountains in the Himalayas are said to be 10 km apart, everybody will agree that these mountains are *"close"* together. However, two cars that have the same quantitative distance are perceived as being *"far"* apart from each other. This example shows that perceived distance, and with it the influence area around spatial objects, depends on the relative size and the relative importance associated to the object.

Topology

Topological relations are another class of spatial relations that can be used to make assumptions about spatial relevance. Which topological relations between two spatial objects can be evaluated depends on the dimensionality of their representations. Because this work is focussed in information retrieval with place names, i.e. with objects that have a regional extension in geographic space, we will mainly look at topological relations between 2-dimensional regions. Here, relations of equality, containment, overlap, and connection are the most interesting with respect to spatial relevance.

If we look at the topological relations possible between two 2-dimensional regions, they seem to show a partial ordering with respect to spatial relevance that follows common-sense reasoning: two regions that are connected are more relevant to each other than two regions that are disconnected. Likewise, two overlapping regions seem to have more in common (i.e., are more relevant) than two connected regions. And a region that is part of another region (i.e., that is fully overlapped by the region) is probably more relevant than a region that is only partially overlapped. For two overlapping regions, we can make the assumption that their spatial relevance is proportional to the degree of overlap.

Following this approach, we can establish an ordering of topological relations between 2-dimensional regions where equality indicates the highest, and disconnectedness the lowest spatial relevance between two regions. This ordering is closely related to the natural sequence of topological relations often expressed in terms of *conceptual neighborhoods* (see section 4.2.3).

Orientation

Orientation defines the position of spatial objects with respect to a reference system. In geographic space, the dominant reference system is the one of *cardinal directions* [Frank, 1996]. Cardinal directions are used both to reference information and to specify spatial queries: We may say that a location is *"south of the border"*, or we may look for locations that are *"to the West"* of a location of interest.

Spatial relevance can be linked to the difference between the requested and the actual orientation of a location with respect to a location of interest. For example, given the typical quantitative 360° reference system of cardinal directions, a spatial query for locations *"to the West"* of a location of interest will consider a spatial object at 270° (i.e., 0° divergence) to be more relevant than a location at 300° (i.e., 30° divergence).

Hierarchical Partonomies

Cognitive studies indicate that humans tend to organize geographic space according to hierarchical principles (e.g., [Hirtle and Jonides, 1985, Freksa, 1991, Fotheringham and Curtis, 1992]). An example is the organization of countries into hierarchical partonomies of administrative units. In such a hierarchy, the spatial relevance of two regions (as well as the regions, locations, and geographic objects contained in these regions) is affected by the position of the region within the partonomy: Two regions that belong to the same super-region are spatially more relevant than two regions that belong to different super-regions. As a consequence, we can establish a partial ordering of spatial relevance on the basis of the hierarchical closeness of regions. The hierarchically closer two regions are, the more spatially relevant they are.

However, like in the case of distance (see above), the effect of hierarchical closeness on overall spatial relevance depends on the nature of the information object as well as the context of the spatial query. For instance, it makes a difference whether we are looking for the closest restaurant to have lunch at, or the closest school to send our children to. In the first case, hierarchical closeness does play only a minor role: The most relevant restaurant is the one that is the closest (in terms of Euclidean distance), or the most accessible in terms of travel time. On the other hand, the closest or the (geographically) most accessible school may not be within the limits of our school district, i.e. it does not belong to the same administrative super-unit as our home. Therefore it may not be feasible to send our children to the closest school in a geographic sense, but rather to the closest school in an *administrative* sense.

In summary we can say that all four types of spatial relations (i.e., *distance*, *topology*, *orientation*, and *partonomy*) described above seem to influence the spatial relevance of locations, regions, and geographic objects with respect to spatial queries. In the following we will examine these spatial relations in more detail. We will discuss qualitative vs. quantitative methods for spatial reasoning in general and then briefly outline how these relations can be evaluated quantitatively, for instance in geographic information systems. We will then describe the most important qualitative methods that were developed for spatial reasoning with these relations.

4.2 Methods to Handle Spatial Relations

4.2.1 Quantitative vs. Qualitative Methods

The main focus of this work is to develop a representation scheme, and based on this scheme, appropriate reasoning mechanisms that support reasoning about the spatial relevance of geographic objects. These tools are to be used in the domain of information retrieval to enhance spatial information requests. Given

a set of objects distributed in geographic space, where one object has been selected as *object of interest*, our goal is to find an appropriate ordering that reflects the relevance of all objects in the set with respect to the object of interest. This ordering is based on the assumption that the objects can be assigned a *spatial relevance* relative to the object of interest. In our approach, spatial relevance is a function of a number of spatial relations, among them *metric* relations (denoting *distance* and *proximity*), *topologic* relations (used to express *connectedness* and *neighborhood*), *ordinal* relations (used to express the relative *positions* of objects), and *partonomic* relations (reflecting the *hierarchical ordering* of geographic objects).

Computational geometry as used by standard Geographic Information Systems (GIS) and spatial databases offers a number of methods to compute and to quantify spatial relations. All these methods have in common that they rely on exact representations embedded in a coordinatized space (i.e., in most cases the Euclidean space).

In a coordinatized space it is trivial to sort a set of objects on basis of their *absolute distances* relative to an object or location of interest. Likewise, topologic relations between objects can be computed with the help of *"spatial operations that can determine whether one feature touches, coincides with, overlaps, is inside of, or is outside of another feature"* [Zeiler, 1999]. Algorithms to compute metric and topologic relations between objects in coordinatized spaces therefore belong to the basic functions used in Geographic Information Systems (GIS) and spatial databases to solve spatial queries [Günther, 1998].

Combined with a suitable metric, these operations can be extended to give a quantitative account of some topological relations, like for example the *overlap* relation. If we can compute by *how much* an object a overlaps an object b, this degree of overlap can be used as a measure for spatial relevance. An example for a system that uses such operations for spatial queries is the gazetteer use by the *Alexandria Digital Library (ADL)* (section 3.2). In the *ADL* gazetteer, spatial relevance can be computed as *"a measure of the comparative size and overlap of two spatial areas - the query area and the spatial footprint of the collection object. The higher the value, the more similar the areas are in terms of size and overlap"* [ADL, 2002]. In the same context, spatial relevance is defined as *"a measure of how closely the item's footprint matches the query box (the area of overlap is included in the calculation as well as the sizes of the two areas)"* .

Quantitative representations of geographic objects and the respective algorithms for spatial reasoning seems to be simple and straightforward. However, they do have a number of shortcomings that limit their applicability, in particular in the domain of (spatial) information retrieval. Some of these shortcomings were already mentioned earlier in the discussion of gazetteers and spatial footprint representations (see section 3.2). In summary, these are:

Unavailability of exact data : Quantitative methods for spatial reasoning rely on exact representations of objects in coordinatized spaces. Both the complexity of the spatial reasoning and the quality of the reasoning results depend heavily on the quality and precision of the available quantitative representations. In section 3.2.3 we have seen that for a number of technical and non-technical reasons, exact quantitative representations of geographic objects (e.g., in the form of polygons) are not always available, or are only available at relatively high costs.

We also mentioned that many state-of-the-practice applications circumvent this problem by using more or less crude spatial footprint approximations. Bounding boxes and point representations are used as a substitute for detailed polygon data. The obvious result of compromising on data quality is a decline in the complexity and quality of the reasoning process. Complex spatial queries on the basis of coarse, highly approximated spatial representations will not yield adequate results. For example, it does not make sense to compute the relative overlap of two bounding boxes, given the amount of error introduced by such a coarse approximation of a region.

Bad data quality : A limiting factor closely related to the previous one is the problem associated with the quality of geospatial data. With quality we refer mainly to the precision and consistence of the digitalization. Even in GIS applications where exact polygonal representations of regional objects are available, problems with data quality are endemic [Worboys, 1998]. This can affect the quality of topologic and other spatial computations. Two neighboring objects (i.e., polygons) that, due to a sloppy digitization, only slightly overlap will cause an error in an evaluation of topologic relations.

Common-sense spatial reasoning is qualitative : Spatial queries based on quantitative spatial representations fail to *"comprehend"* the semantics of what the human user wants to express in the query. A typical spatial query like *"Show me all restaurants close to the railway station"* can easily be implemented by drawing a circular spatial buffer with a radius r around the centroid of the polygon representing the railway station. But what distance expressed in meters or kilometers should this radius r represent? Certainly the common-sense notion of closeness is heavily influenced by the context of the query. *Close to the railway station* may not be the same as *close to the ocean*.

Similar problems arise then we talk about topological and partonomic relations: Given an object c that is part of a larger object a. If a touches another object b, but c does not, is c relevant to b? A simple evaluation of the topologic relationships between the three objects would indicate that c is disjoint from b, and therefore not relevant. Our intuition, however, tells us that c as part of the relevant larger object a may be relevant as well.

We see in particular the last issue, i.e. the discrepancy between exact spatial representations in coordinatized spaces and the human preference to model geographic space qualitatively, as a serious shortcoming of GISs and other purely quantitative systems in the domain of information retrieval. This relates not only to the expressiveness of the information requests supported by such systems[3], but also to their inability to provide an intuitive and cognitively sound

[3]For example, exact metric distances and the use of (boolean) buffer operations as used in GIS is not particularly well suited to support intuitive spatial queries because they introduce a level of precision that in many cases is not needed and even counter-productive. A spatial query based on a boolean evaluation of a buffer zone ignores all *"near hits"*, i.e. all locations that are located just outside the buffer zone. However, the best answer to a multi-component query that is often only vaguely specified by a human user may be just among these near hits.

user-interface to spatial information, a problem that has been recognized for some time within the GIS community [Egenhofer and Mark, 1995]. An alternative could be the use of qualitative spatial models and qualitative spatial reasoning.

Qualitative Spatial Reasoning

To overcome the shortcomings of quantitative representations and computational geometry, we will examine representations and reasoning methods that were developed in the field of *Qualitative Spatial Reasoning (QSR)*. *QSR* is a science that has developed as a branch of qualitative reasoning *(QR)*(e.g., see [Freksa, 1991, Freksa and Roehrig, 1993, Cohn, 1997, Cohn and Hazarika, 2001]). Qualitative reasoning tries to model the physical world, and to enable a computer to make predictions about this world, without having to resort to quantitative representations. Interest in qualitative reasoning has been growing lately because high-precision quantitative measurements proved to be not as useful as was initially believed. This is true for both the analysis of complex systems, as well as for common-sense reasoning. There are indications that higher cognitive mechanisms employ qualitative rather than quantitative mechanisms, which makes qualitative knowledge representation and reasoning an interesting topic for many applications in Artificial Intelligence (AI) [Freksa and Roehrig, 1993].

QR can be applied to integrate the relatively simple common-sense knowledge about the physical world with the complex quantitative models developed by scientist and engineers. To do so, qualitative reasoning has to *"find ways to represent continuous properties of the world by discrete systems of symbols"* [Cohn and Hazarika, 2001]. A general approach is to group continuous values into equivalence classes. A set of qualitative values based on equivalence classes is called a *quantity space*, and typically exhibits a partial or total natural ordering. Exploiting the transitivity of this ordering is frequently used for qualitative inferences.

Another reason for the use of *QR* is to lower computational costs through a reduction of data without loss of (vital) information. Because qualitative representations *"make only as many distinctions as necessary to identify objects, events, situations etc. in a given context"* [Hernandez, 1994], they allow for more compact representations of problems than quantitative approaches do. This is true in particular for geographic objects, where exact quantitative representations (e.g., in the form of high-precision polygons) typically result in large data volumes[4].

Qualitative Representations

Qualitative reasoning has been studied extensively in the temporal domain. Allen's approach of one-dimensional qualitative reasoning with time intervals [Allen, 1983, Allen, 1984, Allen, 1991] has become a paradigm for qualitative

[4]With the ever-increasing capacity and power of today's computers, computational cost may seem to be a minor problem that will be solved soon. But the general complexity and non-linearity of many problems on the one side, and the trend towards distributed and mobile applications on the other side make light-weight methods for knowledge representation and reasoning a relevant research topic.

reasoning. Based on this work, a number of similar methods were developed for the spatial domain (e.g., [Freksa and Roehrig, 1993, Schlieder, 1999]).

The choice of primary primitives is one of the most fundamental characteristics of a representation scheme for qualitative spatial knowledge. Most mathematical theories of space use points as primary primitives and model extended spatial entities as sets of points. A number of researchers with strong ties to the worlds of GIS and spatial databases adopted this approach to define *"point-set topological spatial relations"* (e.g., [Egenhofer and Franzosa, 1991, Egenhofer and Franzosa, 1995], and also [Worboys and Bofakos, 1993]). On the other hand, there is strong tendency to directly use *regions* rather than points as primary primitives (e.g., [Randell and Cohn, 1989, Randell et al., 1992, Cohn and Hazarika, 2001]).

Which embedding space is used is another important characteristic for the representation of qualitative spatial knowledge. While high-level approaches favor continuous space, application-oriented low-level approaches often prefer a discrete embedding space. An attempt to bridge the gap between these two fundamentally different approaches was made by Galton [Galton, 1999]. He developed a *"mereotopology of discrete space"* by modifying representations of topological relations that were developed for continuous space to be able to work in discrete space.

There is no general approach to qualitative representation of space and qualitative spatial reasoning. The representation schemes and reasoning methods that were developed by different schools within the scientific *QSR* community are targeted at specific types of problems and spatial relations. For each type of spatial relation identified in the previous section to be a key-parameter for spatial relevance reasoning (i.e., distance, topology, orientation and partonomy), a number of different and sometimes competing qualitative approaches exist.

In the following, we will examine some of the applicable quantitative and qualitative methods for spatial representation and reasoning. This will include a brief overview of the respective theories and the state-of-the-art in the field of *QSR*.

4.2.2 Metric Relations: Distance and Proximity

Quantitative Distance and Buffer Operations

The selection of locations and spatial objects based on their metric distance from a location of interest is a standard method in geographic information systems and spatial databases. Given a set P of point locations p_i (i.e., objects represented as tuples of Cartesian point coordinates), a circular buffer that has a radius r and is centered in q (the location of interest) may be drawn (Figure 4.1). All point locations $p_i \in P$ with a distance $d(p_i, q) \leq r$ (i.e. all points that are located *within* the circular buffer) belong to the result set R of the query.

In 2-dimensional Euclidean space, which in the context of this work is the standard frame of reference for the representation of geographic objects, it is trivial to compute the metric distance between point-locations. Under the assumption that metric distance is inversely proportional to spatial relevance, a partial ordering for all points $p_i \in R$ with respect to q can be established, with those points being spatially most relevant that are closest to q. Accordingly, in Figure 4.1, the spatial relevancies $\sigma(p_i, q)$ of the points p_1, p_2, p_3, p_4 with respect

to q show the following order: $\sigma(p_1, q) > \sigma(p_2, q) > \sigma(p_3, q) > \sigma(p_4, q)$.

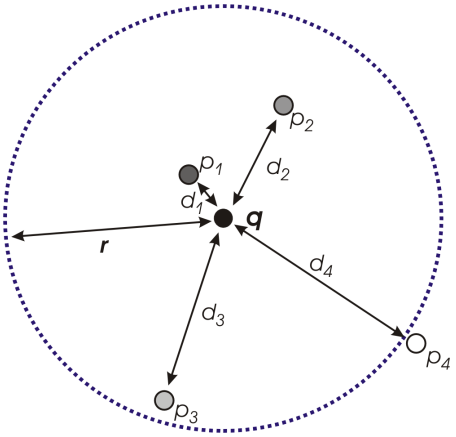

Figure 4.1: A circular buffer

It is important to note that in a standard buffer operation, p_4 will not be in the result set of a buffer query with radius r because $d_4 > r$, i.e. the point is located *outside* the buffer area.

The example above shows that the selection of spatial objects through a buffer operation combines two types of inferences, namely a distance measurement and the evaluation of the topological relation of the object with respect to the buffer area. This is even more obvious when a buffer is used to select objects of higher dimensionality, i.e. line objects or areal objects. In general, a buffer operation can be performed in two distinct steps:

1. A buffer region with a specific width w is defined. For radial buffers, w is incident with the radius r. For non-radial buffers (e.g., buffer zones around line or 2-d areal objects), w refers to the metric length of the perpendicular drawn between the outer boundary of the buffer and the outer boundary of the areal object.

2. To decide wether a spatial object belongs to the result set of the buffer query, the topological relation of the object with respect to the buffer zone is evaluated. This evaluation is typically a binary boolean function

$$B : b \longrightarrow \{0, 1\} \tag{4.1}$$

where a spatial object x is either said to be inside ($b(x) = 1$) or outside the buffer ($b(x) = 0$). For point-objects, B is given by the evaluation of the $is - contained$ relation. For line- and areal objects, B may also be interpreted as the $intersects$ and even $is - connected$ relations.

Qualitative Distance

Measures in Euclidean space express distances in exact quantitative numbers. Humans, on the other hand, tend to use distance in a more qualitative way. To describe distances between spatial objects, we prefer qualitative statements

like *"near"* and *"far"* over exact numbers in kilometers and meters. Likewise, spatial queries expressed in natural language have a preference of qualitative expressions. For example, we would rather specify a query like *"Which pubs are close to the beach?"* instead of *"Which pubs are located within a 500 m linear buffer along the beach?"*.

Qualitative spatial reasoning offers a number of methods to express and reason about *qualitative distance*. On an imaginary scale of the different abstraction levels of spatial relations, qualitative distance, together with orientation and shape, can be found somewhere *"between topology and fully metric spatial representation"* [Cohn and Hazarika, 2001], i.e. somewhere between the most qualitative and the most quantitative type of spatial relations. In *QSR* there are two different families of methods for the representation of qualitative distance: Those that measure distance on some *absolute* qualitative scale, and those that apply some kind of *relative* distance comparison.

Absolute qualitative distance: Distance is normally thought of as being a one dimensional, linear concept. Therefore traditional *QR* techniques based on qualitative algebras and linear quantity spaces can be applied. One method is to compare *orders of magnitude*[Mavrovouniotis and Stephanopoulos, 1987] : If a qualitative distance a is said to be *much larger* than distance b, then a large number of units of b have to be summed up to reach the same magnitude as a.

Another method to compare qualitative distances on an absolute scale is the *Delta Calculus* developed by Zimmermann [Zimmermann, 1995]: He uses a triadic relation $x(>, d)y$ to encode the fact that x is larger than y by the amount of d. Consequently $x(>, y)y$ denotes that x is more than twice as big as y.

Relative qualitative distance: Relative representations of qualitative distance try to describe qualitative distance in terms of named distances like *"near"* and *"far"*. As early as 1922, De Laguna [DeLaguna, 1922] specified the triadic relation $CanConnect(x, y, z)$. This relation is true if a region x is able to connect two regions y and z without having to apply a simple translation (i.e. scaling, rotation or shape change). Using the *CanConnect* relation, specify notions like *equidistance*, *nearer-than*, *farther-than* and *connected-to* can be expressed.

A general question related to the measurements of distances between regions is *where to measure from*. This question is not only interesting for quantitative distance measurements, but also for qualitative distances. One way to measure the distance between two regions a and b is to take the minimum distance between two points that belong to the boundaries of a and b, respectively. This approach is used in de Laguna's *CanConnect* relation. Another option is to measure the distance between the centroids of a region, or between subregions and points within a region.

As we pointed out above, named qualitative distances are context dependent [Escrig and Toledo, 1998]. One possible context is *scale*: Two cities A and B appear to be *close* on a map of European scale, while they are *far apart* on a map of regional scale. Other contexts are accessibility (e.g. given by the existence or non-existence of road connections) and travel time. In the latter, distance

is assymetrical because the travel time in one direction may be different from the travel time in the other direction. Distance may even depend on human perception. [Holyoak and Mah, 1982] confirmed our intuition that distances in areas that are not well known tend to be underestimated.

4.2.3 Topologic Relations

The term *"topologic"* refers to a group of spatial relations that remain invariant under a group of homeomorphisms, or *"rubber-sheet"* transformations. If we visualize the 2-D plane as a sheet of flexible material, rubber-sheet transformations are those transformations that stretch and distort the plane without folding or tearing it. The *connection* relation, for example, is a topologic relation because two objects remain connected no matter what rubber-sheet distortions are applied to them.

Topology is said to be *"perhaps the most fundamental aspect of space"* [Cohn and Hazarika, 2001]. Because topology knows only qualitative relations such as *region a is connected to region b*, it is often considered to be the most qualitative of all types of spatial relations. Topologic relations between spatial objects can be described without reference to the position, orientation, shape or size of the objects [Molenaar, 1998].

Which topological relations are possible between two spatial objects depends on their dimensionality. Between two point objects, only the *equality* and the *disjunction* relation hold. Between 2-dimensional regions (which are the dominant class of spatial objects considered in this work), a number of other topological relations are possible. Among these are obviously the equality and the disjunction relations, but in addition relations of *connection*, *overlap*, and *containment* are possible.

The evaluation of topologic relations plays an important role in geographic information systems. As we have seen above, standard buffer operations are based on the evaluation of the topological relations between spatial objects and a buffer zone (e.g., *"Select all objects that are contained within a 5 km buffer around location x"*). Likewise, the selection of spatial objects based on their topologic relations with respect to a region of interest (e.g., *"Select all polygons that overlap with the polygon p"*) is a standard operation in GIS.

In *QSR*, there are two distinct schools of thought that govern research about topology. They are different in their choice of primary primitives and formalizations, but were developed at approximately the same time and come to remarkably similar conclusions: On the one hand, a number of researchers with strong ties to the world of GIS and spatial databases built their theories on mathematical formalisms like point-set topology and algebraic topology. Following the approach of most mathematical theories of space which use points as primary primitives, and model extended spatial entities as sets of points, they define *"point-set topological spatial relations"* (e.g., [Pullar and Egenhofer, 1988, Egenhofer, 1989, Egenhofer and Franzosa, 1991, Egenhofer and Franzosa, 1995, Worboys and Bofakos, 1993]). They model geographic regions as point sets and distinguish between the *interior*, *boundary*, and *exterior* of a region. Using this representation, the set of topologic relations between two regions is given by all possible intersections of their boundaries, interiors, and exteriors. This approach resulted in the development of the so-called *4-intersection* and *9-intersection* models [Egenhofer and Franzosa, 1991,

Egenhofer et al., 1993].

On the other hand, there is a strong tendency within the *QSR* community to directly use *regions* rather than points as primary primitives [Cohn and Hazarika, 2001]. Such *pointless geometries* follow the paradigm that *"the spatial world consists of regions"* [Stell, 2004] which is based on a philosophical discussion about the representation of geographical objects and primitive spatial entities initiated by de Laguna [DeLaguna, 1922] and Whitehead [Whitehead, 1978]. Based on the work of Clarke [Clarke, 1981, Clarke, 1985], this led to the development of the so-called *Region-Connection-Calculus (RCC)* [Randell and Cohn, 1989, Randell et al., 1992].

Intersection Models

In this approach, a region x is represented by three sets of points - its interior (x°), its boundary (∂x), and its exterior (x^-). Between two regions a and b defined in this fashion, nine relationships, the so-called *9-intersection* $R_9(a,b)$ can be defined and expressed in a 3x3 matrix:

$$R_9(a,b) = \begin{pmatrix} a^\circ \cap b^\circ & a^\circ \cap \partial b & a^\circ \cap b^- \\ \partial a \cap b^\circ & \partial a \cap \partial b & \partial a \cap b^- \\ a^- \cap b^\circ & a^- \cap \partial b & a^- \cap b^- \end{pmatrix}$$

Setting the values of this matrix to empty (0) and non-empty (1), a total of $2^9 = 512$ binary topological relations are possible. This number, however, is reduced when the objects of concern are embedded in $\mathbb{R}^2$. For two homogeneous, 2-dimensional and connected regions with connected boundaries in $\mathbb{R}^2$, eight relations provide a mutually complete coverage [Egenhofer and Franzosa, 1991]. As it turns out, these eight relations correspond to the eight relations defined in *RCC-8*.

The Region-Connection-Calculus (RCC)

The primitive spatial entity in *RCC* is the *region*, and the primitive spatial relation is that of *connection*: Two regions a and b are connected $C(a,b)$ if their topological closures share at least one point. Out of this basic relation, sets of *jointly exhaustive and pairwise disjoint (JEPD)* relations were developed. Between a given pair of regions, one and only one of these relations will hold.

One of the most popular *RCC* calculi is *RCC-8* [Randell et al., 1992] which defines eight *JEPD* relations (Figure 4.2): DC ($D$is$C$onnected), EC (Externally Connected), PO (Partially Overlapping), TPP (Tangential Proper Part), $NTPP$ (Non-Tangential Proper Part), EQ (Equal), $TPPI$ (Tangential Proper Part Inverse), and $NTPPI$ (Non-Tangential Proper Part Inverse).

A smaller *JEPD* set of relations is given by *RCC-5* [Bennett, 1994]. In *RCC-5*, some of the relations defined in *RCC-8* are combined to form new relations (Figure 4.2): TPP and $NTPP$ are lumped together as PP (Proper Part), $NTPPI$ and $TPPI$ as PPI (Proper Part Inverse), and EC and DC as DR (Distinct Regions). The latter simplification means that the relation C

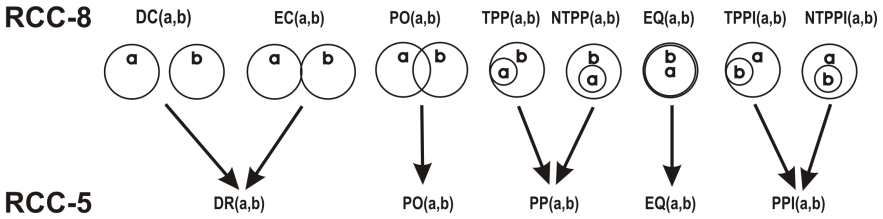

Figure 4.2: *RCC-8* and *RCC-5* relations (after [Cohn and Gotts, 1996a])

(*C*onnected) as a primitive is no longer available, and other relations like PP take its place [Cohn and Gotts, 1996a].

Conceptual Neighborhoods

If we consider two 2-dimensional regions that move in the plane (e.g., the action-radii of two moving robots) there is a certain natural order to the sequence of the possible topologic relations of these regions. For example, for two disconnected regions to become overlapping (i.e., two robots colliding), they have to be connected (however briefly) first. Based on previous work on the sequence of 1-dimensional (temporal) intervals [Freksa, 1992a], so-called *conceptual neighborhoods* for region-region relations expressed in *RCC-8* were developed [5]. They describe the sequence of topological region-region relations that has to be followed to get from one state to another (Figure 4.3).

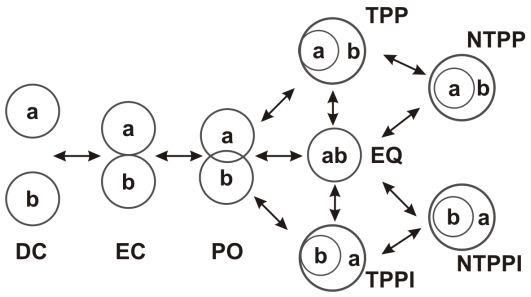

Figure 4.3: Conceptual neighborhoods of *RCC-8* relations

If we look at the resulting graph for the eight *RCC-8* relations, we can see that the ordering of the *JEPD* relations obtained by depicting the continuous transitions from $DC(a,b)$ to $NTPP(a,b)$ and $NTPPI(a,b)$, respectively, can be interpreted as an ordering of their connectedness: Two disconnected regions a and b ($DC(a,b)$) are further apart (i.e., less connected) than two regions that are externally connected ($EC(a,b)$), which are less connected than two regions that overlap ($PO(a,b)$). For the remaining four relations, the different levels of connectedness are less distinct. However, we may agree on stating that if a region a is *"only"* a tangential proper part of a region b or vice-versa

[5] Conceptual neighborhoods are not confined to region-region relations or RCC. [Egenhofer, 1995], for example, describes conceptual neighborhoods for topological line-region relations using the 9-intersection model.

($TPP(a,b)$ or $TPPi(a,b)$) they are somehow less connected than if a is a non-tangential proper part ($NTPP(a,b)$ or $NTPPi(a,b)$), or if both regions are equal ($EQ(a,b)$).

If we link the spatial relevance of two regions to their level of connectedness, the conceptual neighborhood graph of the topologic *RCC-8* relations can be translated into an ordering of spatial relevance. On one end, this ordering is delimited by the disconnection relation, on the other end by the equality relation. A partial ordering of topologic relations with respect to spatial relevance can be established, where $DC(a,b) < EC(a,b) < PO(a,b) < TPP/TPPi(a,b) < NTPP/NTPPi(a,b) < EQ(a,b)$.

Following this heuristic, we can for example say that two regions a and b, where a is contained in b ($TPP(a,b)$), are spatially more relevant than two regions b and c, where c is disconnected from b ($DC(c,b)$). However, we cannot make a judgement about which of two pairs of regions for which the disconnected relation holds (i.e., $DC(a,b)$ and $DC(a,c)$) is spatially more relevant because the topologic disconnected relation does not encode any information about the (quantitative or qualitative) distance of the regions involved. To fill in this gap it may be useful to look at alternative ways to encode adjacency and neighborhood relations.

4.2.4 Adjacency and Neighborhood

Topologic reasoning based on the 9-intersection model mentioned above is popular in GIS, spatial databases, and other fields where regions are represented as polygons. The general definition of a polygon is very similar to that of a region used in intersection models: Polygons have a *boundary* that is defined by a closed polyline. They have an *interior* (or *face*) that is the area enclosed by this polyline, and they have an *exterior* that is the Cartesian plane outside the polygon-boundary. The boundary itself consists of nodes, or vertices, connected by edges, or line segments.

For regions represented as polygons, Molenaar [Molenaar, 1998] gives a general definition of connection: Two spatial objects represented as polygons are topologically connected if they have some common geometric elements, i.e. common nodes, edges, or faces. The connectivity of spatial objects can be represented in a *connectivity graph*, where each node represents an object, and an edge between two nodes indicates that the respective objects are topologically connected.

In addition to connectivity, Molenaar defines *adjacency* for polygon-faces: Two faces f_1 and f_2 are adjacent at an edge e_i if their boundaries share that edge. This relationship can be represented in an *adjacency graph* where there is at most one edge between a pair of nodes, and where self-adjacency is not represented. Clusters of adjacent objects can be grouped to regions, and an adjacency graph can be drawn for these regions as well. In the case of such cluster-regions, two regions are called adjacent if they contain objects that are adjacent but do not contain common objects.

Applied to regular or irregular tessellations (see section 5.2 for a definition), the concepts of connectivity and adjacency can be used to define *neighborhoods* for spatial sub-regions of the tessellation. In general, two types of neighborhoods are distinguished [Gonzalez and Wintz, 1987, Molenaar, 1998] (Figure 4.4):

(Simple) Neighborhood: The (simple) neighborhood of a sub-region x of a tessellation consists of all sub-regions that are *connected* to x.

Full neighborhood: The *full* neighborhood of a sub-region x of a tessellation consists of all sub-regions that are *adjacent* to x.

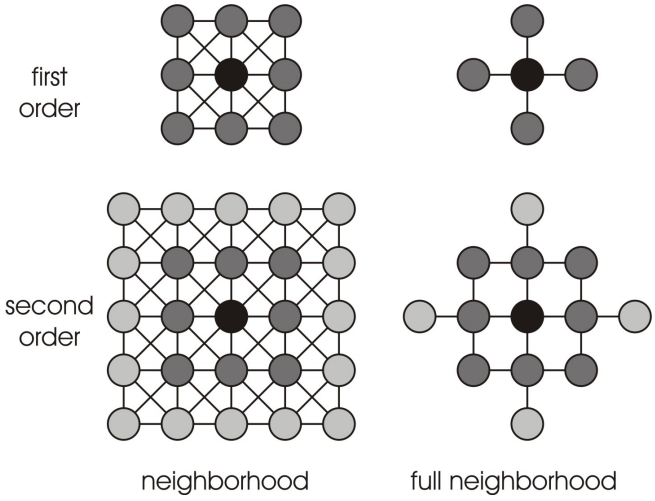

Figure 4.4: First and second order neighborhood, full neigborhood (after [Molenaar, 1998])

The direct neighbors of x are called *first-order* neighbors. The neighbors of the first-order neighbors of x are called *second-order* neighbors of x. In general, the neighbors of the $n^{th} order$ neighbors of x are called $n + 1^{th} order$ neighbors of x.

As we will see in section 7.1, the *neighborhood distance* of two regions derived from their order of neighborhood provides a semi-qualitative measure of distance that can be very useful for reasoning about spatial relevance.

4.2.5 Ordinal Relations

Ordinal information refers to an intermediate level of abstraction, between topological and metric information. In general, ordinal relations describe the location of points relative to a reference system. Orientation describes the relative position of objects in space. Unlike topological relations, orientation is not a binary relation of two spatial objects, but is used to express the orientation of a *primary object (PO)* with respect to a *reference object (RO)* within a common *frame of reference (FofR)* [Cohn and Hazarika, 2001].

Depending on which frame of reference is used, an *deictic*, an *intrinsic*, or an *extrinsic* view on orientation is taken. The deictic view describes orientation with respect to the originator of a spatial query. This is an interesting perspective for location based applications, where the *RO* (e.g., the user of a mobile navigation tool) wants to reason about the orientation of objects with respect to his or her current location. The intrinsic view of orientation uses intrinsic properties of the *PO*. For example, a ship has a *front* (the stern) and

a *back* (the aft). In the the extrinsic view, both PO and RO are embedded in an external, fixed frame of reference. Typical examples for extrinsic frames of references are coordinate systems and cardinal directions. The latter is one of the most frequently used *FofR* in Geographic space.

In the extrinsic frame of reference of cardinal directions, the orientation of two point locations on the surface of the Earth is measured as the angle α between the lines $RO - PO$ and $RO - NP$. NP signifies the location of the North Pole which is used as a standard global reference point. While quantitative descriptions of cardinal directions (i.e., angles measured in degrees) are typically limited to a few application domains (e.g., navigation), qualitative descriptions of cardinal directions are used rather frequently. They associate named directions with the main sectors of the unity circle, i.e. the named cardinal direction *"North (N)"* is associated with 0 degrees, *"East (E)"* with 90 degrees, *"South (S)"* with 180 degrees and so on. In most applications including spatial information retrieval, qualitative descriptions with four (N,E,S,W) or eight (N,NE,E,SE,S,SW,W,NW) named cardinal directions are used.

Because only distances that are in the same direction can be summed up, it is obvious that orientation is closely related to distance. In spatial information retrieval, evaluating orientation without information about distance (and vice-versa) does not make much sense. A query for *"a town north of Rome"*, for instance, may retrieve *"Perugia"* (at 180 km), *"Berlin"* (at 1500 km) and *"Oslo"* (at 2450 km), while obviously a closer location (i.e. *"Perugia"*) is much more relevant with respect to this query than a very distant location (i.e., *"Oslo"*).

There have been a number of calculi that combine distance and orientation. In a simple and straightforward approach, Frank [Frank, 1992] combined cardinal directions represented as compass segments with a simple distance metric able to express relative distances like *"far"* and *"close"*.

Freksa and Zimmermann [Freksa, 1992b, Zimmermann and Freksa, 1996] used two points and the line segment between them to describe the position of a third point. This description contains the position and the distance of the point relative to the line segment.

However, representations that adopt a deictic or intrinsic view and that use relative reference systems are needed for many reasoning tasks, for example in the domain of robot navigation. Schlieder applies a deictic view and triadic relations to describe the relative orientation of points [Schlieder, 1993]. He developed a calculus for reasoning about the relative orientation of line segments [Schlieder, 1995]. This approach can be used for qualitative navigation, reasoning about visible locations in navigation, and shape descriptions [Schlieder, 1996].

4.2.6 Partonomic Relations

Mereology

Mereology as the theory of parts and wholes and was initially introduced by Lesniewski [Lesniewski, 1916] as an alternative to set theory. In formal ontology, mereology has been used to model generic part-whole relations (e.g., [Simons, 1987, Varzi, 1996, Smith, 1998]). In qualitative spatial reasoning, mereology is often combined with topology to model spatial inclusions between regions (e.g., [Asher and Vieu, 1995, Smith, 1996, Pratt and Lemon, 1997]).

Such theories are referred to as *mereotopology* because they combine mereological notions like *"a is part-of b"* with topological relations like *"a is connected to b"*.

With respect to the representation of geographic objects, mereologic relations are particularly important because studies in cognitive science provide evidence that hierarchical concepts play an important role in the way humans organize knowledge, and in particular spatial knowledge [Hirtle and Jonides, 1985, Freksa, 1991, Fotheringham and Curtis, 1992]. In this context, *"hierarchization"* is a major conceptual mechanism to model the (geographic) world [Timpf, 1999]. The general idea behind hierarchization is to deduce knowledge at a coarse representation level without having to consider facts on more detailed representation levels. This approach reduces the amount of facts that have to be taken into consideration.

Administrative subdivisions are a good example for our cognitive preference to organize geographic objects along the lines of hierarchical partonomies: A district *is part of* a city, the city *is part of* a state, the state *is part of* a country. Such partonomic hierarchies form cognitive spatial models that can be used as a framework to geo-reference other geographical objects. For instance, spatial objects are often geo-referenced by defining their spatial relation with respect to an administrative unit. A park a may for example be said to be *part-of* a city b. If we know the relation of an object with respect to an administrative unit at a specific level of the hierarchy, we can draw conclusions regarding the object's relation to administrative units at other levels of the hierarchy: If park a is *part-of* city b, the fact that b is *part-of* country c implies that a is *part-of* c as well.

In geographic space, *spatial partonomies* are defined by the topologic relations between the (2-dimensional) regional extensions of the respective geographic objects: Park a is *part-of* city b if the regional extension of a is *contained-in* the regional extension of b. The close connection between mereology and topology has been pointed out by a number of authors, and theories have been developed to combine the two (e.g., [Smith, 1996]). Most authors see mereology as a sub-theory of topology [Randell and Cohn, 1989, Gotts, 1996], while others consider topology to be a domain-specific extension of mereology [Eschenbach and Heydrich, 1995]. In both approaches, the colocation of two regions implies that they also share parts. In a third approach, mereology is extended with topologic primitives [Varzi, 1994, Varzi, 1996]. This allows for the modelling of regions that are colocated but do not share parts.

The position of spatial objects within a spatial hierarchy of regions determines the implicit topological relations between the objects, and therefore their relative spatial relevance. For two regions a and b, where a is *part-of* a third region c, and b is not, the implicit topological relations are a *contained-in* c and b *disjunct* c. Therefore, a can be considered to be spatially more relevant to c than b.

We claim that for geographic objects which are geo-referenced within the framework of a hierarchy of regions (e.g., a partonomy of administrative units), spatial relevance is also influenced by the semantics of both the spatial query, the spatial objects to be retrieved, and the spatial partonomy itself. To illustrate this hypothesis we use the example of a tourist who is sunbathing on a beach and, around lunchtime, feels a strong desire to have some food. Suppose the beach

in question is along the shores of *Lake Constance* [6], and the boundary between Germany and Austria cuts right through the beach. We further assume that our tourist sits on the German side of the beach, while a number of restaurants and pubs are located on both sides of the border. For a spatial query with the intention to find a restaurant for lunch it is irrelevant which of the administrative units *"Germany"* or *"Austria"* the restaurant is contained in[7]. Consequently, two equidistant restaurants on either side of the border should lead to identical relevance ratings.

However, if the intention of the query were to apply for a job as a waiter, the same query for the same geographic object (i.e. *"restaurant"*) would probably yield a different result: Because the tax laws and working conditions are still different in both countries, our job-seeker might prefer the restaurants on one side or the other. In this case, the location of the geographic object within the spatial partonomy, i.e. whether the location of the restaurant is part of *"Germany"* or *"Austria"*, would play a role. This role is even more important in a query where our tourist is not interested in finding a restaurant, but a tax office. Obviously, because the public administrations in Austria and Germany are only responsible for their own citizens, the partonomic location of the next relevant tax office is very important and actually outweights it's geographic location: A tax office right across the border may be less relevant (or not relevant at all) than a tax office which is further away, but on this side of the border.

4.3 Uncertainty and Vagueness

4.3.1 Sources and Types of Vagueness

Because *"almost all the information that we possess about the real world is neither certain, complete, nor precise"* [Worboys, 1998], imprecision, uncertainty, and vagueness are phenomena that are endemic in almost every domain of knowledge representation [Cohn and Hazarika, 2001]. This applies in particular to the domain of geographic information, where one of the main problems associated with the representation of regional geographic objects is the *vagueness of region boundaries*[Burrough and Frank, 1996]. This vagueness may be due to the fact that the precise boundary of a regional geographic object is unknown (*epistemological vagueness*), or it may be associated with the inherent *ontological vagueness* of many geographic objects that is an expression of *"peoples conceptual understanding of the world as vague entieties"* [Montello et al., 2003]. In any case, when talking about geographic regions, and the representation of geographic regions in the context of information retrieval, we have to consider vague representations, and spatial reasoning on the basis of vague representations.

For the representation of 2-dimensional regions, the definition of the region-boundary plays an important role. In quantitative representations, the region boundary is typically represented as a closed polyline. In qualitative representations, the region boundary is given either explicitly (e.g., in the 9-intersection model), or implicitly (e.g., in *RCC*). Both the quantitative and the qualitative

[6] Lake Constance is a large water body on the boundary between Germany, Switzerland, and Austria

[7] There are no border controls between Austria and Germany anymore, and both countries belong to the EURO zone

representations mentioned so far assume regions to be crisp partitions of space, with crisp region boundaries. This view fits very well with the mathematical concepts and the data models needed to process spatial information in a computer. As a consequence there has been a strong tendency to force reality into crisp representations, i.e. to use crisp region-boundaries even where regions are vague and the exact location of their boundaries is unknown [Schneider, 1999].

This approach has been only partially successful. In the domain of spatial databases, some of the problems with this approach can be summarized under the concept of *data quality*. In general, there may be a number of different reasons why the boundaries of geographic objects and regions are imprecise, uncertain or vague [Worboys, 1998] [Cohn, 1999]:

Imprecision caused by measurements and representations: Imprecise representations of geographic objects are often caused by inaccurate and erroneous measurements, but also the limitations of computers with respect to the representation of point coordinates. This causes the deviation of measured values (e.g., the coordinate position of polygon vertices) from their *"true"* positions in geographic space.

This problem has long been recognized by the GIS community, and there exists a large amount of related literature that describes solutions to the problem (for an overview, see [Hunter and Goodchild, 1993]). Some authors, however, suggest that digital representations of geographic regions are intrinsically imprecise because the location of a point (and therefore the exact location of a region boundary) can never be measured and represented with unlimited precision [Montello et al., 2003].

A slightly different notion of imprecision is related to the resolution, or granularity, at which a spatial observation is made [Worboys, 1998]. It determines how much detail of a spatial object can be recorded, and in which detail this recording can be represented.

Uncertainty through incomplete information: A lack of information about a spatial object can be a major source of uncertainty about its exact borders. In some cases, there may simply be not enough information (e.g., measurements) available to delineate the exact boundary of a spatial object, or region. We already mentioned the example of an oilfield, of which the crisp boundary cannot be determined due to a lack of sampling data (see section 3.1.3).

Other examples of incomplete information pertain to applications where dynamic scenarios have to be modelled. This is the case in emergency response applications, logistic networks, but also scenarios that exhibit regular temporal variations (e.g., a coastline under tidal influence, or a river changing its path).

Intrinsic vagueness of geographic objects: The vagueness of the regional extent of a spatial object is often the result of an intrinsic ontological vagueness, i.e. the vagueness of the concepts that are used to describe the data. In section 3.1.3 we have seen that many place names exhibit an intrinsic locational vagueness due to the intrinsic vagueness of the geographic objects and regions they are used to reference. Typical examples are place names that refer to "man-made" regions exist as the result of

some (arbitrary) political process, and that often do not have clearly defined boundaries (e.g. *"Southern England"*, *"Northern Germany"*).

Another type of vague region boundaries is caused by field variation. If the values of a parameter that is used as an indicator to separate two regions changes gradually, it is often impossible to determine an exact boundary between these regions. A typical example soil type classifications. Here, a soil type A may slowly diffuse into another soil type B, and the exact boundary between A and B is somewhat arbitrary. Note that this is true even in the (theoretical) case where the soil composition is known throughout the whole area under consideration.

Other authors (e.g., [Schneider, 1999]) distinguish only between objects with sharp boundaries for which position and shape are unknown or cannot be measured precisely, and objects with boundaries that are not well-defined due to the intrinsic properties of the object itself. In case of the first category, *uncertainty* relates either to the lack of knowledge about position and shape of an object with an existing, real boundary (*positional uncertainty*), or to the inability to measure such an object precisely (*measurement uncertainty*). For the second category, the term *fuzziness* is used to describe the intrinsic vagueness of an object which has a certain spatial extent, but which inherently does not have boundaries that can be defined precisely.

4.3.2 Methods to Handle Vagueness

The need for a formal representation of imprecise, uncertain and vague regional (geographic) objects, and methods for spatial reasoning based on such representations, plays an important role in a number of applications, including GIS, spatial databases and gazetteers. In response to this need, a number of different theories to handle such regions have been developed. Based on their governing principals, these approaches can be grouped into different families of models [Erwig and Schneider, 1997, Borgida et al., 1989, Kulik, 2003]:

Probabilistic Models

Probability theory can be used to represent the vagueness of spatial regions that is caused by the uncertainty of positional and measurement information (e.g., [Burrough, 1996], [Shibasaki, 1993]). Probabilistic models define the grade of membership of an entity in a set by a statistically defined probability function. Examples are the uncertainty about the spatial extent of field-based entities, like regions defined by some property such as temperature, or the water level of a lake.

Fuzzy models

A rather popular semi-quantitative approach to represent vague spatial regions is provided by fuzzy set theory. Fuzzy sets were first introduced by Zadeh [Zadeh, 1965] with the intention to model imprecise concepts in a definable way. Fuzzy set theory is an extension or generalization (and not a replacement) of classical boolean set theory and deals only with fuzziness, not with uncertainty. Fuzziness is not a probabilistic attribute, in which the grade of membership of

an individual in a set is connected to a given statistically defined probability function. Rather, it is an admission of the possibility that an individual is a member of a set or that a given statement is true.

In fuzzy set theory, the uncertainty that a location belongs to a a region is modelled by a real number between 0 and 1: A membership value of 0 indicates that the location is definitely not in the region. A value of 1 indicates that the location definitely is in the region. The magnitude of an intermediate value indicates the level of certainty that the location is in the region.

Fuzzy set theory was applied to spatial data by a number of authors (e.g., [Altmann, 1994, Dutta, 1989, Stefanakis et al., 1996]) for a number of different applications. Some authors used fuzzy set theory to create regional representations of point data [Brown, 1998], while others developed fuzzy methods to model imprecise and qualitative spatial relations between geographic objects [Guesgen and Albrecht, 2000]. In a series of papers, Schneider developed fuzzy spatial data types, including fuzzy points, fuzzy lines, and fuzzy regions [Schneider, 1999, Schneider, 2000, Schneider, 2001, Schneider, 2003].

Qualitative Models Based on Crisp Regions

Within the domain of Qualitative Spatial Reasoning, several approaches to model vague regions with the help of exact, crisp regional representations have been developed [Clementini and di Felice, 1996], [Cohn and Gotts, 1996a], [Schneider, 1996]. These methods have in common that they model a vague region as a combination of multiple crisp regions. The result is a zoned model of the vague region, typically with a inner region and an outer region. The inner and the outer region enclose the *"real"* (but indeterminate) boundaries of the spatial object. They therefore represent an approximation of the minimal and the maximal regional extension of the object.

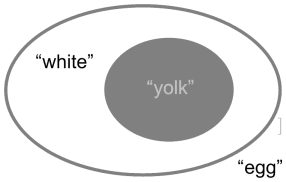

Figure 4.5: The egg-yolk approach

One major advantage of these approaches is that they build upon existing and well-established concepts for the qualitative representation of and reasoning with crisp regions. The so-called *"egg-yolk"* calculus [Cohn and Gotts, 1996a], for example, is an extension of the well-known *RCC-8* calculus (see section 4.2.3). In the egg-yolk theory, vague regions are modelled as a pair of crisp regions: the *"yolk"*, representing the vague regions minimum extension, and the *"egg"*, representing its maximum extension (Figure 4.5).

The egg-yolk representation can be used to describe topological spatial relations between vague regions. On the basis of *RCC-8*, a total of 252 *JEPD* topological relations can be defined between two vague egg-yolk regions [Cohn and Gotts, 1996b]. Through application of constraints imposed by an embedding of the regions in $\mathbb{R}^2$, the number of valid relations can be reduced to 46 (Figure 4.6).

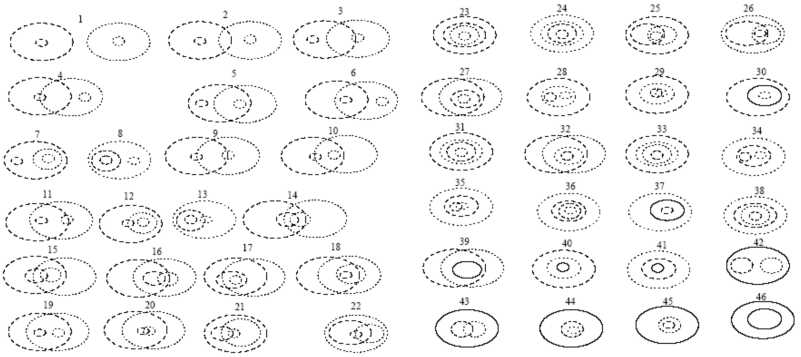

Figure 4.6: The 46 possible topological relations between egg-yolk regions (after [Cohn and Gotts, 1996a])

Analogous to *RCC*, the 9-intersection calculus described above can be extended to allow for the representation of vague regions [Clementini and di Felice, 1996, Clementini and di Felice, 1997]. The result is a calculus of *regions with broad boundaries*. Among those regions, 44 JEPD topologic relations can be distinguished and grouped into a conceptual neighborhood consisting of 18 sets.

Models Based on Discrete Spatial Partitions and Rough Set Theory

Like fuzzy models, models based on rough set theory employ a semi-quantitative representation of vague regions. However, instead of defining a membership function, rough set theory extends the concept of mathematical sets by introducing a boundary region of a set.

In the context of this work, rough set theory is particularly interesting because it is well suited to represent vague data (i.e., vague regions) on the basis of finite spatial partitions. In the following section, we will therefore describe rough set theory and its application to the representation of spatial regions in more detail.

4.4 Discrete Spatial Representations

4.4.1 Continuous and Discrete Space

Most of the methods for the representation of crisp and vague regions described so far are based on a notion of *continuous space*. A continuous space makes it easier to axiomatize the respective spatial relations. However, such axiomatizations are difficult or impossible to use for practical applications. Several authors addressed this problem by developing representations based on finite partitions of space, i.e. *discrete spaces* [Kaufman, 1991, Galton, 1999, Masolo and Vieu, 1999]. Because the representation of a region in discrete space always involves a certain approximation of the region, discrete representations are generally well suited to model vague regions. The methods presented below are therefore closely related to the representation of vague spatial objects.

When talking about discrete spaces and the approximation of regions in discrete space, we also have to consider the representation of information on different levels of granularity, a problem which has been addressed by a number of authors in the *QSR* community [Euzenat, 1995, Bittner and Stell, 1998], as well as within a GIS context [Timpf and Frank, 1997, Worboys, 1998, Stell and Worboys, 1998, Stell and Worboys, 1999, Stell, 2000]. In the following, we will examine approaches for the representation and approximation of regions using discrete spaces.

4.4.2 Extending *RCC-8* to Discrete Space

The Region Connection Calculus *(RCC)* is based on the assumption of continuous space. Regions in *RCC* do neither have a specific location, nor a specific extension. Using a set of *JEPD* relations, the topological relationship between two regions can be modelled in a purely qualitative manner.

While this approach is very useful to reason about topological relations on a highly abstracted level, it is not well suited for the computation of topological relations between potentially many spatial objects represented on the less abstracted level of, for example, a computerized digital map. To bridge the gap between high-level qualitative approaches and lower-level quantitative methods is the intention of the work of Galton [Galton, 1999]. In an attempt to develop a theory of discrete space that parallels established theories of topology and mereotopology of continuous space, he outlined a discrete-space version of the well-known *RCC-8* relations, which he calls *RCC-8D*.

The *RCC-8D* theory uses cells (instead of regions) as spatial primitives, and adjacency (instead of connection) as the primitive relation. Galton defines discrete regions as aggregates of cells, and adds a second primitive relation, namely containment of a cell in a region. The existence of a null region $\emptyset$ is admitted, and there exists a universal region U that contains all cells. In practical applications, good choices for U are (regular or irregular) tessellations, such as rectangular grids or a Triangulated Irregular Networks *(TINs)*.

Galton applies the eight *JEPD* relations of *RCC-8* ($DC(a,b)$, $EC(a,b)$, $TPP(a,b)$, $PO(a,b)$, $NTPP(a,b)$, $TPPI(a,b)$, $NTPPI(a,b)$, and $EQ(a,b)$) to discrete space. It turns out that their definitions in discrete space do not diverge very much from the respective definitions in continuous space. *Qualitative continuity*, however, which is an important property of the *RCC* logic, and is typically depicted in *conceptual neighborhood* diagrams, seems to be much less distinct in *RCC-8D* than in *RCC-8*. Galton therefore concludes that *RCC* relations may be less useful as descriptors of discrete space than of continuous space.

Other properties of discrete space, however, are very useful in the context of practical applications. One of them is the natural metric which is intrinsic to discrete space. Between two cells x and y, a path from x to y can be defined as a sequence of adjacent cells. The length of the path is given by the number of cells in the path. The distance $\delta(x,y)$ is defined as the path with the least cells, i.e. the shortest path between x and y. We will see in section 7.1 that this kind of semi-qualitative distance is very useful for spatial relevance reasoning.

4.4.3 Rough-Set Approximations of Regions

Spatial resolution is fundamental to many aspects of the representation of spatial data. Multi-resolution data models and database interaction languages are often seen as pre-requisites for spatial data integration. With the intention to provide a formal framework for multi-resolution geographic spaces, Worboys [Worboys, 1998] describes a theory of spatial imprecision arising from observations of spatial entities and relationships at multiple finite resolutions. He discusses an approach to handling and reasoning with the uncertainty that results from representations of spatial entities at different finite resolutions. He uses techniques that are very similar to the mathematical theory of *rough sets*.

Rough-set theory was introduced by Pawlak [Pawlak, 1982, Pawlak, 1984, Pawlak, 1991, Pawlak, 1993], and is a technique mainly used to manage uncertainty in databases. The basic idea behind rough-set theory is that spatial entities can only be perceived by making observations about them. Depending on the granularity or resolution at which a specific observation is made it provides information at different degrees of precision and accuracy. Here, the term granularity refers to collections of elements that are indiscernible from each other. The lower the granularity, the better we can discern differences between elements.

Following rough-set theory, Worboys assumes for any observation an indiscernibility relation ρ on a set S, where $t\rho s$ can be read as: *t is indiscernible from s*. An indiscernibility relation ρ on set S leads to a collection of subsets of S:

$$R(s) = t \in S | t\rho s \tag{4.2}$$

Because ρ is interpreted as an equivalence relation, sets of the form $R(s)$ for $s \in S$ form a partition of S. The set of equivalence classes of S with respect to ρ is written S/ρ. Worboys uses the above definition of indiscernibility to define a lower approximation $L(T)$ and an upper approximation $U(T)$ of a set T with respect to S/ρ.

$$L(T) = x \in S/\rho | x \subseteq T \tag{4.3}$$

$$U(T) = x \in S/\rho | x \cap T \neq \emptyset \tag{4.4}$$

The upper and the lower approximation of T bound the *"true"* extent of T, i.e. $L(T) \subseteq T \subseteq U(T)$. Only if the set T can be defined precisely, or crisply, with respect to ρ both approximations are equal, i.e. $L(T) = U(T)$.

4.4.4 Rough Location

The indeterminacy of the location of spatial objects in geographic space, i.e. their *locational vagueness*, is closely related to the vagueness of the objects themselves, i.e. to the vagueness of the underlying human concepts. Because location is an important component of geographic information, concepts to

represent and process vague locational information with the help of a *"rough location"* were developed by Bittner and Stell [Bittner, 1999, Bittner, 2000, Bittner and Stell, 2002]. Like the approach described in the previous section, rough locations use a discrete partitioning of space that is based on general rough-set theory.

The theory of rough location distinguishes between the *exact location* of an object in space and the *approximate location* of the object with respect to regions forming a partitioning of space. Following an approach by Casati and Varzi [Casati and Varzi, 1995], the exact location of a spatial object is defined as *"the region of space taken up by the object"*. Consequently, the notion of *exact location* relates spatial wholes (i.e. spatial objects) to regional wholes. The relation of parts of spatial objects to parts of regions of space is called *part location*. The *rough location* of a spatial object defines its part location with respect to a set of regions forming a *regional partitioning of space*.

Using a set of part location predicates taken from [Casati and Varzi, 1995], three *JEPD* spatial predicates are defined, namely *fully located (FL)*, *overlapping located (OL)*, and *non overlapping located (NL)*. These predicates are the basis for the formalization of rough location. The rough location of a spatial object is assigned through a set of *location mappings*. For a given spatial object, location mappings return approximations of the object in terms of its relationship with respect to the regions of a regional partition. These approximations correspond to what Bittner calls the *overlap sensitive*, *containment sensitive*, and *overlap and containment sensitive* rough location of a spatial object. The method supports the definition of operations on rough locations, such as union and intersection operations, which can be used to formalize binary topological relations between approximated vague objects.

Part II

Qualitative Representation and Reasoning with Place Names

Chapter 5

A Discrete Representation Framework

5.1 Representations of 2-D Spatial Regions

5.1.1 Qualitative Representations

Among the main intentions of this work is the development of a practical framework for spatial relevance reasoning based on simple and intuitive representations of geographic locations and geographic objects. In the domain of information retrieval, such a framework can be used to add the spatial dimension to *"intelligent"* queries, i.e. to queries which are based on an evaluation of the semantics of information items rather than simple string matching algorithms. Such integrated intelligent queries are of the general type *concept* **@** *location* ***in*** *time*, indicating that they address the *conceptual*, the *spatial*, and the *temporal* semantics of an information object.

In section 4.1 we discussed the concept of spatial relevance. We identified a number of spatial relations that can be used to establish a ranking of locations, geographic objects, and the associated information objects based on their relative relevance with respect to a spatial query. Among these relations are *topological* and *mereological* relations, as well as *distance* relations.

One option to implement a system for spatial information retrieval is to use standard tools like Geographic Information Systems (GIS), spatial databases, or digital gazetteers. These tools use quantitative representations of regional geographic objects in Euclidean geometry and Cartesian coordinate spaces. They apply methods rooted in computational geometry to evaluate spatial relations between spatial objects. As we outlined in section 4.2, these methods are very useful in many application areas that require an efficient handling of exact quantitative representations of spatial objects. However, applied to the domain of information retrieval in heterogeneous and distributed environments (which is the focus of this work), quantitative representations and computational methods do have a number of shortcomings (section 4.2).

To develop an intuitive and user-friendly framework for the conceptualization of geographic space, and the effective reasoning about the spatial relevance of information objects, we have to overcome the limitations of quantitative rep-

resentations and algorithms. As a solution we propose to use qualitative or semi-qualitative representation schemes together with suitable qualitative reasoning methods. Such representations and methods have the advantage that they treat spatial phenomena in a way that is much closer to human spatial reasoning than quantitative representations are. Also, qualitative representations are often more compact than comparable coordinate-based representations because they concentrate on the representation of essential properties of space and spatial objects, and eliminate redundant information. This makes qualitative representations interesting for applications in heterogeneous and distributed environments, for which compact and interoperable representations are a prerequisite.

In section 4.2, we reviewed some of the methods that were developed in the domain of Qualitative Spatial Reasoning *(QSR)*. Qualitative formalizations exist for a broad range of spatial relations, including topological, mereotopological, ordinal and metric relations, as well as for the representation of crisp and vague regions.

5.1.2 Semi-Qualitative Representations

Non-withstanding their potential benefits, it is often difficult to integrate purely qualitative representations into practical *"real-world"* applications: the notion of continuous space, which is used by many high-level qualitative theories, is not compatible to the discrete spaces used by many physical recording devices [Galton, 1999].

In section 4.4, we reviewed a number of *"semi-qualitative"* representations that were developed to overcome this problem. These representations have in common that they replace continuous space with *discrete partitionings* of space as the basic framework of spatial reference. For practical applications, discrete spaces do have a number of advantages: While in continuous space, metrics *"do not arise out of the mereotopological structure but have to be defined separately, discrete space comes endowed with a natural metric"* [Galton, 1999]. Here, Galton refers to the semi-quantitative distances given through neighborhood and adjacency relations of discrete partitionings.

Another advantage is that spatial objects can be geo-referenced indirectly through references to parts of a discrete space. Using such *spatial indices*, the data volume needed to represent a spatial object can be considerable reduced.[1]

5.1.3 Spatial Relevance Reasoning in Discrete Space

In chapters 2 and 3 we have seen that a large percentage of geo-referenced information available through distributed networks (i.e., the Semantic Web), large information repositories (i.e., digital libraries and metadata information systems), and even spatial data infrastructures specifically designed for the management of geospatial data is indirectly geo-referenced, often with the help of place names. Place names are natural language terms that are used to reference and identify geographic objects (both physical and non-physical) in geographic space. They provide a natural, user-friendly and, from a cognitive perspective, sound method to conceptualize geographic space. Accordingly, place names are

[1] For this reason, spatial indices are widely used to support spatial queries in spatial database applications.

frequently used in conjunction with spatial metadata, and to specify spatial queries.

We have seen in section 3.2 that (standardized) place names are organized in place name lists, the so-called *"gazetteers"*. Within the last decade, digital gazetteers have become important tools for the access to geo-referenced information, and they already found their way into many information systems. In a gazetteer, the spatial extent of a geographic object is represented as a *spatial footprint*. This spatial footprint is used to reference a place names to a geographic location and is the basis for spatial reasoning [Hill, 2000], [Riekert, 1999]. However, due to the simple spatial encoding of they use, most state-of-the-art gazetteers provide only limited spatial reasoning capabilities.

In this work, we will present a practical solution to spatial relevance reasoning that is based on discrete partitionings of (continuous) geographic space. They will provide the framework for qualitative representations of geographic objects as *qualitative spatial footprints*. In section 5.2, we discuss different types of discrete partitionings that can be used to build *qualitative spatial reference models*.

Compact and interoperable representations, as well as efficient reasoning mechanisms are needed for many *"spatially-aware"* applications in distributed heterogeneous systems and/or light-weight clients (e.g. in mobile location-based applications). We will therefore develop a method to encode qualitative spatial reference models and qualitative spatial footprints as graph-based abstractions (section 5.2.3). Such abstractions do have the advantage that they are a very condensed and light-weight data format and at the same time support standard graph-algorithms which can be used to build efficient reasoning tools.

In section 6.1, we formalize the mapping functions that can be used to define qualitative spatial footprints for regional geographic objects with both crisp and indeterminate boundaries on basis of these abstractions.

In section 6.2, we develop an approach to organize place names and their qualitative spatial footprints in hierarchical *place name structures*. Place name structures are designed to manage both standardized and non-standardized place names and provide intuitive and user-friendly tools to conceptualize and model geographic space. We look at the cognitive aspects of spatial modelling and review some of the empirical work that has been done to describe the use of preferred mental models for spatial modelling. Some of these results are integrated into the architecture of place name structures.

In section 7.1, we integrate graph-based abstractions and reasoning methods with the requirements for spatial relevance reasoning discussed in section 4.1. The result is a simple metric to compute relative spatial relevancies between the units of a qualitative spatial reference model.

In section 7.2, we extend the metric developed to compute the spatial relevance between the units of spatial reference models to reasoning about the relative spatial relevance of place names in a place name structure. Finally we show how this method can be used to integrate multiple heterogeneous place name structures. We prove our concept using a number of real-world examples and use-cases (section 6.3 and chapter 8).

5.2 Discrete Representations of Spatial Regions

5.2.1 Discrete Space and Spatial Indices

Calculi that use qualitative representations of regions (e.g., the *RCC* calculi) are based on the notion of space as a continuum. Regions defined in such spaces either consist of an infinite set of points, or mere representations of boundaries. Such representations are very well suited to reason about general relations of regions. They cannot, however, be used to solve problems in *"natural"* spaces, like geographic space.

Quantitative models of geographic space like the Euclidean space are theoretically continuous, but practically limited by the precision of the underlying digital representation. Here, regions are defined through polygons, i.e. finite point sets delimited by vertices and edges. Discrete spaces are representations that incorporate elements of both models.

Our notion of discrete spaces follows a definition that was put forward by Galton [Galton, 1999] and described in more detail in section 4.4. He defines a discrete space to consist of a finite region of space, called *universal region* U, which can be subdivided into a finite set of smaller regions, or *cells*. The primitive spatial relation between these cells is that of *adjacency*. [2]

To be useful for practical applications, Galton proposes a number of additional conditions that a suitable discrete space should fulfil:

1. Every cell should have a finite number of neighbors (i.e. adjacent cells),
2. every cell should have the same number of neighbors[3],
3. the universe should be self-connected,
4. between any two cells there should be a unique least path, and
5. the universe U should contain finitely many cells.

Regular and Irregular Tessellations

Regular *tessellations* are among the discrete spaces that fulfill the requirements stated above. Following a definition given in Worboys [Worboys, 1995], a tessellation is a *"partition of the plane or portion of the plane as the union of a set of disjoint areal objects. (...) A tessellation of a surface is a covering of the surface with an arrangement of non-overlapping polygons"*. A *regular tessellation* consists only of regular and equal polygons [4].

There is only a limited number of completely regular tessellations in the 2-dimensional Euclidean plane: Orthogonal grids, hexagonal grids, and orthodiagonal grids (Figure 5.1 a, b, and c). Among these, orthogonal grids which partition space into equal squares are the most well-known. Such *rasters* are used by many practical applications to discretize space. The pixels of a computer screen are a prominent example. The number of *irregular tessellations*, on the other hand, is unlimited. The most frequently used irregular tessellation,

[2] Galton notes that this view of discrete space is very closely related to Graph Theory, a fact which will prove useful for the encoding of our spatial reference model.

[3] We will describe later suitable tessellations where this constraint may be relaxed.

[4] In a regular polygon, the length of all edges and all internal angles are equal.

the *Triangulated Irregular Network (TIN)*, uses triangles of different sizes and orientation to tessellate space.

A special type of regular tessellation often used in spatial databases and GIS is the so-called *regional quadtree* [Samet, 1984], [Samet, 1989]. In contrast to other regular tessellations regional quadtrees have a variable resolution. This is achieved by a recursive subdivision of individual (square) cells in the segment of the quadtree where refinement is needed. The decomposition of the grid cells into smaller parts is not arbitrary but follows specific rules: A cell can only be divided into four equal sub-cells. The four subcells of a subdivision can be enumerated in a certain, well-defined order which provides an unambiguous identifier for each cell.

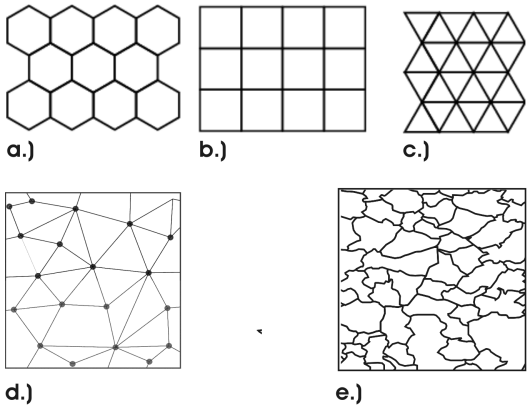

Figure 5.1: Regular (a. hexagonal b. rectangular c. triangular) and irregular (d. triangular e. polygonal) tessellations

Tessellations as Frame of Reference for Spatial Indices

According to Galton's definition, a discrete *region* is an aggregate of cells of a spatial partition. Thus a region can be explicated through references to specific cells of a tessellation. Such references, or spatial indices, are used in GIS and spatial databases as well as in a small number of state-of-the-art gazetteers. In the former, they provide a functionality similar to that of indices in standard relational databases: They support and improve the performance of data access algorithms. In the latter, spatial indices are used to geo-reference geographic objects denoted by place names (see section 3.2).

Tessellations have a number of properties that make them well-suited as a frame of reference for the representation and reasoning with spatial regions:

Natural metric: Because the adjacency relation is transitive, a distinct path can computed between every two cells of a tessellation. In regular tessellations, this path distance provides a qualitative approximation of the Euclidean distance between the centroids of two cells. In inhomogeneous irregular tessellations, the path distance may deviate from the Euclidean distance as a function of the distribution of cell sizes.

Representation of complex regions: Spatial indices based on tessellations

allow for the representation of complex regions without the need for exact, coordinate-based polygonal data.

Representation of vague regions: The discrete representation of a region is by definition an approximation of the regions's true spatial extent. Therefore representations using spatial indices are well-suited for the representation of vague regions.

"Artificial" vs. "Natural" Tessellations

The regular and irregular tessellations discussed so far have in common that they are arbitrary subdivisions of geographic space: All cells are either uniformly shaped and sized (e.g., in regular orthogonal grids), or their subdivision follows strict mathematical rules (e.g., in quadtrees and TINs). In any case, the spatial partitionings created by such tessellations do not have any relation to (natural or artificial) geographic features. They provide no information about the underlying geographical and organizational features. Accordingly, the naming-scheme of such cells is often designed exclusively to be processed by machines, using cell identifiers that do not have any relation to objects or regions in the underlying geographic space.

It is obvious that such *"artificial"* tessellations do not provide an intuitive and user friendly framework for the specification of spatial indices. Without the help of tools that offer a combined visualization of geographic objects and reference grid, a human user will not be able to use an artificial tessellation as a referencing system. The grid cells of a city map, for example, are typically referred to as something like *Quadrant A5*, or *Cell XZ35-2*. Without using a map, or in case of an online system, a map-based graphic user interface, it is virtually impossible to manually define which grid cells can be used as spatial indices for a given geographic object.

In distributed and heterogeneous applications, the usefulness of artificial tessellations as a spatial frame of reference is also limited by their *lack of interoperability*. The regular reference grid used by one gazetteer, for example, is most likely different from the reference grid used by another gazetteer. Artificial reference tessellations can therefore be viewed as geo-referencing standards that cover a specific and limited subsection of geographic space (e.g., a city, a country, a map-quadrant). In most cases, such tessellations do have a fixed resolution that is optimized to represent spatial objects within a specific size-range[5].

In general, all regions that are approximated on basis of one specific tessellation have to follow the same standard. Only regions that are discretized using the same standard can be compared to each other. Unfortunately these standards are not global, but are typically used only within a specific (and rather limited) user group or application[6]. As in any standardization process, unless a generally accepted standard reference grid is established (which is unlikely

[5]Most gazetteer applications use reference grids with a fixed resolution. Each reference grid is optimized only for a certain range of scales which means that only geographic objects that fall within a certain size-range can be represented optimally. Two objects that are smaller than the grid resolution cannot be distinguished. In case of the GEIN gazetteer, only objects that have a diameter larger than 3 km can be distinguished. On the other hand, objects that are much larger than 3 km in diameter will have to be represented by an excessive number of grid cells.

[6]One of the few exceptions is the global grid of geographic latitude and longitude.

because of a number of technical and organizational problems), different user groups will continue to define their own reference grids tailored to meet their individual needs and intentions. This leads to the simultaneous use of many different and incompatible reference grids, which poses a serious problem for the exchange of geo-referenced data and information between user communities. The problem is particularly evident in applications like the Semantic-Web that are intended to provide access to distributed and heterogeneous data sources.

In the following section, we will evaluate a special group of irregular tessellations, namely *polygonal tessellations*. As we will see below, specific types of polygonal tessellations can be referred to as *"natural"* tessellations. In a "natural" tessellation, the outlines of the individual polygons have a well-defined relation to the underlying features of geographic space. By using such "natural" polygonal tessellations to create spatial indices, we hope to overcome some of the shortcomings of the regular and irregular tessellations described above.

Polygonal Tessellations

A polygonal tessellation is an irregular tessellation where the cells constitute of irregularly shaped polygonal regions. Polygonal vector data are frequently used in geographic information systems (GIS) to represent geographic objects in 2-dimensional Euclidean space $\mathbb{R}^2$. In this context, a (simple) polygon in $\mathbb{R}^2$ is defined as the area enclosed by a simple closed polyline that represents the boundary of the polygon and consists of a finite set of line segments, or *edges* [Worboys, 1995]. The end-points of the edges are called *vertices*(Figure 5.2).

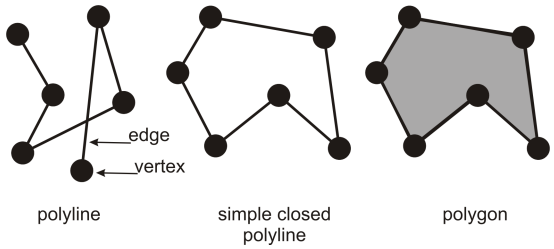

Figure 5.2: Polylines and polygons (after [Worboys, 1998])

In this work, we use a definition where a polygon is a closed sets of points, i.e. edges and vertices belong to the polygon. In [Schlieder et al., 2001] we gave a general account of polygonal arrangements in $\mathbb{R}^2$: If we the polygons $P_1, ..., P_n$ belong to a part of the plane that is bounded by a polygon P, two types of arrangements of the polygons within the containing polygon P can be distinguished:

- A *polygonal covering* $P = P_1 \cup ... \cup P_n$, where a set of generally overlapping polygons covers the whole area of the containing polygon.
- A *polygonal patchwork interior*, where $(P_i \cap P_j) = 0$ for all $i \neq j$ from $\{1, ..., n\}$. In a patchwork, the polygons are either disjoint, or intersect only in edges and/or vertices.

The polygonal tessellation T is a special, but very common type of arrangement. It can be defined as a polygonal covering that also forms a polygonal

patchwork. With respect to the topological relations between the polygons, a tessellation *"is a covering of the surface with an arrangement of non-overlapping polygons"* [Worboys, 1995].

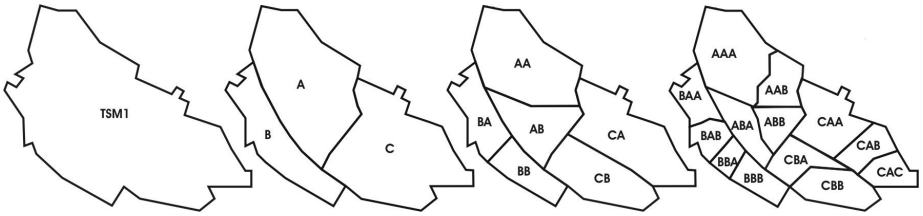

Figure 5.3: Simplified example of a polygonal standard reference tessellation

If we decompose a polygonal covering, patchwork, or tessellation P into its components $P_1, ..., P_n$, we obtain a decomposition which, analogous to a partonomy, is a hierarchical data structure for encoding the *spatially-part-of* relation $\sqsubseteq$ together with the type of arrangement of the parts.

For the special case of a polygonal decomposition by tessellation, we can define the relation $\mathrm{T} \sqsubseteq \Pi \times 2^{\Pi}$, where Π denotes the set of polygons in the plane and $\mathrm{T}(P, \{P_1, ..., P_k\})$ iff $\{P_1, ..., P_k\}$ is a tessellation of P. Using this relation, we can say that a polygon P_1 is *spatially part-of* a polygon P_2 iff P_1 is part of the decomposition by tessellation of P_2, i.e.,

$$P_1 \sqsubseteq P_2 \Leftrightarrow \mathrm{T}(P_2, \{..., P_1, ...\}) \tag{5.1}$$

Applied to the decomposition hierarchy shown in 5.3 we can say for example that $\{AA, AB\} \sqsubseteq A$, and $\mathrm{T}(A, \{AA, AB\})$.

As mentioned above, we refer to polygonal tessellation as *"natural"* tessellations, indicating that the polygons of the tessellation represent meaningful subdivisions of geographic space rather than arbitrary geometric figures. In fact, while they share most of their mathematical properties, what distinguishes such "natural" polygonal tessellations from other irregular tessellations is their intrinsic semantics. In "natural" polygonal tessellations, the individual polygons may represent the spatial extent of administrative subdivisions, postal code zones, climate census districts and other geographic features.

However, to be useful as a generally understood frame of reference, a suitable polygonal tessellation must also have a large user community, and its semantics (i.e., the names and the classification of the individual polygons) has to be well-known. In this work we focus our attention on a relative small number of widely-used and well defined polygonal tessellations, which we call *polygonal Standard Reference Tessellations (pSRT).*

5.2.2 Polygonal Standard Reference Tessellations

Polygonal Standard Reference Tessellations (pSRTs) are tessellations that have a well-known and standardized semantics, i.e. they partition geographic space into a set of well-defined polygonal regions. Tessellations of man-made, or *"fiat"* [Smith and Mark, 1998] partitions of geographic space, such as administrative subdivisions, postal code areas, or census districts, provide the best candidates for pSRTs. They were specifically designed to organize geographic space and

represent meaningful spatial partitions with a well-defined semantics. Consequently, they are frequently used to geo-reference geographic objects, data, and information.

In common-sense spatial reasoning, we may also resort to pSRTs for an intuitive construction of geographic models. In a large city, for instance, an intuitive way to geo-reference a geographic object is to describe its partonomic relation with respect to a unit of a pSRT (e.g., a tessellation of administrative subdivisions): We would say, for example, that the *"Alexanderplatz"* (a popular square in the heart of Berlin) is located *"in Berlin-Mitte"* (a central district of Berlin). Because the names as well as the approximate outlines and locations of the districts of Berlin are well known within the city and beyond, it is easy for a person living in *Spandau* (a suburb of Berlin) to reason that a restaurant located at the *Alexanderplatz* is not really spatially relevant to have lunch because *Berlin-Mitte* and *Spandau* are rather far apart. To come to this conclusion there is no need to know *exactly* how far (in kilometers). It is good enough to know that several districts have to be traversed to get from *Spandau* to *Berlin-Mitte*.

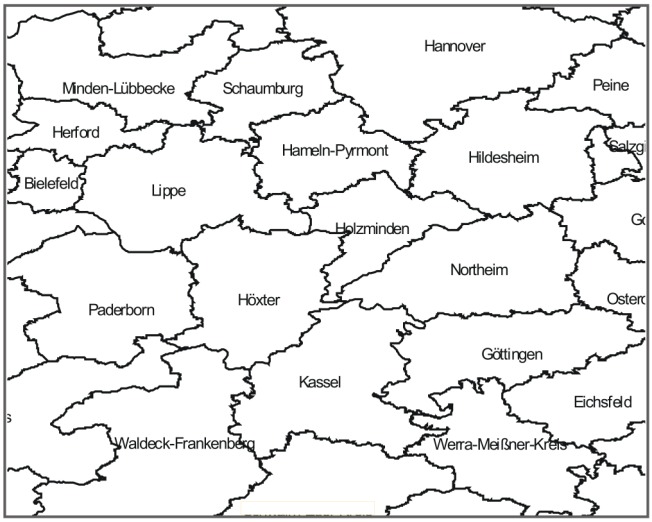

Figure 5.4: Example for a *"natural"* polygonal reference tessellation (*Landkreise* in Germany)

As we have seen in the example above, an important property of a suitable polygonal reference tessellation is that it actually represents a commonly accepted and standardized conceptualization of a geographic region. In fact, like the lists of standardized place names described in section 3.2, many polygonal tessellations (like tessellations of administrative subdivisions) encode *official standards*. Accordingly, they are created, organized and managed by government agencies or other large organizations. And analogous to lists of standardized place names, we can assume that both the names and the outlines of the cells of a reference tessellation remain invariant for an extended period of time[7].

[7]Nevertheless, because political and economic processes are dynamic, abrupt changes involving the re-organization of spatial partitions do occur. The unexpected re-unification of

In addition to administrative subdivisions, a number of other polygonal tessellations like postal code zones, telephone prefix areas, census districts etc. qualify as polygonal standard reference tessellations as well. However, the size of the user community that is familiar with a tessellation of census districts may be much smaller than in the case of administrative subdivisions. Empirical research would be needed to make more detailed assumptions about how large the user community of a given polygonal tessellation on a local, regional, national and international level actually is. In the context of this work, we assume that while administrative subdivisions are the best choice for a discrete spatial frame of reference, any of the natural tessellations mentioned above is better suited for this purpose than most artificially generated reference grids.

Examples for Polygonal Standard Reference Tessellations

The number of polygonal tessellations that qualify as Standard Reference Tessellations is relatively small. Some of the most frequently used pSRTs include tessellations of administrative subdivision and postal code zones:

Administrative Subdivisions: Tessellations of administrative subdivisions partition geographic space into a set of well-defined and typically well-known (both by name and approximate spatial extent) spatial units. In fact, the identifiers of administrative subdivisions are usually standardized official place names (see section 3.1). The polygonal footprints and boundaries of these place names are well-defined, and representations of the data are available in various digital formats. Among all natural polygonal tessellations, tessellations of administrative subdivisions are probably the ones that are best suited to serve as polygonal standard reference tessellations. Consequently, they are widely used in many digital and non-digital applications as an intuitive reference system to annotate data and information with spatial metadata, and to specify spatial queries.

Tessellations of administrative subdivisions typically have a distinct and well-defined hierarchical partonomic structure. This hierarchy reflects the organizational structure of the city, state, or country covered by the tessellation. Depending on the underlying organizational principles, the depth and complexity of the hierarchy may vary considerably.

An example for a tessellation of administrative subdivisions that cover an extended region of geographic space are *SABE*, the *Seamless Administrative Boundaries of Europe* [EuroGeographics, 2003]. *SABE* is the result of an effort to integrate the different administrative hierarchies of the individual EU member-states and many European states that do not belong to the European Union. The result is a continuous tessellation of administrative subdivision that covers 32 European countries and is available at two resolutions, at a 1: 100 000 scale (30 m resolution) and at a 1:1 000 000 scale (200 m resolution).

The data are derived from source data provided by the national mapping agencies of the individual states. These data are compiled and distributed

Germany, for example, caused considerable changes in the administrative subdivision of the country. The length of the period of stability of a specific reference tessellation may therefore vary, and re-adjusting the respective spatial models may be necessary from time to time.

by EuroGeographics [EuroGeographics, 2004], the association of European national mapping agencies. The term *"seamless"* indicates that *SABE* constitutes a continuous tessellation of Europe, i.e., that there are no gaps or overlaps between the data sets initially derived from different sources (Figure 5.5).

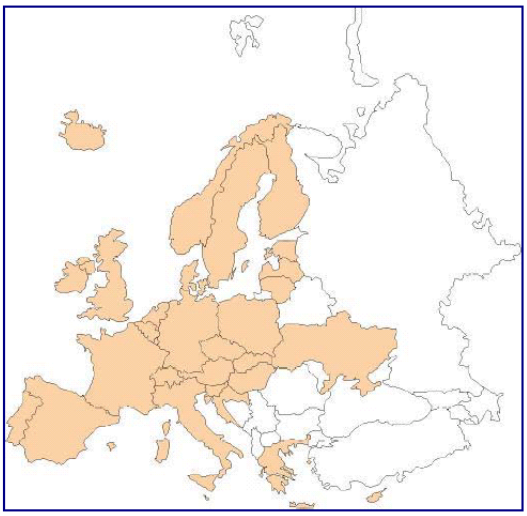

Figure 5.5: Coverage of the SABE dataset

At least for the current 15 EU member states, the hierarchical structure of the *SABE* data is based on the the classification of the European administrative units given by *NUTS*, the *Nomencleature des Unités Territoriales Statistiques* [European-Commission, 2003]. This European standard provides an integrated and uniform breakdown of territorial units for the production of regional statistics for the European Union.

The catalog defines a hierarchical classification of administrative subdivisions up to 5 *NUTS* region-levels. The naming scheme of *NUTS* follows the terminology used by the participating countries (e.g., *Länder* and *Kreise* in Germany, *régions* and *départements* in France, *Comunidades autonomas* and *provincias* in Spain, *regioni* and *provincie* in Italy, etc.). However, due to the still very heterogeneous administrative structures in the different EU member states, not all five *NUTS* levels are available for all countries.

Postal Code Zones: Postal code zones were originally developed in support of the logistics of postal delivery. Together with additional information (i.e., street name and number, flat number etc.) they provide a unique address for mail delivery. Today, an increasing number of (online) information services use postal codes to spatially organize data and information. Users of such services are able to search information (e.g., about the availability of hotels etc.) based on the specification of a postal code region. Often this search exploits the hierarchical organization of postal code zones, i.e. users may specify only the first couple of digits of a code to address a complete postal code region.

Because of the organizational structures of the postal services in different countries vary considerable, the systematics of postal codes are heterogeneous as well. In Germany, a general system of postal codes exists only since 1941. Until 1993, a 4-digit postal code number based on a hierarchy of distribution zones, distribution regions, and distribution districts was in use. German reunification made changes inevitable, and since 1993 a 5-digit postal code number is used.

In Sweden, a system very similar to the German postal code numbers has been in use since 1968. In this system, postal codes of five digits are used. Larger communities can be represented by multiple postal codes. In the Netherlands, a system of four numbers and two letters is used since 1978. The numbers signify communities and districts, the letters streets and city-blocks. Until 1972, postal code numbers in France corresponded to the two-digit numbers that are used to specify the French *departements* on car license plates. After 1972, three more digits were added to allow for a more flexible coding. In addition, the so-called CEDEX (*Courrier d'Enterprise a Distribution Exceptionelle*) numbers were introduced to be used by large enterprises and government organizations.

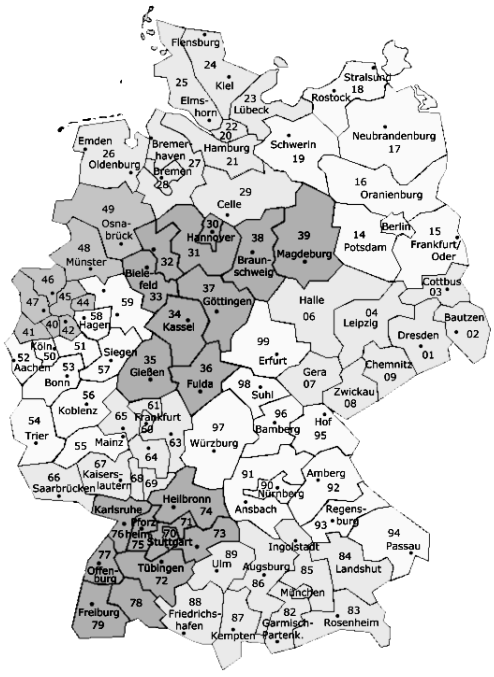

Figure 5.6: Postal code zones used by the *Deutsche Post AG*

For geo-referencing, postal codes are not as intuitive as administrative subdivisions because they provide only numbers and no names. Nevertheless, the organization scheme of most postal code systems reflects features in geographic space as well as administrative subdivisions to a certain extent. In the German system, for example, the first digit of the five digit code refers to 10 regions which, starting with 0 in the eastern states of *Sachsen*

and *Thüringen*, follow a counter-clockwise ascending order. Following this order, the first digit of given to the northern states of *Schleswig-Holstein*, *Hamburg*, and *Bremen* is a 2, while most parts of the southern states of *Baden-Württemberg* and *Bayern* are assigned the numbers 7 to 9. As a consequence of this ordering, most people do have at least some notion about the approximate location of a region within Germany if they are given the associated postal code.

The Partonomic Structure of a pSRT

The recursive decomposition of a polygonal standard reference tessellation yields a hierarchical structure that can be depicted in a *decomposition tree*. An important property of this decomposition hierarchy is that it is not arbitrary, but reflects the organizational semantics of the tessellation. In general, the levels of the hierarchy refer to one (or multiple) classification schemes that are used to organize types of geographic objects. Figure 5.7 shows the hierarchical structure of a tessellation of administrative subdivisions for the Federal Republic of Germany. Following the national levels of administration (*USE*), this hierarchy has 6 levels. With respect to the European *NUTS* classification scheme (see below), the tessellation can be decomposed into a tree with five levels [8] : While the object on the top level (*NUTS 0*) is of type *"Bundesrepublik"*, we find objects of type *"Bundesland"* on *NUTS 1*, objects of type *"Regierungsbezirk"* on *NUTS 2*, objects of types *"Landkreis"* and *"Stadtkreis"* on *NUTS 3*, and objects of types *"Gemeinde"*, *"Stadt"*, *"verbandsangehörige Gemeinde"*, *"verbandsangehörige Stadt"*, *"gemeindefreies Gebiet"*, and *"Wasserfläche"* on *NUTS 5*.

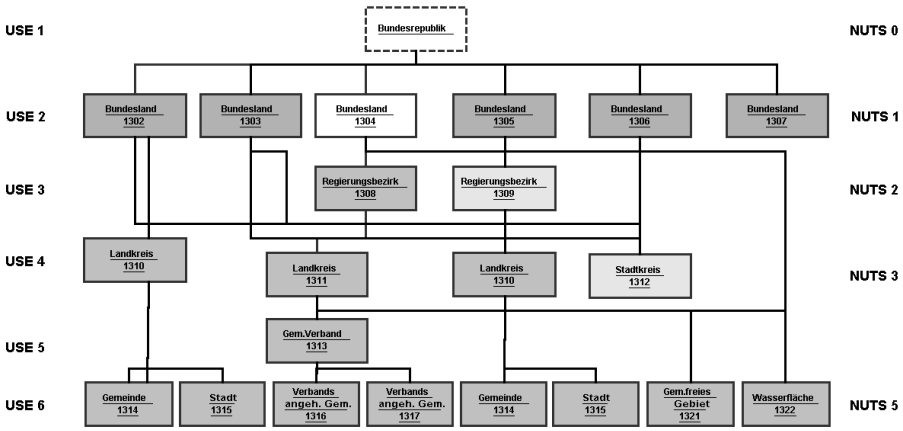

Figure 5.7: The hierarchy of *NUTS* and *USE* regions for Germany

Each level in the decomposition hierarchy represents a distinct level of granularity, with the resolution of the spatial discretization increasing from the top downward. Polygonal standard reference tessellations can therefore be used as *multi-resolution spatial reference models*. We will see in sections 7.1.1 and 7.1.2

[8] *NUTS 4* is missing because as a European classification scheme, *NUTS* tries to integrate the classification schemes of all European countries. For Germany, there does not exist a type that corresponds to *NUTS 4*.

how this property, together with the well-defined semantics of the decomposition tree, can be used to evaluate the spatial relevance of approximated regions.

Quantitative Representation of a pSRT

To be used in computerized information systems, the polygonal standard reference tessellations described above have to be available digitally. Some of them, like the *SABE* data, can be obtained easily but do require a substantial financial investment. A the full coverage of *SABE* data is currently priced at several thousand EURO. Other digital reference tessellations are not available through regular distribution channels. To obtain a comprehensive map of the German postal code zones, for example, proved to be extremely difficult as the *Deutsche Post AG*(the German postal service) does not sell digital version of the map to the public[9].

In addition to availability and cost, the use of vector-based digital representations of pSRT does often result in a number of technical problems, especially in cases where multiple pSRT have to be integrated. Among these problems are incompatibilities of data formats, coordinate systems, and geographic projections, as well as differences in the quality of the data. These problems are typical for polygonal vector data in general, but in the case of a vector-based digital pSRT, they can severely limit its usefulness as an interoperable frame of reference for spatial indices.

In particular issues associated to differences in the quality of the data may be problematic. While in theory it is straightforward to compute topological and other spatial relations between the polygons of a tessellation, bad data quality can result in serious errors. For example, to determine the neighborhood of a polygon P we have to select all polygons that share a common vertex with P (or a common boundary segment in case of a full neighborhood). In a data set where polygons are represented as individual and unconnected spatial objects (e.g., the popular ESRI *Shape* format), digitalization errors can cause topological inconsistencies that make this simple operation highly error-prone. Among the most frequent digitization errors are:

- Two polygons that appear to be neighbors but do not share a common vertex (Figure 5.8a).
- Vertices that represent the same point but do not have exactly the same coordinates (Figure 5.8b).
- *"Sliver"* polygons, i.e. adjacent polygons that slightly overlap in some areas (Figure 5.8c).
- Edges of adjacent polygons that run in parallel.

Standard GIS tools provide solutions to these problems, like the application of error-tolerances to recognize two vertices with almost the same coordinates as identical, or the creation of new vertices in the case of two polygons that are supposed to touch but do not share a common vertex. However, all these technical solutions can reduce, but not completely eliminate errors in the data sets.

[9] A digital map of German postal codes is included in the sample data delivered with ESRI's ArcGIS 8.X software, which is priced at several thousand EUROs.

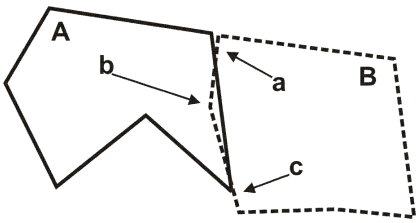

Figure 5.8: Common digitization errors: a.)Vertices do not match, b.)lines overlap (sliver), c.)vertex in polygon A is not matched by vertex in polygon B

In general, the complexity and large data volumes typical for digital vector representations may cause high computational costs. For example, the polygonal tessellation of the administrative units contained in the *SABE* data set covering the area of Germany has a size of more than 11 MB (in ESRI shape format) and comprises 14753 polygons. Theoretically, to find the neighbors of one administrative unit in this data set, all other 14752 polygons have to be accessed and evaluated. Modern GIS and spatial databases use elaborate spatial indices and optimized algorithms to reduce the computational costs for such operations as much as possible. However, run-time performance of spatial queries is still an issue in the GIS community.

In addition, to avoid redundancies, data sets like the *SABE* data do contain only polygons at the lowest administrative level covered. In the case of *SABE* this is *NUTS 5*, i.e. the polygons in the data set represent administrative units on a community level. Polygons of administrative units at higher levels of the *NUTS* hierarchy are not included. They have to be created through aggregation of the *NUTS 5* polygons. This aggregation is time-consuming and elaborate process that requires appropriate GIS tools and expertise.

To be able to use the resulting polygonal coverages in a GIS, the different levels of the hierarchy have to be organized in layers. Each layer L represents a specific level of the tessellation's decomposition tree, and each polygon P on L represents the union of a set of polygons $\{P_n, ..., P_m\}$ on layer $L+1$. However, in the GIS, both P and $\{P_n, ..., P_m\}$ are represented as different and unrelated spatial objects (i.e. polygons). Apart from the problem of data redundancies and overly complex spatial queries[10], problems related to data-quality as described above may cause the union of $\{P_n, ..., P_m\}$ to not exactly match P. The result are erroneous answers to spatial queries. A query supposed to select all polygons on layer $L + 1$ that are part of polygon P on layer L, for example, may yield polygons that slightly overlap with P, but do not belong to the decomposition of P.

In summary we can say that the representation of polygonal standard reference tessellations as digital vector data has a number of shortcomings that limit the useability of such data as the basis for interoperable spatial reference models and spatial relevance reasoning. These shortcomings are closely related to the intrinsic properties of exact, coordinate-based vector data. Although solutions to these problems can be found with the help of state-of-the-art GIS

[10]In traditional, layer-based GIS, partonomic relations between polygons can only be computed for two active layers at a time. Modern geodatabases offer an integrated access to multiple layers, but computational costs grow with the number of layers and polygons

technology, they may require considerable investments in high quality data, powerful hardware, as well as high-end software. An indication that practical constraints often prevent the solution of these problems in real-world applications is the observation that, in spite of the ongoing standardization process initiated by the OpenGIS Consortium and other organizations, true interoperability of digital geospatial data (e.g., vector-based pSRT) has not been achieved so far [GINIE, 2003].

On the other hand, for the type of spatial relevance reasoning we are concerned with in this work, the vector-based digital representation of a pSRT holds a lot of redundant information: To determine the neighborhood distance between two polygons P_1 and P_2, or to find the set $\{P_1, ..., P_n\}$ that constitutes the decomposition of P_3 we need not know the exact outline of P_1,P_2,P_3 and all the other polygons in the tessellation. A representation that is optimized for simple and effective spatial reasoning algorithms has to formalize the topologic and partonomic relations between the polygons explicitly, but can get rid of the bulk of superfluous data, like the exact coordinates of polygon vertices.

In the following section, we will present an approach to build optimized and light-weight representations using graph-based abstractions of polygonal tessellations. The main building-blocks for such models are *connection graphs* to encode the topological, and *decomposition trees* to encode the partonomic relations between the units of a polygonal tessellation.

5.2.3 Graph-Based Qualitative Spatial Reference Models

A Brief Introduction to Graphs

Graphs have long been used for the representation and study of topological relations. Leonard Euler's *Königsberg Bridge Problem* (Figure 5.9) dates back to 1736 and is probably the most famous example for qualitative spatial reasoning with the help of graphs. There is a wealth of literature that discusses graphs and graph theory in computer science (e.g.,[Dijkstra, 1959, Harary, 1971, van Leeuwen, 1990, Turau, 1996]) and applied to GIS (e.g., [Maguire et al., 1991, Laurini and Thompson, 1992, Worboys, 1995, Molenaar, 1998]). In this work we will very briefly recapitulate the terminology and basic ideas of graph theory, following the terminology proposed in [Harary, 1971]. This discussion will be focused and those concepts and algorithms that are relevant for our approach to graph based spatial models and spatial relevance reasoning.

A *graph* G is composed of a finite nonempty set $N = N(G)$ of n *nodes* (often also referred to as *vertices* or *points*) together with a prescribed set E of m unordered pairs of distinct nodes of N (Figure 5.10). Each pair $x = \{u, v\}$ of nodes in E is called a *line*, or *edge*, of G. The edge x is said to join the nodes u and v. Given an edge $x = \{u, v\}$ we say that u and v are *adjacent*, and that node u and edge x are *incident*.

In a *directed graph*, the pairs of u and v in E are ordered. The elements in E are referred to as *directed lines*, or *arcs*. In a *multigraph*, two nodes can be joined by more than one edge, called *multiple edges* or *multiple lines*. A graph is called *labelled* if its nodes are given distinct names, such as $v_1, v_2, ..., v_n$.

A *walk* of a graph G is defined as the sequence of nodes and edges that have to be traversed to move from one point to another. A walk from v_0 to v_n

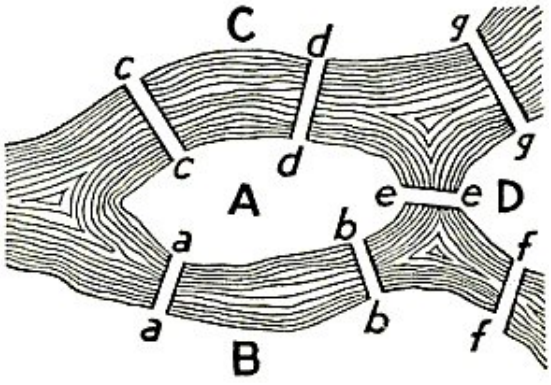

Figure 5.9: The Koenigsberg Bridge Problem (after [Kraitchik, 1942])

thus results in the sequence $v_0, x_1, v_1, ..., v_{n-1}, x_n, v_n$, or $v_0 v_1 ... v_n$ if the explicit representation of the respective edges is omitted. If all nodes and the respective edges in a walk are distinct, it is called a *path*. The length of a walk (and thus of a path) is equal to the number of edges in it, i.e., the length of a path $v_0 ... v_n$ is n. The shortest path between two nodes is the path where n is minimal. A graph is called *connected* if a path can be defined for every pair of nodes.

The length of the shortest path between two nodes u and v in a graph G is called the *distance d(u,v)*. In a connected graph, distance is a *metric*, i.e. for all nodes u, v, and w we can say that:

1. $d(u, v) \geq 0$, with $d(u, v) = 0$ iff $u = v$,
2. $d(u, v) = d(v, u)$, and
3. $d(u, v) + d(v, w) \geq d(u, v)$.

For a connected graph G, the diameter $d(G)$ of G is defined as the length of any longest distance $d(v, u)$ between two nodes u and v in G.

A *tree* is a special case of an *acyclic* and *connected* graph. Most authors define a tree as being *directed* as well, with a *root* node at the top and *leaf* nodes at the bottom. The diameter of a tree T is often called the *depth* $D(T)$.

The *embedding* of a graph in the Euclidean plane encodes additional spatial information. A *planar graph* is a graph that is embedded in the plane in a way that the edges do only intersect at nodes. Usually, there are a number of possible planar embeddings for a given planar graph. These embeddings are equivalent with respect to the connection relations they encode. Topologically and metrically, however, they are not equivalent. Analogous to the polygons discussed above, the planar embedding of a planar graph subdivides the plane into *faces* (also called *regions*). The *Euler formula* defines the relation of the number of faces f, the number of arcs or edges e, and the number of nodes n that a consistent planar graph has to meet: $f - e + n = 1$.

If we represent each face in a planar embedded graph G by a node and draw an edge between all nodes that are adjacent, we obtain the *dual* $G*$ of that graph (Figure 5.11).

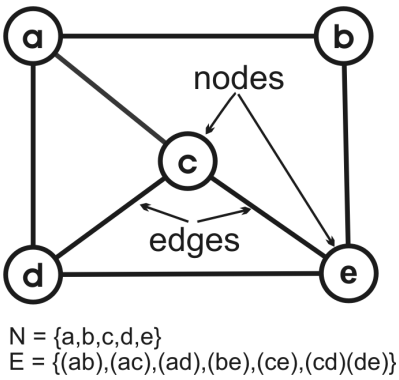

Figure 5.10: Features of a graph

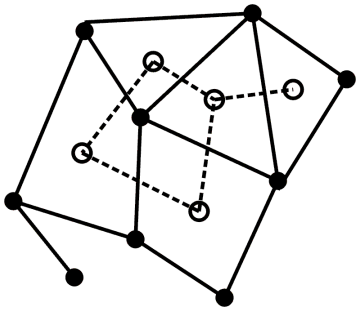

Figure 5.11: Planar graph and its dual

We will see below that this is an important concept in the context of adjacency graphs in polygonal tessellations. A well-known pair of graph and dual is the *Voronoi diagram* and the *Delaunay triangulation* (e.g., [Dirichlet, 1850], [Voronoi, 1908], [Okabe et al., 1992]): Given a set of arbitrarily distributed points in the plane, an area of closest proximity can be defined for each point (i.e. all locations in this area are closest to the point in question). These areas are polygons (also called *Thiessen polygons*) and constitute a tessellation, the *Voronoi diagram*. The dual of a Voronoi diagram is called a *Delaunay triangulation* (Figure 5.12).

Graph representations support a number of search algorithms that can be applied to find the shortest path between two node, or to compute the n^{th} order neighborhood of a node. For undirected graphs, one particularly useful method is the shortest path algorithm developed by Dijkstra [Dijkstra, 1959]. The Dijkstra- and other graph algorithms are described in more detail for example in [Turau, 1996] and [Harary, 1971].

Connection Graphs

Polygonal tessellations like the polygonal standard reference tessellations described in section 5.2.2 can easily be abstracted as graph-based representations. In fact, the dual of a polygonal tessellation yields a graph that encodes the adjacency relations between the individual polygons of a polygonal standard

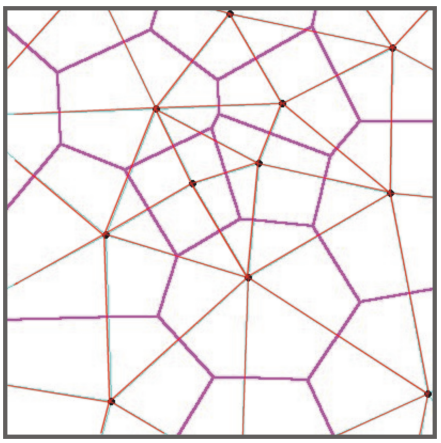

Figure 5.12: Voronoi diagram and Delaunay triangulation

reference tessellation. The result is an adjacency or *neighborhood graph* as described in section 4.2.4.

It is obvious that any adequate qualitative abstraction from the spatial information contained in a polygonal reference tessellation must preserve at least the information encoded in the neighborhood graph. However, to support reasoning about spatial relevance, the qualitative representation of the decomposition must preserve even more information, namely some type of ordinal information. If only the neighborhood graph is used, then the two arrangements of polygons shown below cannot be distinguished (Figure 5.13). Both have the same graph, but they differ fundamentally with respect to neighborhood: Neighbors of P_1 and P_3 can never be neighbors of P_2 if the polygons are arranged as in Figure 5.13-a, while they can in the arrangement depicted in Figure 5.13-b. The problem is caused by the existence of a dual neighborhood between P_1 and P_3 in Figure 5.13-a., which have two (disconnected) edges in common. An adequate graph-based representation of a polygonal tessellation should be able to encode multiple neighborhood relations between polygons.

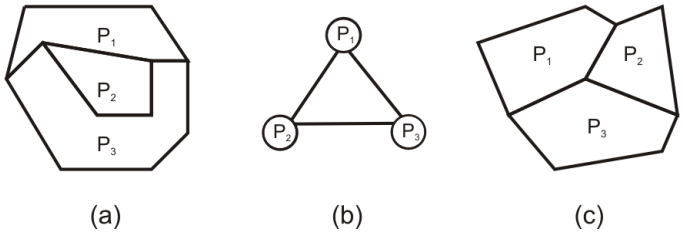

Figure 5.13: Multiple neighborhood relations

In [Schlieder et al., 2001] we proposed the *connection graph* as an abstraction of a polygonal tessellation that supports the specification of multiple neighborhood relations between polygons.

DEFINITION 5.1
The connection graph of a homogeneous decomposition by tessellation with

neighborhood graph $\mathcal{N} = (V_\mathcal{N}, E_\mathcal{N})$ *is a graph* $\mathcal{C} = (V_\mathcal{C}, E_\mathcal{C})$ *together with the combinatorial embedding of* $\mathcal{C}$ *in the plane.* $V_\mathcal{C} = V_\mathcal{N} \cup \{E\}$ *where* E *is the exterior, unbounded polygonal region.* $E_\mathcal{C}$ *contains an edge* (P_i, P_j) *for each connected sequence of polygon edges that* P_i *and* P_j *share. The combinatorial embedding of* $\mathcal{C}$ *consists in the circular ordering of the edges from* $E_\mathcal{C}$ *at each vertex from* $V_\mathcal{C}$*.*

Figure 5.14 shows the connection graph $\mathcal{C}$ of a homogeneous decomposition by tessellation D. Each polygon of D is represented by a vertex from $\mathcal{C}$. In addition there is the node 1 representing the external polygonal region. The edges from $\mathcal{C}$ which are incident with a vertex are easily obtained together with their circular ordering by scanning the contour of the corresponding polygon. For polygon 10 the following circular sequence of neighbors is obtained: $1, 2, 3, 4, 6, 8, 9, 8$. Note that polygon 8 appears twice in this list because it shares with 10 two disconnected polygon edges. On the other hand polygon 9, which shares three edges with 10, appears only once because the three edges are connected. As the example shows, the connection graph is a multi-graph in which several edges can join the same pair of vertices. Technically speaking, the connection graph consists of the dual of the tessellation together with the combinatorial embedding of the dual.

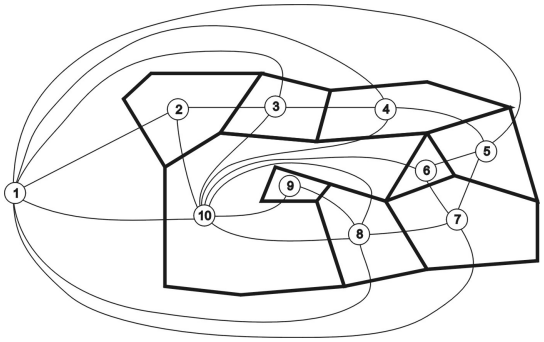

Figure 5.14: Connection graph representation of a decomposition by tessellation

Decomposition Trees

In section 5.2.2 we said that *"natural"* polygonal tessellations are typically organized along the lines of hierarchical partonomies which reflect the intrinsic semantics of the tessellation. The partonomic structure of a polygonal tessellation can be accessed through its recursive decomposition. In the general case, the decomposition hierarchy can be represented as a *directed acyclic graph (DAG).*

For the special case of a polygonal standard reference tessellation, the decomposition hierarchy is a *decomposition tree* because each polygon at level $L+1$ belongs to only one polygon at level L. The total depth D of the decomposition tree is the distance from the deepest leaf to the root of the tree, i.e., the minimum number of nodes that have to be traversed vertically upward from a polygon at the highest decomposition level to the polygon that represents the undecomposed tessellation.

The set of nodes in the decomposition tree $\mathcal{D}$ of a reference tessellation pSRT represents the set of *all* polygons $P \in pSRT$. A reference unit r_2 at depth $d+1$ is said to be *spatially-part-of* a unit r_1 at depth d if the two nodes representing the units are connected in the tree, i.e. $r_2 \sqsubseteq r_1$ iff r_2 is a child of r_1.

Figure 5.15 shows the connection graphs $\mathcal{C}_\mathrm{i}$ and the decomposition tree for $\mathcal{D}$ a simplified polygonal reference tessellation.

Architecture of a Qualitative Spatial Reference Model

In this work, we use graph-based abstractions of polygonal standard reference tessellations (pSRTs) to create *qualitative spatial reference models*. A pSRT represents a discrete and hierarchical partitioning of a finite region of geographic space. The hierarchical structure of such a partitioning can be represented by a decomposition tree $\mathcal{D}$.

Because we use them to geo-reference place name regions, we refer to the individual polygons in a polygonal standard reference tessellation as *reference units* r. Reference units can be grouped into partonomic sets S, where $S = \{r_i, ..., r_n\}, r \in pSRT$. A partonomic set S of reference units is called *non-redundant* ($NR(S)$) if none of the reference units in S is *spatially part-of* another reference unit in S,

$$NR(S) \Longleftrightarrow \forall r_i \in S \forall r_j \in S : r_i \sqsubseteq r_j \longrightarrow i = j, \tag{5.2}$$

A partonomic set S of reference units can be called *normalized* ($NO(S)$) if all reference units $r \in S$ have the same graph-theoretical depth d with respect to the decomposition tree of pSRT:

$$NO(S) \Leftrightarrow \forall r_i \in S, \forall r_j \in S | d(r_i) = d(r_j) \tag{5.3}$$

An important property of a valid polygonal reference tessellation is that it can be decomposed into normalized partonomic sets of reference units, with each set representing a specific level of the partonomic hierarchy of the tessellation. Because each level of the hierarchy represents a specific spatial resolution of the tessellation, we speak of a *level-of-detail* (L) of the tessellation which is incident to a depth d of the respective decomposition tree. The decomposition tree in Figure 5.15, for example, has four levels-of-detail, with L_0 being the least, and L_3 being most detailed representation.

In a valid pSRT, a normalized and non-redundant set S_L of polygons exists at every level L of $\mathcal{D}$. Each normalized and non-redundant set S_L of polygons $\{P_1, ..., P_i\}$ at level L can be represented as an individual connection graph $\mathcal{C}_\mathrm{L}$. Consequently, the graph-based abstraction $\mathcal{S}$ of a pSRT consists of $D+1$ layered connection graphs, where D is the total depth of the decomposition tree $\mathcal{D}$ of the pSRT.

Definition 5.2
A qualitative spatial reference model $\mathcal{S}$ is a graph-based representation of a polygonal standard reference tessellation (pSRT). Each resolution level L in pSRT is represented by a connection graph $\mathcal{C}_\mathrm{L}$ in $\mathcal{S}$, while the decomposition hierarchy of pSRT is represented by a decomposition tree $\mathcal{D}$.

A qualitative spatial reference model based on a polygonal standard reference tessellation is therefore built as a structure that combines the layered set of

connection graphs with the respective decomposition tree (Figure 5.15). The connection graphs represent the coverage of a finite region of geographic space by polygonal tessellations at different levels of granularity, while the decomposition tree represents the partonomic relationships of the individual units of these tessellations. The model is valid for a specific geographic region and up to a specific level of detail.

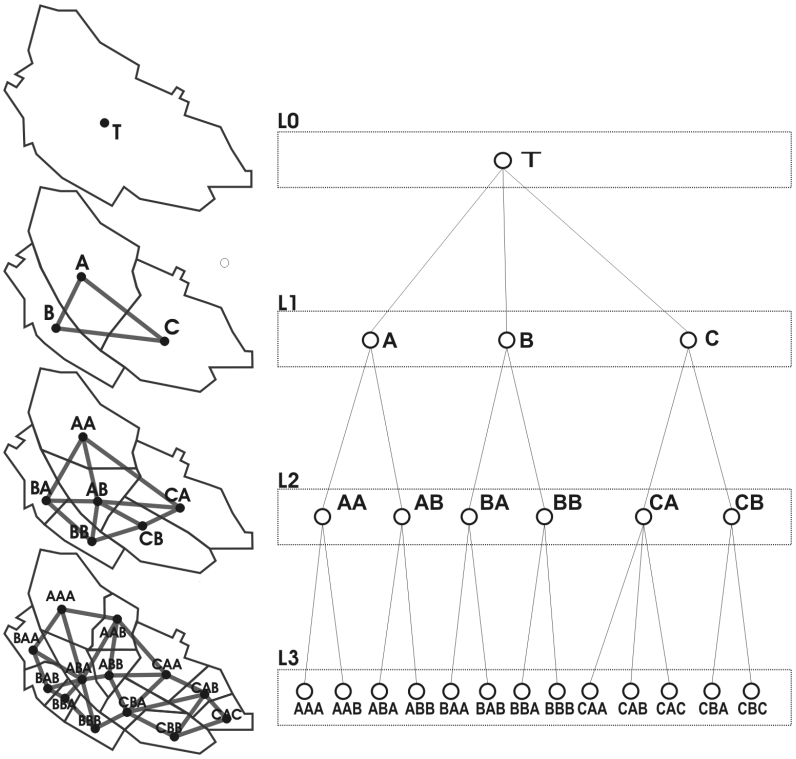

Figure 5.15: Architecture of a qualitative spatial reference model $\mathcal{S}$

Encoding of Qualitative Spatial Reference Models

The purpose of a qualitative spatial reference model is to provide a light-weight, portable, and interoperable spatial frame of reference. One appropriate encoding for such a model is to use a schema based on the *Extensible Markup Language (XML)* [W3C, 2000]. Figure A.1 shows an extract from a spatial model of administrative units of Germany (using the NUTS classification) that we encoded in an *XML* schema developed for the purpose. For each node of the reference model, the *XML* representation encodes a unique identifier, the name of the administrative unit (i.e. the place name), an ordered list of the object's neighbors (in terms of their unique identifiers), and the identifier of the immediate parent of a node.

Although the underlying *XML* schema has not been optimized yet, we already achieve a considerable reduction of data volumes compared to the respective digital vector data (see section 7.1.4). By choosing an open and non-

proprietory data format we want to further improve the interoperability of qualitative spatial reference models.

Qualitative spatial reference models encode both topological, ordinal, and partonomic relations between polygonal regions. In the following section we will discuss how this information can be used for spatial relevance reasoning.

Chapter 6

Approximated Regions

6.1 Discrete Representation of Place Name Regions

6.1.1 Discrete Approximations of Regions

In the previous section we outlined the architecture for qualitative spatial reference models based on discretizations of geographic space provided by polygonal standard reference tessellations. We showed that through a graph-based representation of such tessellations, we can considerably reduce the associated complexity and data volumes. In comparison to quantitative coordinate-based vector data (which represent a standard data format in GIS), qualitative models are light-weight and interoperable.

The purpose of qualitative spatial reference models is to provide a framework for the representation of regional geographic objects (i.e., place name regions), and to support algorithms for the evaluation of the relative spatial relevance of these geographic objects. In this chapter we will discuss how both crisp and indeterminate regions can be approximated on the basis of qualitative spatial reference models. In chapter 7 we will show how these models can be used to reason about the spatial relevance of approximated regions.

Discrete approximations of regions in general, and of geographic regions in particular have been discussed by a number of authors (e.g., [Worboys, 1998], [Bittner and Stell, 1998], [Galton, 1999], see also section 4.4). As indicated in section 5.2, our intention is to use discrete approximations of regions based on qualitative spatial reference models for spatial relevance reasoning. In this work we focus on geographic space and geographic objects. Therefore we use the term *"region"* to denote a geographic region, i.e., the regional extent of a geographic object. And because we only look at *named* geographic objects, or place names, the regional extent of a geographic object is also referred to as a *place name region*.

Analogous to the spatial footprint used in gazetteers and other place name lists (section 3.2), we call the discrete approximation of a place name region the *qualitative spatial footprint* of the place name. In the simplest case, a qualitative spatial footprint is created through a binary *"thematic"* mapping between the place name region and the respective units in a spatial reference model

[Schlieder and Vögele, 2002] [Vögele and Schlieder, 2003]. We call the resulting discrete approximation, where place name region is represented by the union of all intersected reference units, a *simple* qualitative spatial footprint.

Whether a simple qualitative spatial footprint is good enough to represent a given place name region depends very much on the *resolution* of the qualitative spatial reference model and the degree of *vagueness* of the place name region in question. In the (frequent) case where a place name region corresponds directly to the units of a reference tessellation (e.g., if the outline of a place name region was defined on basis of administrative units), a simple qualitative spatial footprint may yield a sufficiently accurate approximation.

However, as we have seen in section 3.1.3, many place name regions are intrinsically vague. For the representation of vague regions, and in the case where there is a large discrepancy between the (maximum) resolution of the spatial reference model and the (average) dimensions of the place name regions to be represented, simple binary mappings do not suffice. To improve the quality of the spatial footprint representations we have to develop more elaborate approximations, which we call *complex* qualitative spatial footprint [Vögele et al., 2003b].

In the following, we first introduce simple qualitative spatial footprints. We discuss the general impact of the resolution of the underlying reference tessellation on the quality of a spatial footprint. We then outline an approach for the representation of complex spatial footprints in which we extend simple spatial footprints through the specification of an upper and a lower approximation.

6.1.2 Simple Qualitative Spatial Footprints

Discrete partitions of geographic space as provided by a qualitative spatial reference model can be used to define place name regions in terms of *spatial indices*. A place name region p is the 2-dimensional extent of a geographic object denoted by a place name in 2-dimensional Euclidean space $\mathbb{R}^2$. Following an idea developed by Schlieder and presented in [Schlieder et al., 2001, Vögele and Schlieder, 2003], we use a projection of p to a discrete representation of $\mathbb{R}^2$ (i.e., a qualitative spatial reference model $\mathcal{S}$) to obtain a set of partitions of $\mathbb{R}^2$ that are called the *qualitative spatial footprint (S_p)* of p.

If only one set of spatial indices is used to extensionally-define a place name region p, we call this set the *simple* qualitative spatial footprint of p[1]. The set of spatial indices that is used as an adequate approximation of p in the discrete space of the $\mathcal{S}$ is derived through a binary mapping of p onto $\mathcal{S}$. The resulting simple qualitative spatial footprint S_p includes all reference units $\{r_1, ..., r_n\} \in \mathcal{S}$ that intersect with the 2-dimensional projection of p onto $\mathcal{S}$.

Formally, to express the regional extent of a place name $p \in P$ (where P is the set of all place names) in terms of a set of reference units $S_p = \{r_i, ..., r_n\}$, $r_i \in R$, where R is the set of all reference units in the spatial reference model, we use a function $L : p \rightarrow 2^R$, where 2^R is the power set of all reference units.

To define which reference units of a spatial reference model belong to the approximation of p, the function L evaluates the topological relations between p and all $r_i \in R$. Because both p and r_i represent regions in $\mathbb{R}^2$, we can use the

[1] Note that the spatial indices in a spatial footprint do not have to form a set of (spatially) connected cells of a discrete partitioning. It is possible to define the spatial footprint of a region in terms of multiple disjunct clusters reference units.

region connection calculus *RCC-5* [Bennett, 1994] to evaluate the topological relations between them. Function L assigns every reference unit $r_i \in R$ to the simple qualitative spatial footprint of p for which one of the following *RCC-5* relations, or the respective inverse relation, is true: $EQ(r_i, p)$,$PP(r_i, p)$, or $OL(r_i, p)$.

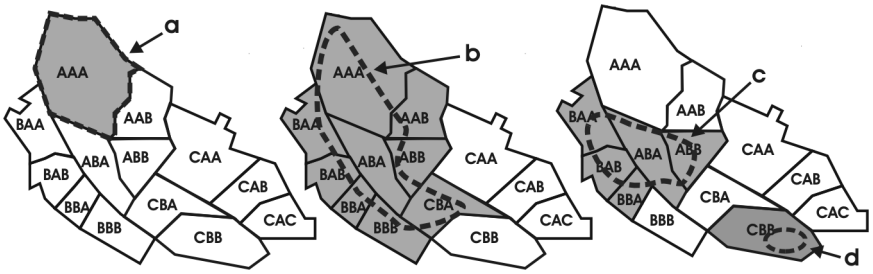

Figure 6.1: Examples for simple qualitative spatial footprints

For the case where p is equal to, or a proper part of, a single reference unit r_i, S_p contains only one element, namely r_i. If p overlaps with or contains multiple reference units, S_p consists of a set $\{r_i, ..., r_n\}$ of reference units. On the other hand, an empty set is assigned to S_p for the case where p and r_i are either disconnected, or only externally connected[2].

$$L(p) = S_p = \begin{cases} \{r_i\} & : \quad EQ(p, r_i) \vee PP(p, r_i) \\ \{..., r_i, ...\} & : \quad PO(p, r_i) \vee PP^{-1}(p, r_i) \\ \{\} & : \quad DR(p, r_i) \end{cases} \tag{6.1}$$

Applied to the spatial reference model depicted in Figure 6.1, we can define the following simple qualitative spatial footprints for the four place name regions *a*,*b*, *c*, and *d*: $S_a = \{AAA\}$, $S_b = \{AAA, AAB, ABA, ABB, BAA, BAB, BBA, BBB, CBA\}$, $S_c = \{ABA, ABB, BAA, BAB, BBA\}$, and $S_d = \{CBB\}$.

To define the qualitative spatial footprint of a place name region, function L can be applied either automatically or manually. An automatic mapping of a place name region onto a spatial reference tessellation can be performed if both the region and the tessellation are available in digital format. Provided the availability of suitable digital data, using an automatic mapping procedure allows for the definition of a large number of spatial footprints with relatively little effort.

On the other hand, the spatial footprint of a place name region may as well be defined manually through a visual or cognitive matching of the region with the reference units. Because the user typically knows the names and approximate outlines of the units of a polygonal standard reference tessellation (see section 6.2.3), defining a spatial footprint manually is straightforward. However, unlike in the case of an automatic mapping, it may be more intuitive and greatly simplify the process to use reference units from different levels of the spatial

[2]In practical terms this means that if a place name region is located outside the geographic region covered by the spatial reference model, the respective qualitative spatial footprint is an empty set.

reference model. In the example given above, place name b is defined by a set of 9 reference units.

In a polygonal standard reference tessellation, each polygon at a lower level of resolution can be decomposed into a set of tessellating polygons at a higher level of resolution. Likewise, polygons may be aggregated to form their "parent" polygon at the next lower resolution level. This makes it possible to simplify the representation of S_p if one or multiple subsets of S_p at L can be aggregated to polygons at $L-1$. The representation of S_b, for example, can be compacted from $S_b = \{AAA, AAB, ABA, ABB, BAA, BAB, BBA, BBB, CBA\}$ to $S_b = \{A, B, CBA\}$. It is obvious that the compacted version of S_b requires less storage space and is easier to build manually.

The representation of a qualitative spatial footprint may be compacted as long as the resulting set of reference units remains *non-redundant* (section 5.2.3). As indicated by the term *non-redundant*, none of the elements in this set is allowed to be *spatially-part-of* another element of the spatial footprint. The fact that a place name region can be approximated using reference units from different levels of a multi-resolution reference model has significant practical implications. It simplifies the specification and representation of spatial footprints, especially in cases where large and small place name regions are to be evaluated together.

6.1.3 Resolution of the Reference Model

A qualitative spatial footprint approximates the regional extent of a geographic object in a discrete geographic space. The quality of this approximation, i.e. the error introduced through the mapping function L, depends very much on the spatial resolution of the underlying reference model. Given a multi-resolution spatial reference model, the selection of the appropriate level of resolution is important. To make a judgement about which level of resolution should be used to approximate a place name region p, the relation between the size (in terms of its spatial extension) of p and the average size of the units of the reference model in question, r_ϕ has to be considered. In general, we can distinguish three distinct cases with respect to this relation:

Case 1 - $r_\phi \ll p$: The place name region p is much larger than the average size of the reference units $r \in \mathcal{S}$, i.e. the reference model is fine-grained with respect to the regions that are to be approximated (Figure 6.2).

In this case, the error introduced by an approximation of p in terms of $\mathcal{S}$ is likely to be small. In the worst case, applying the binary mapping function L to p will result in an approximation S_p that slightly *overestimates* the actual size of p (Figure 6.2-Region A). This is due to the fact that L includes in S_p those reference units that only partially intersect p. In the best case, i.e. if the outline of the approximated region follows more or less the outlines of the reference units, the error can be neglected (Figure 6.2-Region B).

Case 2: $r_\phi \gg p$: The place name region p is much smaller than the average reference unit r. In this case, the approximation of p is equal to the respective reference unit, i.e. $S_p = \{r\}$. If r cannot be decomposed into smaller units, no further resolution is possible. For example, project areas a and

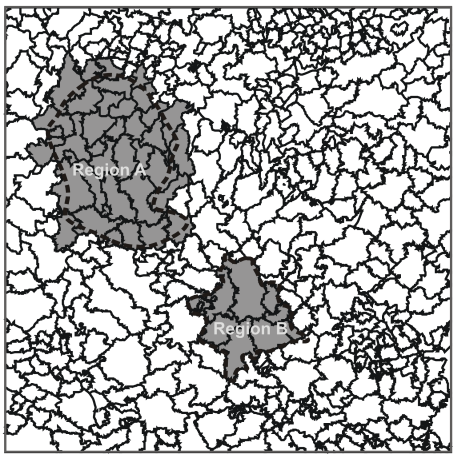

Figure 6.2: pSRT resolution Case 1: $r_\phi \ll p$

b depicted in Figure 6.3 have the same spatial footprint, namely reference unit r_1. In a Case 2 situation, the quality of a footprint approximation depends very much on the resolution of the reference tessellation. If the reference tessellation is coarse, a large locational and extensional error will be introduced.

Figure 6.3: pSRT resolution Case 2: $r_\phi \gg p$

It is therefore often not possible to create adequate footprint approximations in a Case 2 situation. Nevertheless, this situation is frequently encountered in practical applications: Many place names are associated to geographic objects that constitute *landmarks*. Examples are parks, squares, and buildings in a city, or small lakes and mountain tops in a natural area. Such landmarks do typically have a much smaller spatial extent than the available reference units. Often the only way to represent a landmark in a qualitative spatial reference model is to say that it is located somewhere within a specific reference unit.

Case 3 - $r_\phi \approx p$: Both the units of the reference tessellation and the place name region have approximately the same size. Here, the locational and extensional uncertainty of S_p may vary considerably, depending on specific properties of p. We distinguish two sub-cases:

Case 3a - The outline of p follows the boundaries of r: The error introduced by an approximation of p in $\mathcal{S}$ is small if there is an (more or less) exact overlap of p with a set of reference units $K = \{r_1, ..., r_n\}, r_i \in \mathcal{S}$. In terms of *RCC-5* we can say that the relation $EQ(p, K)$ holds, i.e. if the spatial extent of p is equal to the spatial extent of S. Here, the qualitative spatial footprint $S_p = K$ is an exact representation of the spatial extent of p (Figure 6.4).

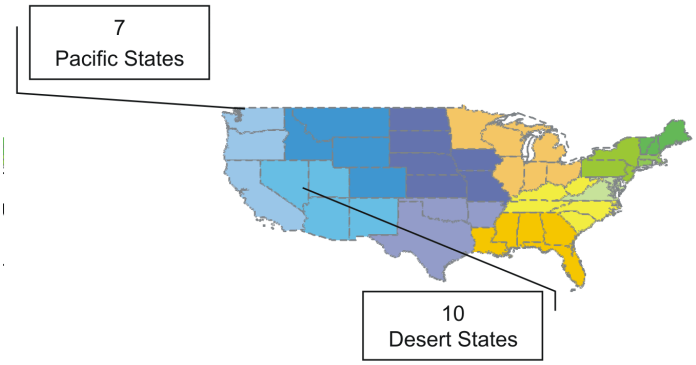

Figure 6.4: pSRT resolution Case 3a: $r_\phi \approx p$ and congruence of boundaries

This case occurs frequently in real-world applications because many regions with man-made boundaries are defined as aggregations of other regions of the same type. For instance, the region called *"Pacific States"* in Figure 6.4 is defined as an aggregation of three man-made subdivisions of geographic space, namely the states of *"Washington"*, *"Oregon"* and *"California"*.

Case 3b - The outline of p does not follow the boundaries of r: For the case where a place name region p partially overlaps multiple reference units, but not fully contains any of them, S_p is a very crude approximation of p that considerably overestimates the spatial extent of p (Figure 6.5).

This situation is often encountered when natural regions are approximated on basis of tessellations of man-made subdivisions (e.g., administrative units). Here, the boundary of the natural region may be found to cut through several reference units, sometimes without fully overlapping any of them. An example is shown in Figure 6.5, where the mountain region of the *Weserbergland* partially overlaps several reference units (*"Landkreise"* of the Federal states of *Northrhine-Westfalia* and *Lower Saxony*). The resulting (simple) qualitative spatial footprint for the *Weserbergland* region is a very coarse approximation that highly overestimates the extent of the mountain range.

Figure 6.5: pSRT resolution Case 3b: $r_\phi \approx p$ and divergence of boundaries

Obviously the quality of a discrete spatial footprint approximation depends on the level of resolution at which it is defined. In general we can say that the finer the resolution, the better the approximation. Best results are obtained if $p/r_\phi \gg 1$, i.e., if the ratio between the size of the place name region and the average size of the reference units is much larger than 1. In this case, we approximate a large place name region using a fine-grained reference tessellation. This could lead to the conclusion that the best strategy to achieve optimal footprint approximations is to always use the highest available resolution of a spatial reference model.

However, this approach has some serious shortcomings: For one, given a fine grained reference model and a large place name region, the set of reference units that constitutes the respective qualitative spatial footprint may become very large. Without the help of appropriate tools like graphical user interfaces or full-blown GIS applications, the manual definition of such a spatial footprint will be tedious, if not impossible. From a computational point of view, a large spatial footprint will increase the size of the representation and cause unnecessary computational overhead.

Another problem is implicit in the design of the qualitative spatial reference model itself: Because the reference model is built using a *"natural"* polygonal standard reference tessellation (e.g., administrative subdivisions), there is only a finite (and often rather small) set of reference layers at different, but fixed resolution levels available. In general the maximum resolution of the model is determined by the characteristics of the underlying reference tessellation and cannot be extended easily.

A spatial reference model based on a tessellation of administrative subdivisions following the *NUTS* classification for *Germany* (see section section 5.2.2), for example, consists of five granularity levels. In this model, the maximum resolution is set by the tessellation of reference units of type *"Gemeinde"*, which

refers to *NUTS 5*. Based on the maximum resolution provided by this model, two geographic objects located in the same *"Gemeinde"* cannot be spatially disambiguated. Consequently the model is not well-suited to approximate regions on a *"sub-Gemeinde"* level (e.g., a municipal park). At the same time, the *"Gemeinde"* level may be too detailed to approximate a larger region like a national park, whereas the next less-detailed level available (the *"Landkreis"* level) may be too coarse for the purpose.

6.1.4 Complex Spatial Footprint Approximations

To overcome the limitations of simple qualitative spatial footprints and to obtain better approximations for place name regions, we introduced the concept of *complex qualitative spatial footprints* [Vögele et al., 2003b]. This approach is built on work by Worboys [Worboys, 1998] who applied rough set theory to define the *upper* and the *lower* approximation of a discretized region (see section 4.4). We applied this idea to the concept of a qualitative spatial footprint S_p of a place name p and distinguish between an *upper approximation* $\bar{S}_p$ and a *lower approximation* $\underline{S}_p$ of p.

The upper approximation $\bar{S}_p$ defines the maximum extent of p in terms of the units of a qualitative spatial reference model $\mathcal{S}$. $\bar{S}_p$ is incident with the simple spatial footprint of p, i.e., the approximation of p obtained by applying function L to map p onto $\mathcal{S}$. It represents the largest possible footprint for p and consists of all reference units that *probably* belong to the "true" spatial approximation of p. Analogous to equation 6.1, we obtain $\bar{S}_p$ by applying a mapping function $\bar{L} : p \rightarrow 2^R$. Expressed in terms of *RCC-5*, $\bar{L}_p$ assigns all reference units $r \in \mathcal{S}$ to $\bar{S}_p$ that *partially overlap* p, are *proper part of* p, *contain* p, or are *equal to* p:

$$\bar{L}(p) = \bar{S}_p = \begin{cases} \{r_i\} & : \; EQ(p, r_i) \\ \{..., r_i, ...\} & : \; PP^{-1}(p, r_i) \vee PP(p, r_i) \vee PO(p, r_i) \\ \{\} & : \; DR(p, r_i) \end{cases} \tag{6.2}$$

Likewise, the *lower approximation* $\underline{S}_p$ of p is obtained through a mapping function $\underline{L} : p \rightarrow 2^R$. $\underline{S}_p$ denotes the minimum extent of the place name region in discrete reference space, i.e. all reference units that *definitely* belong to its spatial footprint. $\underline{S}_p$ only comprises those reference units r that are *equal to* or a *proper part of* p:

$$\underline{L}(p) = \underline{S}_p = \begin{cases} \{r_i\} & : \; EQ(p, r_i) \\ \{..., r_i, ...\} & : \; PP^{-1}(p, r_i) \\ \{\} & : \; PP(p, r_i) \vee PO(p, r_i) \vee DR(p, r_i) \end{cases} \tag{6.3}$$

The lower approximation of a region p is always a subset of its upper approximation. The complex spatial footprint S_p of p is therefore given by the tuple

$$S_p = (\underline{S}_p, \bar{S}_p), where \underline{S}_p \subseteq \bar{S}_p \tag{6.4}$$

In Figure 6.6, region a is approximated using a qualitative spatial reference model based on a tessellation of US administrative subdivisions (i.e., US counties). The figure depicts a situation that corresponds to the trivial case

described in *Case 1* in section 6.1.3 where the average size of reference units is significantly smaller than the region to be approximated. If we apply equations 6.1 and 6.3, we obtain two non-empty sets of reference units $\bar{S}_a$ and $\underline{S}_a$ that fulfill the condition $\underline{S}_a \subset \bar{S}_a$.

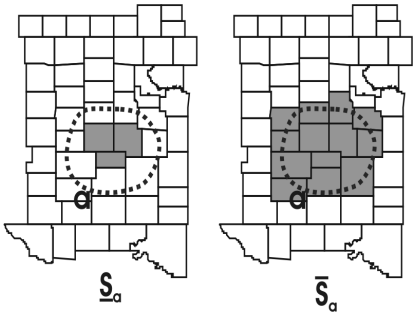

Figure 6.6: The upper and lower approximation of a region a

The trivial case applies also to region b shown in Figure 6.7. The complex spatial footprint of b is given by the tuple $S_b = (\bar{S}_b, \underline{S}_b)$, where $\bar{S}_b = \{AAA, AAB, ABA, ABB, BAA, BAB, BBA, BBB, CBA\}$, and $\underline{S}_a = \{ABA\}$.

Region c in Figure 6.7 is an example for the case where the average size of the reference units is similar to the size of the region (*Case 2* in section 6.1.3) and where none of the reference units fully overlap with the region. Here, the upper approximation for the region is $\bar{S}_c = \{ABA, ABB, BAA, BAB, BBA\}$, while the lower approximation is an empty set, i.e. $\underline{S}_c = \{\}$.

For the case where the place name region is an exact match of a reference unit (*Case 3-a* in section 6.1.3), the upper and the lower approximations are identical, i.e. $\underline{S} = \bar{S}$. This is the case for region a, where both $\bar{S}_a = \underline{S}_a = \{AAA\}$.

Finally, region d is an example for the case where the reference unit is much larger than the region to be represented (*Case 3-b* in section 6.1.3). Here, the lower approximation is an empty set, while the upper approximation consists of the relevant reference unit, i.e., $\underline{S}_d = \{\}$, $\bar{S}_d = \{CBB\}$.

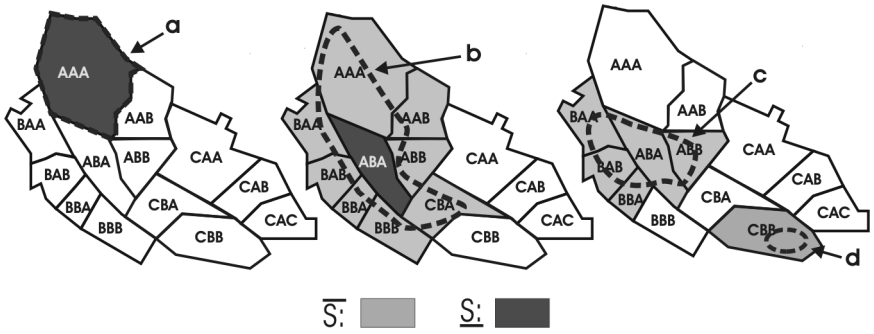

Figure 6.7: Complex spatial footprints for regions a, b, c, and d

6.2 Place Name Structures

6.2.1 Heterogeneous Models of Geographic Space

In section 3.1 we showed that the use of *place names* to disambiguate geographic locations is consistent with the intuitive human approach to conceptualize and model geographic space with natural language identifiers for geographic objects and regions. Place names are frequently used to geo-reference data and information, i.e. to annotate them with a reference to a geographic object or region. *Gazetteers*, i.e., place name lists, are tools to manage and disambiguate place names. Digital gazetteers play an important role in information retrieval, and for the access to *indirectly geo-referenced* information and data.

Most state-of-the-art gazetteers are used to manage *standardized place names*. Standardized place names have undergone a rigorous standardization process, and the components of a standardized place name often do have an official and normative character. This includes primary and variant names, as well as type classes and reference locations. Location information is given by the spatial footprint of a place name, which is typically expressed in terms of a geographic coordinate.

On the other hand individuals as well as specific user groups and organizations tend to create their own models of geographic space. Such *personalized* geographic models reflect user- and application-specific conceptualizations of geographic space. In addition to standardized place names they may include place names that do not appear in the official place name lists. We called these place names *"ad-hoc"* place names to account for the fact that they were created for a specific purpose, and that their application is limited in terms of the size of the respective user group as well as the intended life-span of the place names.

It has long been understood within the gazetteer community that there is a growing need for a solution to the problem of how to integrate standardized and ad-hoc place names [Goodchild, 1999]. Distributed, service-based geodata infrastructures [OGC, 2001b] [Kuhn et al., 2000] provide a framework to establish de-centralized gazetteer services that allow specific user groups to host customized place name lists. With the growing importance of mobile and location based applications, such distributed gazetteer services are increasingly important for location disambiguation. To illustrate the need for and the use of such distributed gazetteer services we consider the following use-case scenarios:

Use Case 1 - The Logistics Manager: The logistics manager of a large German software company (let's call them *Semi-Advanced-Products AG*) is looking for new storage facilities to add additional storage capacity in *Region North*, one of the company's five marketing regions (*Region North, Region South, Region East, Region West, Region Center*). [3] To minimize transportation costs, an important constraint for the search is that storage facility is located within the region in question, i.e. within *Region North*.

To find an appropriate property, the manager evaluates listings of commercial rentals put up by a number of different real estate agencies. Each agency uses its own (spatial) reference system to annotate their listings with spatial information: While some cite the respective administrative

[3] The company uses this spatial subdivision into marketing regions to internally organize their operations.

unit ("5000 sqm storage space available in *Landkreis Lüneburg*"), others refer to specific locations ("warehouse in the *Technologiepark*, close to *Universität Bremen*") or reference systems like postal code area ("property located within the *24xx-PLZ area*.").

To decide which offer fulfills the given locational constraint ("located within *Region North*"), the manager has to match his company-specific conceptualization of geographic space (marketing areas) with the geographic models used by the various real estate companies (administrative subdivisions, street-addresses, postal code zones, etc.). Most likely, this matching is done manually, i.e. the manager either knows from experience that a given location (e.g., "*Landkreis Lüneburg*") is located within the target area (i.e., "*Region North*"), or he has to consult a map or use GIS tools.

For an efficient computerized search of multiple (real estate) listings that are maintained by different (real-estate) agencies and are available through a number of (distributed) organizations, we have to be able to automatically map one model of geographic space onto another. An appropriate search tool has to be able to resolve a geographic region expressed in the terms of the spatial model used by the information seeker (e.g., marketing regions of *Semi-Advanced-Products*) into regions expressed in terms of the spatial models used by the information providers (e.g., place names, postal code areas, administrative units, etc.).

Use Case 2 - Natural vs. historic America: The *Smithsonian Institution* [Smithsonean, 2004], one of the largest science foundations in the United States, issues (among many others) a *Guide to Natural America* and a *Guide to Historic America*. Both guides point to documents related to the natural history of the North-American continent and to the historic development of the United States, respectively. Because many historic events are rooted in conditions related to natural history, a researcher interested in a specific topic may encounter the need to consult both guides to solve one question. An example for the close relation of natural and human history is the 1849 gold rush in California, where specific natural conditions (i.e., rich alluvial gold deposits) had a significant impact on the development of the state.

Unfortunately, each of the guides follows its own terminology with respect to the spatial subdivision of the North American continent. For example, to find information about the gold mining districts in California and Nevada, the user of the *Guide to Natural America* would have to look for the *"Far West"*. In the *Guide to Historic America*, however, this region is at the intersection between what is called the *"Pacific States"* and the *"Desert States"* (Figure 6.8).

The examples above show that in different information collections, even if they are maintained by the same organization, different models for the same geographic region may be used as a result of the different purposes and intentions of the information collections. In order to be able to obtain an integrated view on multiple heterogeneous information collections, methods for an automatic integration of heterogeneous spatial models are needed.

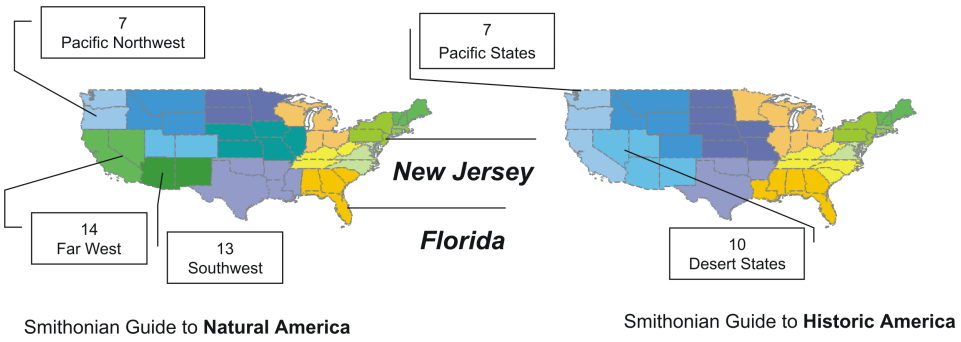

Figure 6.8: Natural and historic America according to the Smithsonian Institute

6.2.2 Integration of Personalized Spatial Models

The use-cases outlined above underline the need for systems that are able to integrate spatial models which cover the same geographic area, but were created by different users for specific purposes. Most of the large gazetteers (e.g., the *ADL* gazetteer and the *TGN*) try to integrate as many sources for place names as possible into one comprehensive and very large place name list. The *ADL* gazetteer, for example, consists of about 6 Mio. place name entries that were compiled from a number of different sources, including the U.S. Geological Survey's *Geographic Names Information System (GNIS)* [USGS, 2004] and the *Geonames Server* maintained by the National Imagery and Mapping Agency (NIMA) [NIMA, 2004].

This centralized approach works very well for lists of official and standardized place names. However, with the growth of distributed and heterogeneous information collections on the Internet and the rise of mobile and location-based services, there will be a growing need to integrate a large number of personalized gazetteers [Goodchild, 1999]. These will be small, purpose-driven, and subject to frequent change.

Most likely, the large and centralized gazetteers mentioned above will not be able (and willing) to integrate large numbers of highly volatile and potentially error-prone *"personalized"* gazetteers. What is needed are more flexible architectures and methods to integrate heterogeneous spatial models on-the-fly and as needed. We already mentioned the specifications for distributed gazetteer services developed by the *OGC* as a first step in the right direction. However, in an environment where many users want to build their own personalized gazetteers, simply providing the means for their gazetteer integration is not good enough. We also have to take into consideration the efforts that are required for their construction and maintenance. Methods are needed that help the user with this task in that they make the creation of a personalized gazetteer as simple as possible. We can define the following requirements for such methods and the resulting tools:

Intuitive and user-friendly modelling framework: The modelling effort required to create a personalized gazetteer has to be as small as possible. It should not depend on specialized expertise, such as proficiency with geographic information systems (GIS). A modelling framework for

personalized gazetteers should allow the user to define a spatial model without having to resort to un-intuitive and complex spatial representations and concepts. This includes geographic coordinates, complex polygons, coordinate projections and a number of other concepts related to the quantitative representation of geographic locations.

Interoperable and effective representation: To be useful in distributed and heterogeneous environments, a personalized gazetteers has to use an efficient and interoperable representation scheme. This representation has to be sparse and condensed in order to keep data volumes to a minimum. On the other hand, it has to be expressive enough to support meaningful spatial relevance reasoning. To be interoperable, the representation should use open formats and standards.

Flexible semantics: A personalized gazetteer reflects a user- and application-specific conceptualization of geographic space. It has to be flexible enough to support the encoding of different spatial models without the need for changes or additions to the overall representation scheme.

In section 5.2 we described our approach to build spatial reference models using discretizations of geographic space in the form of polygonal standard reference tessellations. We showed how crisp and vague geographic (place name) regions can be approximated using such discrete reference models. In the following section, we will show how these approximations can be used to build flexible gazetteers, or *Place Name Structures (PNS)*, that fulfill the requirements outlined above.

6.2.3 Architecture of a Place Name Structure

A *Place Name Structure (PNS)* $\mathcal{P}$ is a central component of a personalized gazetteer [Vögele et al., 2003b]. It represents a conceptualization of a finite sub-section of geographic space with the help of place names. Through the use of (simple and complex) qualitative spatial footprints (section 6.1), the place names in a place name structure are referenced to a discrete partitioning of geographic space provided by a qualitative spatial reference model (section 5.2). This representation supports effective graph-based algorithms that can be used to reason about the relative spatial relevance of individual place names.

Analogous to a qualitative spatial reference model, a place name structure is used to encode both the topological and the partonomic relations of its components. In fact, each qualitative spatial reference model that is based on a polygonal standard reference tessellation of named geographic regions (like administrative subdivisions) can be seen as a special case of a place name structure. Analogous to a qualitative spatial reference model, each place name structure has a *horizontal* (i.e., topologic) and a *vertical* (i.e., partonomic) dimension (Figure 6.9).

The Horizontal Dimension of a PNS

The horizontal dimension of a place name structure relates to the spatial extensions of place name regions, and to the topologic relations that exist between these regions. The spatial extension of a place name region is given by its (simple

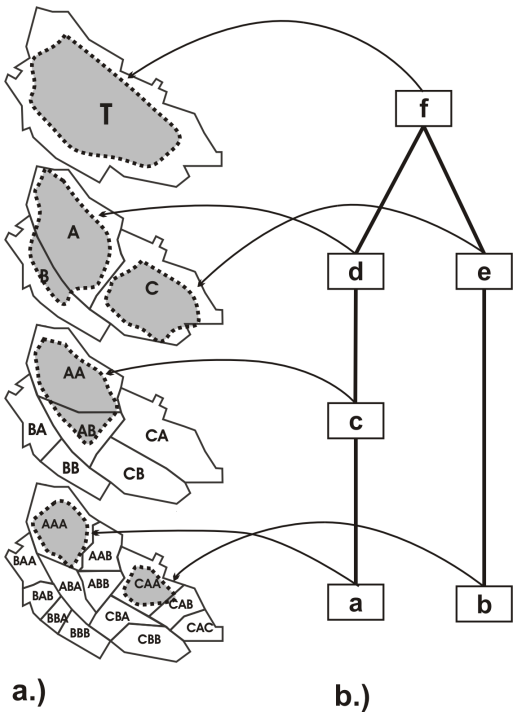

Figure 6.9: Architecture of a place name structure a.) spatial footprints, b.) hierarchical partonomy

or complex) qualitative spatial footprint. A qualitative spatial footprint defines a place name region *extensionally* using spatial indices taken form a qualitative spatial reference model.

Using their qualitative spatial footprints, we can draw inferences about the spatial relations between place name regions. For one, set-theoretic operations allow us to make assumptions about the topologic relations of regions. On the other hand, we can compute qualitative distance on basis of the underlying spatial reference model. We cannot compute neighborhood distances directly because, in contrast to the units of the underlying spatial reference model, the regions of a place name structure do not have to form a tessellation. We may as well find regions that overlap and regions that are disconnected.

We can only infer the topological relations between two place name regions and reason about their spatial relevance if the two regions are comparable. In general, only place name regions that are approximated using the same common frame of reference (i.e., the same qualitative spatial reference model) can be compared. In a place name structure, all regions are extensionally defined using spatial indices taken from the same qualitative reference model. Consequently, the fact that two place names a and b belong to the same place name structure $\mathcal{P}$ implies that their spatial footprints S_a and S_b are based on the same spatial reference model $\mathcal{S}$.

The Vertical Dimension of a PNS

The vertical dimension of a place name structure encodes the hierarchical organization of the respective place name regions. To organize geographic objects denoted by place names along the lines of hierarchical structures follows the human preference to conceptualize geographic space in terms of hierarchical partonomies (see section 4.2).

However, in contrast to the regular and standardized hierarchy of a qualitative spatial reference model, the hierarchy of a place name structure may be irregular and unbalanced. A hierarchical classification scheme may not exist, and the hierarchy may directly reflect the mereotopologic relations between regions: A region b is said to be *part-of* a region a if b is perceived as a sub-region of a.

In our view, a method for the intuitive modelling of geographic space should account for such perceived mereotopological relations. It should provide methods to add new regions to a model on basis of their mereotopological relations with other regions. For example, it should be possible to add the region *"Lüneburger Heide"* (a natural region) to a spatial model of Europe by stating that this region is *"part-of Northern Germany"*, i.e. another region already defined in the model.

6.2.4 Extensionally- and Intensionally-Defined Regions

Place name regions that are extensionally defined as qualitative spatial footprints are the basic building blocks of a place name structure. The degree to which the spatial extent of the regions is known can vary considerably. On one hand, there are regions with well-defined, exact boundaries that are available as digitized polygonal vector data. An example for such data are the *Natural Regions of Germany*, which have been compiled by an expert team and are published by the federal agency in charge of cartographic products, the *Bundesamt für Kartographie und Geodäsie (BKG)*[BKG, 1994][4]. On the other hand, the spatial extent of a region may be very vague and the region may have indeterminate boundaries that cannot be cast into any exact polygonal shape. Typical example for such regions are neighborhoods in cities (e.g., "das Viertel" in Bremen) of which the exact extent is not well defined, and for which polygonal representations do not exist.

To facilitate the intuitive and user-friendly creation of spatial models, place name structures have to be able to model not only well-defined place name regions, but also vague regions with indeterminate boundaries. To fulfill this requirement, two fundamentally different methods to define a place name region in a place name structure have to be available:

Extensional definition: A place name region may be extensionally defined by specifying a simple or complex qualitative spatial footprint. A method has to be provided to infer the position of the region within the hierarchy of the place name structure based on its spatial footprint.

Mainly place name regions with with well-defined boundaries can be defined extensionally. If digital polygonal data are available, GIS-based

[4]Note, however, that even in this official document based on many man-years of work, the exact boundaries for some natural regions could not be established!

methods for an automated mapping of the region onto the spatial reference model can be applied. If polygonal data are not available, the manual definition of the spatial footprint (even complex ones) should be as simple and user-friedly as possible.

Intensional definition: It should be possible to add a region to a place name structure without having to explicitly define a spatial footprint, i.e., by simply defining its position within the partonomic hierarchy of the place name structure. A method to approximate the spatial footprint of a region as a function of its position within the PNS hierarchy should be provided.

Mainly vague place name regions with indeterminate boundaries and complex regions for which no polygonal data are available are defined intensionally.

To be able to compute the relative spatial relevance of place name regions, both their extensional and their intensional definitions have to be known. This implies that for an intensionally defined place name a spatial footprint, and for an extensionally defined place name region a position within the PNS hierarchy has to be inferred.

Adding Extensionally-Defined Regions

To define a place name region extensionally, it has to be assigned a qualitative spatial footprint and a position within the hierarchy in the place name structure:

Definition of a Qualitative Spatial Footprint: Both an automatic and a manual definition of this footprint are possible, depending on user preferences and the available data. If both the spatial reference tessellation and the place name region are available as digital polygonal vector data, standard GIS software can be used to map the place name region onto the reference tessellation. If this is not the case, the mapping has to be done manually.

In any case, while a simple spatial footprint for the place name region is obtained through a single mapping procedure, the definition of a complex spatial footprint is a 2-step process: First, the upper approximation of the region has to be determined by selecting all reference units that intersect with the region. Then the lower approximation is specified by selecting those reference units that are fully contained in the region.

The resulting qualitative spatial footprint is encoded in an XML-based format very similar to that used for the underlying qualitative spatial reference model (see Figure B.1 for an example).

Inferring the hierarchical position of a place name: If we add an extensionally-defined region to an existing place name structure, the position of this region within the hierarchy of the structure has to be determined. This can be done manually, for example by specifying the respective *RCC* relations, or it can be computed through an evaluation of the mereotopologic relations of the new region with respect to the other regions already in the PNS. However, it is important to note that the second approach implies a purely *"spatial"* view on the partonomic

relations of place names. It is neither applicable for spatial footprints that consist of non-connected clusters of reference units, nor for place name regions in which *part-of* relations are defined functionally rather than mereotopologically.

In the special case of a PNS that consists of regions approximated as simple qualitative spatial footprints, the automatic determination of partonomic relations is based on the eight *JEPD* relations defined in *RCC-8* (see section 4.2.3). In the general case of a PNS that consists of regions approximated by complex qualitative spatial footprints, the evaluation is more difficult. Depending on the underlying assumptions, the complexity of topological relations that can be described between such regions varies considerably. Cohn et al. [Cohn and Gotts, 1996a] [Cohn and Gotts, 1996b] distinguish 46 possible relations between vague regions of the *"egg-yolk"* type (see section 4.3.2).

A much simpler approach is taken by Worboys [Worboys, 1998] who, for two regions a and b represented as rough sets (see section 4.4), developed a three-valued logic where a is *definitely-part-of* b if all of the four relations hold:

$$\bar{S}_a \subseteq \bar{S}_b, \bar{S}_a \subseteq \underline{S}_b, \underline{S}_a \subseteq \bar{S}_b, and \underline{S}_a \subseteq \underline{S}_b. \tag{6.5}$$

The region a is *definitely-not-part-of* b if all four relations are false. For all other possible combinations, a is said to be *maybe-part-of* b .

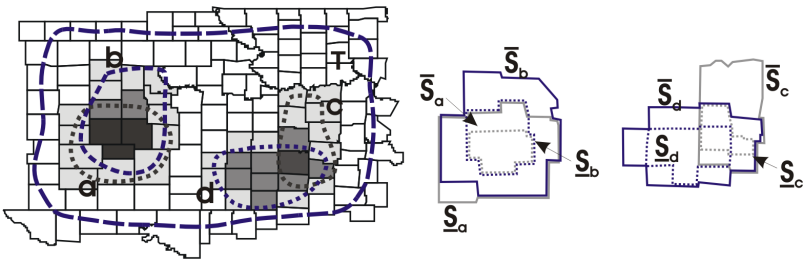

Figure 6.10: Configurations of extensionally-defined place name regions

For the trivial case where the *boundary region* $B_a = \bar{S}_a \cap \underline{S}_a$ of a discretized region a is not more than one reference unit thick, we can make the assumption that B_a is small compared to $\bar{S}_a$[5]. Under this assumption, we can simplify the approach adopted from Worboys even further by saying that only if a is *definitely-not-part-of* b, the two regions are disjunct.

This allows us to develop a practical solution to infer the partonomy of place name regions on the basis of their discretized approximations: We can say that $PP(a, b)$ holds as long as the lower approximation of a is a subset of the upper approximation of b. Or, expressed in terms of the reference units of a polygonal tessellations:

$$PP(a, b) \rightarrow \{r | r \in \underline{S}_a \wedge r \in \bar{S}_b\} \tag{6.6}$$

[5]This refers to *Case 1* in section 6.1.3 and represents the most frequent case encountered in real-world applications.

Applying equation 6.6 to the five place name regions a, b, c, d and t depicted in Figure 6.10, we can infer the following partonomic relations: $PP(a,t)$, $PP(b,t)$, $PP(c,t)$, $PP(d,t)$, and $PP(a,b)$. To obtain a meaningful hierarchical partonomy that can be encoded in a *directed acyclic graph (DAG)*, these partonomic relations have to be ordered, and redundant arcs have to be removed. Because the *spatially-part-of* relation is transitive, we can say that

$$a \sqsubseteq b \wedge b \sqsubseteq t \longrightarrow a \sqsubseteq t \tag{6.7}$$

As a consequence, $a \sqsubseteq t$ does not have to be stated explicitly, i.e., it represents redundant information. To obtain a redundance-free representation, we set the rule that every region can only be spatially part-of its direct parent.

Based on the definition of *spatially-part-of* (see section 5.2), we can say that for three regions a, b and t, if a is spatially part-of b as well as of t, and b is smaller than (i.e., a subset of) t, then b is the direct parent of a. Consequently, the relation $a \sqsubseteq t$ is redundant and can be removed.

The following sequence of procedures can be used to compute the partonomic hierarchy and remove redundant information:

1. Compute partonomic relations between all discretized regions. The result of this operation is encoded as a directed acyclic graph G.
2. Compute the transitive reduction of G, i.e. the minimal graph G^* with the same connectivity as G[6].

The result is a hierarchical place name structure like the one shown in Figure 6.11.

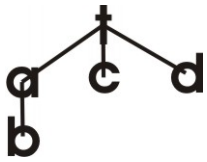

Figure 6.11: Inferred hierarchy of place name regions

Thus the position of a place name x within the hierarchy of a place name structure $\mathcal{P}$ can be inferred on the basis of it's qualitative spatial footprint, S_x, provided that S_x and $\mathcal{P}$ are based on the same spatial reference model $\mathcal{S}$.

Adding Intensionally-Defined Regions

In section 6.2 we introduced place name structures as a step forward towards intuitive tools to model geographic space. An important provision to reach this goal is to provide methods to handle *incomplete spatial information*. For

[6] A number of algorithms can be found in the literature to compute the transitive reduction of a directed graph (e.g., see [Turau, 1996]).

example, if the boundaries of a place name region are indeterminate, the user of a spatial modelling tool should be able to specify this region not by giving an extensional definition, but rather by stating a few relations to known place names such as: *"The (new) place name b is part of the (old) place name a"*. For this purpose, we have to provide the option to define place name regions within a place name structure *intensionally*.

The problem is that incomplete descriptions of configurations of spatial regions are generally compatible with a large number of geometrical realizations. Consider the *RCC-5* formulas $PP(a, d)$, $PP(b, d)$, and $PP(c, d)$ stating proper-part relations between regions a, b, c, and d. Each configuration of the C_1, C_2, and C_3 shown in Figure 6.12 constitutes a valid geometrical model of the three formulas.

To circumvent problems with valid but conflicting spatial models most computational approaches (e.g. constraint satisfaction systems or logical theorem provers) avoid to construct such models explicitly. [7] Human problem-solvers, on the other hand, proceed differently: when solving spatial relational inference tasks, they try to build a mental model that represents a specific configuration compatible with the information that is given to them [Knauff et al., 1998]. It seems as of the first-built mental model dominates the further reasoning process: relations valid in such a preferred model tend to be considered as being valid generally, i.e. valid in all other models too. Mental model-based reasoning therefore shows characteristic reasoning errors (preference biases) [Vögele et al., 2003b].

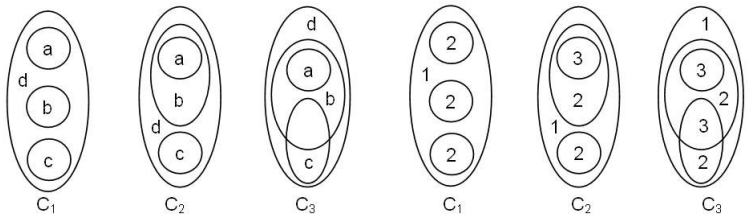

Figure 6.12: Spatial configurations and their boundary overlap complexity, from [Vögele et al., 2003b]

Based on an approach developed by Schlieder and first described in [Vögele et al., 2003b], we use the cognitive preference of certain models over others as a feature of an intuitive spatial modelling tool. We showed that this feature can be exploited to reduce the amount of what needs to be said during model specification. A speaker stating that three regions a, b, and c are part of a region d without adding any further information about the relationship between a, b, and c may trigger a pragmatic reasoning process in the listener which attempts to reconstruct the configuration intended by the speaker. Very likely, the process will come up with a configuration similar to C_1 rather than to C_3. To describe C_2, the speaker would have to make explicit that a is a part of b. To describe C_3, even further information would be necessary. In other words, if a simple configuration is intended, it can be described with a few statements, whereas complex configurations need more details.

[7] A rare exception is the work of [Frank, 1996] who describes the advantages of a preference-based approach to reasoning with cardinal directions.

What determines the simplicity of a spatial configuration and which are the preferred mental models in spatial relational inference tasks? The issue has been studied most comprehensively for the spatial version of the one-dimensional interval calculus [Rauh et al., 2000]. However, this line of research has not yet addressed the *RCC* calculus, which is the most interesting for spatial relevance computation. In [Vögele et al., 2003b], we briefly summarized the main results and derived from this review a hypothesis about preferences for *RCC-5*.

The best explanation for preferences currently available is that they are the result of spatial chunking in working memory which significantly reduces memory load for the preferred mental models, whereas other models require more memory resources. Details of the spatial chunking process have been described by [Schlieder, 1999]. Two of its characteristic features are:[8]

1. A configuration with singularities, i.e. touching objects, never acts as preferred model.

2. There is a preference to avoid overlaps where possible and to keep objects disjoint.

We used these principles to develop what we called an *"educated guess"* about preferences among *RCC* relations [Vögele et al., 2003b]. Because of (1), we restricted attention to *RCC-5*, the region-based topological calculus that does not treat configurations with singularities as special cases (TPP, $TPPI$, and EC in *RCC-8*). In order to deal with (2), we proposed the following measure for boundary overlap complexity in a configuration of connected regions without holes in the Euclidean plane:

Definition 6.1
$T(R)$ is the tessellation of the plane induced by the boundaries of the regions $R = \{r1, ..., rn\}$. Each cell c of $T(R)$ is assigned an overlap number $o(c)$ which is the number of regions from R of which c is a part. The boundary overlap complexity $bc(R)$ is the sum of $o(c)$ for all c from $T(R)$.

For an example refer to Figure 6.12 where the overlap numbers are shown for the tessellations arising from configurations C_1, C_2, and C_3. The following boundary overlap complexities are found: $bc(C_1) = 7$, $bc(C_2) = 8$, $bc(C_3) = 11$. Sorting according to increasing complexity (diminishing simplicity) results in the ordering $C1$, $C2$, $C3$ which matches well with the intuitive notion of complexity of an arrangement of regions. With the measure for boundary overlap complexity, we can now state our hypothesis about preferred models for *RCC-5* in precise terms: In a set of models, the models with least boundary overlap complexity are going to be preferred.

With respect to place name structures and intuitive spatial modelling, this notion of preferred models helps to solve the problem of processing a specification of a new place name that is input into the system by a user. Consider the situation depicted in Figure 6.13. Three place names are already defined in the place name structure with extensions a, c, and d all of which are aggregates of reference units in a spatial reference model. To the right of the

[8]Note that these principles do not suffice to account for all preferences empirically found. They fail, for instance, to explain how the process of mental model construction depends on the order in which the information is processed. An explanation accounting for almost all preferences is given in Schlieder [Schlieder, 1999].

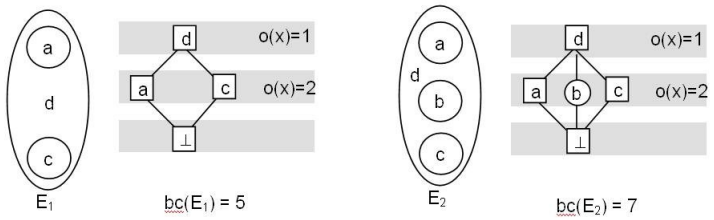

Figure 6.13: a): Starting configuration; b): Result of processing $PP(b, d)$, from [Vögele et al., 2003b]

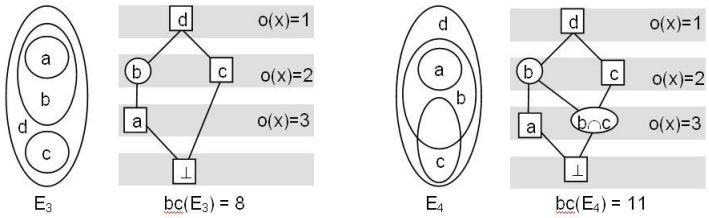

Figure 6.14: c): Result of processing $PPI(b, a)$; d): Result of processing $PO(b, c)$, from [Vögele et al., 2003b]

geometrical model, the partonomic structure is depicted. It is a directed acyclic graph representation of the proper-part relations holding between the cells of the tessellation induced by the boundaries of the regions a, c, and d. The user starts to specify a new place name d and the system incrementally processes the user's input. As first input a *RCC-5* formula - or its natural language equivalent - stating $PP(b, d)$ is processed. The new region b is integrated into the partonomic structure *in such a way that the boundary complexity increases least* (Figure 6.13b). Similarily, the input $PPI(b, a)$ is processed. It requires making a proper part of b, which involves a local rearrangement of the partonomic structure (Figure 6.14c). Finally, after processing $PO(b, c)$, the geometrical model shown in (Figure 6.14d) is obtained.

Note that the algorithm proceeds incrementally following a greedy strategy which always minimizes boundary overlap complexity. It may well be that at a certain stage, an input from the user cannot be integrated into the model so far constructed, although there exists a model satisfying the constraints represented by the place name structure as well as those stated by the user. This failure of not always finding a solution when a solution exists does not come as an surprise given the complexity of the underlying constraint satisfaction problem which is NP-complete. However, the algorithms works - just as human model-based reasoning - as a very fast approximation to the instantiation problem. With the kind of simple place name specifications that users typically provide, we do not consider the incompleteness of the algorithm a real problem for application.

Estimating the Extension of Intensionally-Defined Regions

Using the complexity measure described above, a set of simple rules can be derived that allow us to infer rough approximations for intensionally-defined

place names. [9]

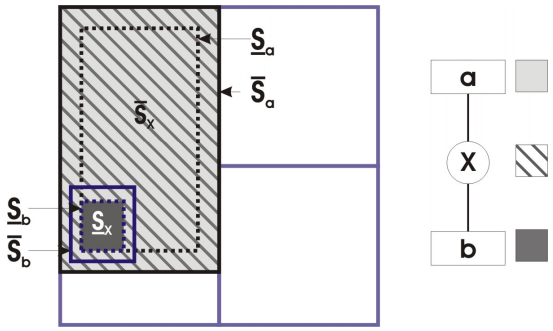

Figure 6.15: Upper and lower approximation of region x

In Figure 6.15, we give the example of a simple place name structure that consists of 2 extensionally-defined place name regions a and b. We can add a new place name region x by specifying two *RCC-5* relations, namely $PP(x, a)$ and $PP^{-}1(x, b)$. We can then say about the spatial extent of x that it must be at least as large as the lower approximation $\underline{S}_b$ of b, but cannot be larger than the upper approximation $\bar{S}_a$ of a. Consequently, we can say about the upper and the lower approximation of x that $\bar{S}_x \subseteq \bar{S}_a$, and $\underline{S}_x \supseteq \underline{S}_b$.

The approximation of x thus derived in the example is very coarse. In fact, it may differ too much from the *"true"* extent of x to support any reasonable inferences. In a less simplified spatial model, our estimate for the extension of x will be better due to the existence of other extensionally-defined regions. It can be further improved through the specification of additional spatial relations to other place names and by adding new extensionally-defined place name regions to the PNS.

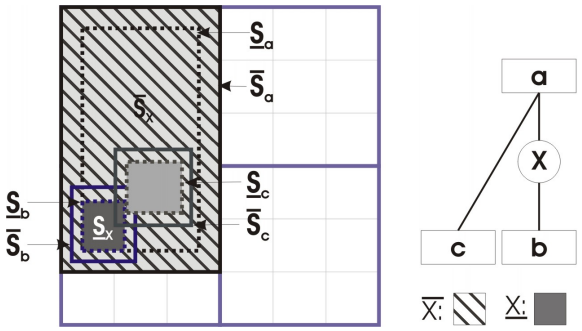

Figure 6.16: Refining the approximation of region x by adding a region c

In Figure 6.16, a new extensionally-defined region c, which is a child of a, is added to the structure. While the lower approximation of x remains the same as in Figure 6.15, the upper approximation of x is refined by the addition

[9]Note: Intensionally-defined place names cannot be used to define a complete place name structure. A valid PNS has to exist before a new place name can be added through intensional definition!

of c: $\bar{S}_x$ cannot be larger than the upper approximation of a minus the lower approximation of c.

$$\bar{S}_x = \bar{S}_a \cap \underline{S}_c \tag{6.8}$$

Within the hierarchy of a PNS, the upper approximation $\bar{S}_x$ of an intensionally-defined place name x can be estimated to be equal to the upper approximation of its parent node p_P, minus the union of the lower approximations of all children of p_P, with exception of the children of x.

$$\bar{S}_x = \bar{S}_{p_P} - \bigcup \underline{S}_p\{p|PP(p, p_P)\} \backslash \underline{S}_p\{p|PP(p, x\} \tag{6.9}$$

The lower approximation $\underline{S}_x$ of x can be estimated to be equal to the union of the lower approximations of all place names p_i that are *part-of* x.

$$\underline{S}_x = \bigcup \underline{S}_p\{p|PP(p, x)\} \tag{6.10}$$

For regions that do not have children (i.e., regions that are leaves of the hierarchical tree), the lower approximation is undefined. In this case we set the lower approximation equal to the upper approximation. Likewise, for regions that do not have parents (i.e., the region that is the root of the hierarchical tree) we set the upper approximation to the union of the upper approximations of all their children.

By applying simple set arithmetics and the rules established in equations 6.9 and 6.10, we can approximate the extension of any place name x. This allows us to find the extensional definition of any intensionally-defined place name added to a place name structure. The result is a place name structure where all regions do have a qualitative spatial footprint.

6.3 Application Example: Regions of *Franken*

6.3.1 Polygonal Projection

In a real world example we demonstrate the discrete approximation of place name regions and the computation of a partonomic hierarchy for such regions. For this purpose, we have chosen an area in the north-eastern part of the state of *Bavaria* in Germany. Our spatial reference system is a polygonal tessellation of *"Gemeinden"*, i.e., administrative units on *NUTS* level 5 (see Figure 5.7). We want to use this reference system to create qualitative spatial footprints for a number of *natural regions*. These place name regions have well-defined boundaries and are even available as polygonal digital data [BKG, 1994] (Figure 6.17).

Because both the reference tessellation and the regions are available as digital data, we can use a standard GIS to project the polygons onto the reference tessellation. For each region, we obtain a spatial footprint with an upper and a lower approximation. Figure 6.18 shows the upper and the lower approximation of the *Frankenwald (FW)* region. Figure 6.19 depicts the upper and lower approximation of the *Münchberger Hochfläche (MH)* region. As we can see in Figure 6.19c , the lower approximation of region MH is a subset of the upper approximation of region FW. Applying the rules described above we can conclude that the relation $PP(MH, FW)$ holds. Consequently, MH is a

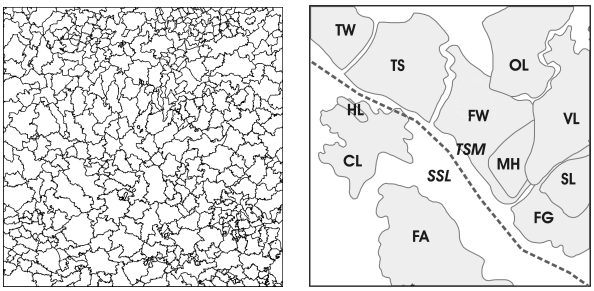

Figure 6.17: a) Polygonal Standard Reference Tessellation (*NUTS* 5), b) Polygonal representation of natural regions (source: [BKG, 1994]

child of FW. Following the same procedure, we can compute the partonomic structure for all regions shown as polygons (solid line) in Figure 6.17 and obtain the hierarchical structure of extensionally-defined place names (indicated by rectangular boxes) in Figure 6.20.

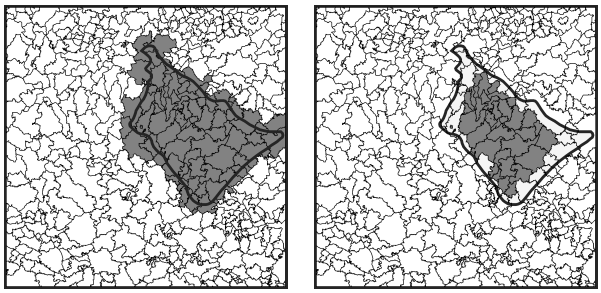

Figure 6.18: (a) Upper and (b) lower approximation for the place name region *Frankenwald* (FW)

6.3.2 Defining New Regions

In a next step, we add two super-regions, TSM and SSL to the place name structure developed above. In Figure 6.17b, these regions are outlined by a dotted line to express the fact that their exact extent is not known. The only information available is the identity of the sub-regions that TSM and SSL consist of.

We add TSM and SSL to the structure by specifying their partonomic relations with respect to the other regions. In the case of SSL we use to following RCC relations: $PP(CL, SSL)$,$PP(FA, SSL)$,$PP(HL, SSL)$, and $PP(SSL, T)$. Following the rules established in section 6.2.4, we can then estimate the lower and upper approximations of TSM and SSL based on the lower and upper approximations of their children.

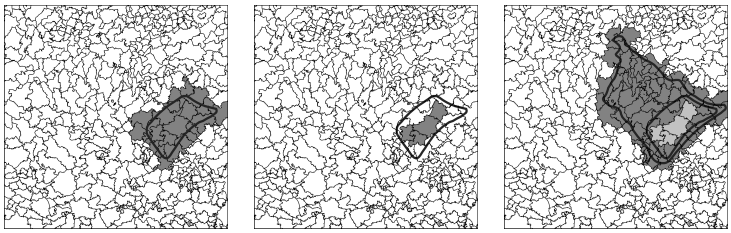

Figure 6.19: (a) Upper and (b) lower approximation for the place name region *Münchberger Hochfläche*(MH), (c) MH is a proper part of FW

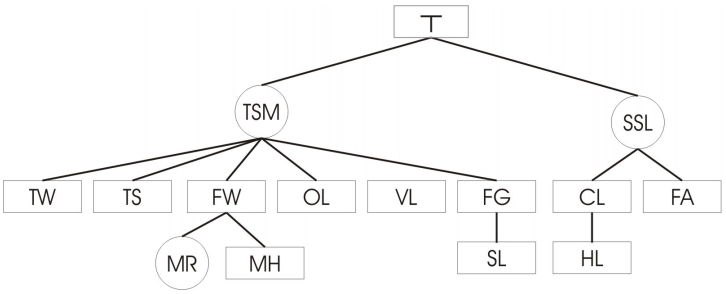

Figure 6.20: Hierarchy of place name structure of natural regions

Finally, we define a new region, the *Frankenwald mining region (MR)* and add it to the place name structure. About this region we only know that it belongs to the north-western part of the *Frankenwald (FW)*. We add MR to the structure as a child of FW by specifying $PP(MR, FW)$ (Figure 6.20). Note that MR is modelled to be a sibling of MH because our algorithm tries to minimize overlap complexity. If MH were to be a sub-region of MR, we would have to explicitly state this fact. Following the rules defined in section 6.2.4, the upper approximation of MR is estimated to be equal to the upper approximation of FW, minus the lower approximation MH. Because MR has no children, its lower approximation is undefined and defaults to the upper approximation.

The final result is the place name structure shown in Figure 6.20. It represents the hierarchical structure of the place name regions based on both the evaluation of the mereo-topological relations between extensionally-defined regions (shown as rectangular boxes), and the inclusion of intensionally-defined regions (depicted as circles). For each place name region in this structure, a user-defined or computed discrete approximation is available.

Chapter 7

Reasoning about Spatial Relevance

7.1 A Simple Metric to Compute Spatial Relevance

In section 4.1.2 we identified topological, partonomic, ordinal and metric relations between geographic objects as the key parameters that have to be evaluated in order to rank place name regions based on spatial relevance. The qualitative spatial reference models and place name structures described in the previous chapter encode information about all four parameters:

Topological information is encoded as the the neighborhood relations of polygons in the connection graph.

Partonomic information is provided by the decomposition hierarchy of the tessellation, which is encoded in a decomposition tree.

Distance information is given as a semi-qualitative measure derived from the neighborhood relations of polygons in the tessellation, which is given by the shortest-path between nodes of the respective connection graph.

Ordinal relations can be derived from the the relative ordering of neighbors of a given polygon encoded in a connection graph.

We argue that this information is sufficient to make a judgement about the relative spatial relevance of reference units in a qualitative spatial reference model. Our goal is to compute a relative spatial relevance $\sigma(r_q, r)$ for two reference units r_q and r. Based on this relative measure, we want to develop a partial ordering of all units in a spatial reference model with respect to a unit of interest.

For example, given three locations r_q, r_i and r_j, we want to know whether r_i or r_j is more relevant with respect to r_q. We may obtain an ordering of the spatial relevance σ like $\sigma(r_q, r_q) > \sigma(r_i, r_q) > \sigma(r_j, r_q)$.

In section 4.1 we developed a notion of spatial relevance that is closely linked to the concept of proximity: The closer two (geographic) objects are, the higher

their mutual spatial relevance. Consequently, among spatial relations identified as key-parameters of spatial relevance, spatial proximity (i.e., *horizontal distance*) expressed as a qualitative distance measure plays a prominent role. Another important measure is the proximity of two reference units within the hierarchical structure of a spatial reference model (i.e., *vertical distance*). We use this dual notion of proximity to develop a simple metric for the computation of the relative spatial relevance between two units of a polygonal tessellation.

7.1.1 Horizontal Distance

In a polygonal tessellation, a semi-qualitative measure for the horizontal spatial distance of two polygons can be derived from their minimal *neighborhood distance*. For two polygons P_i and P_j, the neighborhood distance is the least number n of *adjacent* polygons $P_1, ..., P_n$ that have to be traversed to get from P_i to P_j. The notion of adjacency used in this context is that of *(simple) adjacency*. In contrast to *full adjacency*, (simple) adjacency considers only such polygons to be adjacent that share at least a fraction of a face (see also section 4.2).

In a graph-based abstraction of a polygonal tessellation, the horizontal (neighborhood) distance $\delta(r_i, r_q)$ of two reference units r_i and r_q reflects the *shortest path* between the two respective nodes in the connection graph.

$$\delta(r_i, r_q) = SP(r_i, r_q) \tag{7.1}$$

Neighborhood distances in weighted and un-weighted graphs can be computed using standard methods like Dijkstra's shortest path algorithm. In an un-weighted graph, $\delta(r_i, r_q)$ is equal to the *order of neighborhood* of r_i with respect to r_q. Consequently, $\delta(r_i, r_q)$ is always a natural number between 0 (for $i = q$) and N, where N is the maximum order of neighborhood with respect to q that is possible in a given connection graph. In a weighted graph, which we may want to use non-equally connected geographic regions (see below), $\delta(r_i, r_q)$ can assume an arbitrarily large number depending on which weights were assigned to the edges [1]. To make horizontal distances comparable, we normalize $\delta(r_i, r_q)$ with $\delta(r_N, r_q)$ (i.e., with respect to the longest path in the graph that originates in q) in both weighted and un-weigthed graphs.

$$\delta_{NORM}(r_i, r_q) = \frac{\delta(r_i, r_q)}{\delta(r_N, r_q)} \tag{7.2}$$

In the example shown in Figure 7.1, the maximum order of neighborhood N is 4 as denoted by *path A*. To reach polygon BBB from polygon AAA, we have to traverse at least 2 other polygons , including BBB (*path B*). Consequently, $\delta(AAA, BBB) = 2$.

Computing Horizontal Distances for Multi-Resolution Reference Models

In section 5.2.3 we described a multi-resolution spatial reference model that consists of a set of layered polygonal tessellations derived from a recursive decomposition of a polygonal standard reference tessellation, and the decomposition

[1] Negative labels are not allowed.

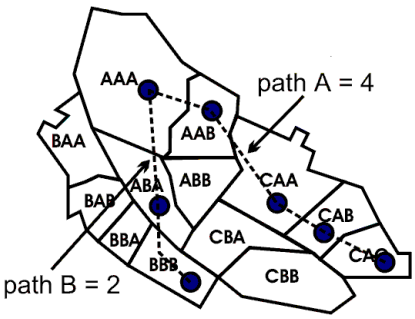

Figure 7.1: Neighborhood (*"horizontal"*) distance in a polygonal tessellation

tree reflecting the hierarchical structure of the $\mathcal{S}$. Each layer of such a model is represented through a connection graph and refers to a distinct level L of resolution, or granularity.

Given a specific reference unit $r_q \in \mathcal{S}$ which belongs to a connection graph C_L at resolution L, we can use equations 7.1 and 7.2 to compute the normalized horizontal distances relative to r_q of all units $\{r_i, ..., r_j\}$ that belong to C_L. We call the set of horizontal distances $\{\delta(r_q, r_1), ..., \delta(r_q, r_n\}$; $r_q, r_i \in C_L$ the *horizontal distance field* $\mathcal{HD}_{\mathrm{q}}$ of r_q within $\mathcal{S}$.

The horizontal distances (relative to r_q) of reference units located on resolution levels above or below L can be computed by back- and forward-propagation of $\mathcal{HD}_{\mathrm{q}}$. Units that belong to connection graphs at higher levels of resolution $L + n$ directly inherit horizonal distance from their parents. This is possible because, by definition, each unit r on resolution L decomposes into a set of units $\{r_1, ...r_n\}$ on resolution $L + 1$.

A reference unit $r_x \in C_{L-1}$ can be decomposed into a set of reference units $\{r_1, ..., r_n\} \in C_L$. Because the children of r_x are on the same level L as r_q, they belong to the distance field $\mathcal{HD}_{\mathrm{q}}$ of r_q. The horizontal distance of r_x relative to r_q can therefore be computed as a function of the respective distances of it's children given in $\mathcal{HD}_{\mathrm{q}}$. A straightforward heuristic is to use the arithmetic mean of $\{\delta(r_1, r_q), ..., \delta(r_n, r_q)\}$ as an approximation of $\delta(r_x, r_q)$:

$$\delta(r_x, r_q) = \frac{\sum_1^n \delta(r_i, r_q)}{n}; \forall r_i | r_i \in C_L \wedge r_i \sqsubseteq x \tag{7.3}$$

Context-Dependent Horizontal Distance

In a connection graph, polygons are represented as nodes, with the arcs connecting the nodes representing adjacency between polygons. In the example tessellation shown in Figure 7.1, the arcs of the connection graph are not weighted, which implies that all polygons in the tessellation are equally connected. Here, the neighborhood distance between two nodes of the graph can be interpreted as an a semi-qualitative measure that is a rough approximation of the *Euclidean distance* between the central points of two polygons in a tessellation[2].

[2]This is true at least in the case where the reference units of a polygonal reference tessellation are homogeneous, i.e. do have approximately the same size. Most standard reference tessellations are in fact homogeneous, at least on a regional level.

In section 4.2.2 we have seen that distance is highly context dependent. Whether two locations are perceived as being *"close"* or *"far"* from each other depends not only on Euclidean distance, but (among others) also on their *accessibility*. In practical applications, the lack of appropriate means of transportation and road connections, or the existence of (natural or artificial) barriers can cause two regions that are connected and direct neighbors in the topological sense be less accessible (and therefore spatially less relevant) than two regions that are further apart.

An example is shown in Figure 7.2: Here we look at accessibility within the practical context of a classical way-finding problem. The three regions a, b, and c are part of a tessellation of administrative units. Because all three regions are first-order neighbors, the neighborhood distances between them are equal (i.e., $\delta(a,b) = \delta(a,c) = \delta(b,c) = 1$), and an algorithm based on the evaluation of neighborhood distances will assign b and c the same spatial relevance relative to a.

However, within the context of a motorized tourist located in region a and looking for a hotel to stay overnight, we may have to re-evaluate the spatial relevance ranking. To do so, we need to add context specific knowledge to the scene. If we look at the road network connecting the three regions we see that while regions a and b are connected by a wide and comfortable motorway, the only road connection between regions a and c is a small and windy road. This is due to a high mountain range located between regions a and c.

Looking at this (road) map, our motorized tourist would intuitively know that region b is much faster and easier to reach than region c. He would decide that, within the given application context (i.e., having to find a place to stay for the night), region b is the better choice to go, i.e., that region b is spatially more relevant than region c.

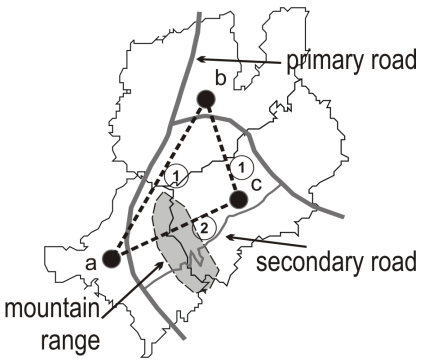

Figure 7.2: Non-equally connected geographic regions

To encode context-dependent knowledge about accessibility in our representation of polygonal tessellations we use *weighted connection graphs*. In a weighted connection graph, the context-specific accessibility of two nodes is represented by a numerical weight attached to the arc connecting these two nodes. Using Dejkstra's algorithm, the shortest path between the nodes is then a function of the weighted neighborhood distances, where the degree of accessibility decreases with an increasing numerical weight.

In our example (Figure 7.2), we use a weight of 1 for polygons that are

connected by primary roads, and 2 for polygons connected by a secondary road. With respect to a, the horizontal distances of the other polygons will then be $\delta(a, b) = 1$ and $\delta(a, c) = 2$, i.e., b is spatially more relevant to a than c.

7.1.2 Vertical Distance

Two polygons r_i and r_q that are part of a spatial reference model are also part of a decomposition hierarchy. The graph distance between r_i and r_q in the decomposition hierarchy is called *vertical distance* $\nu(r_i, r_q)$. This distance reflects the intrinsic semantics of the underlying polygonal reference tessellation formalized in the decomposition tree. For three units r_q, r_a, and r_b that belong to a spatial reference model of administrative units $\mathcal{S}$, we can say that r_a is *administratively closer* (and, as a consequence, administratively more relevant) to r_q if $\nu(r_a, r_q) < \nu(r_b, r_q)$.

We compute the vertical distance $\nu(r_i, r_q)$ between two units r_i and r_q of the same reference tessellation by starting in r_q and recursively traversing the decomposition tree until the first common parent of both units, r_P, is reached. $\nu(r_i, r_q)$ is then defined as the shortest path $SP(r_P, r_q)$ from r_q to r_P.

$$\nu(r_i, r_q) = SP(r_P, r_q) \tag{7.4}$$

In the decomposition tree $\mathcal{D}_\mathcal{S}$ of a spatial reference model $\mathcal{S}$, the maximum vertical distance is given by the depth D of $\mathcal{D}_\mathcal{S}$. Analogous to the horizontal distance, the vertical distance is normalized with D:

$$\nu_{norm}(r_i, r_q) = \frac{SP(r_P, r_q)}{D} \tag{7.5}$$

The decomposition hierarchy shown in Figure 7.3 refers to the simplified polygonal tessellation depicted in Figure 5.3. Using 7.4, the vertical distance $\nu(AAA, AAB)$ is computed to be $1/3 = 0.33$ because both units are in the same sub-tree and the distance to the root T is 3. Accordingly, the vertical distance $\nu(CBA, AAB)$ is $3/3 = 1.0$.

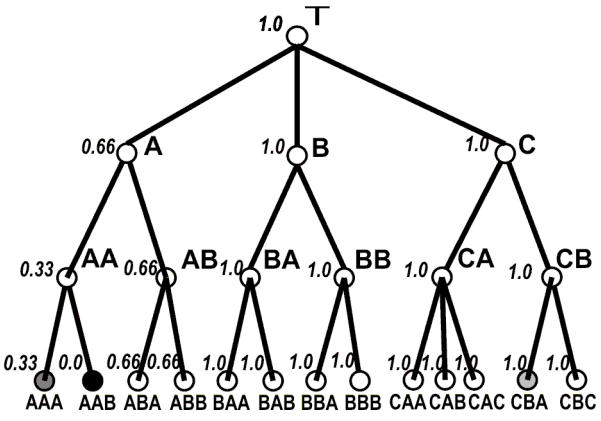

Figure 7.3: Vertical distances in $\mathcal{S}$ (normalized)

7.1.3 Cumulative Distance and Spatial Relevance

The relative spatial relevance of two units r_i and r_q of a spatial reference model $\mathcal{S}$ is a function of both their horizontal and their vertical distance as defined above. To account for this dual dependence, we integrate both parameters into a combined measure which we call the *cumulative distance* $\varepsilon(r_i, r_q)$ of r_i with respect to r_q.

The cumulative distance $\varepsilon(r_i, r_q)$ of two reference units r_i and r_q that are part of the same spatial reference model $\mathcal{S}$ is computed as a simple linear combination of their (normalized) horizontal and vertical distances, $\delta(r_i, r_q)$ and $\nu(r_i, r_q)$ [Schlieder et al., 2001]:

$$\varepsilon(r_i, r_q) = \alpha\delta(r_i, r_q) + (1 - \alpha)\nu(r_i, r_q) \tag{7.6}$$

In equation 7.6, the weighting factor α can take values between 0 and 1. By manipulating α, the cumulative distance can be biased either in favor of the horizontal distance (for $\alpha = 1$) or the vertical distance (for $\alpha = 0$). This means that a spatial query can be fine-tuned to favor either objects that are spatially close to the location of interest, or objects that are hierarchically related to the location of interest.

The spatial relevance $\sigma(r_i, r_q)$ of a unit r_i relative to a unit r_q is defined as the *inverse* of $\varepsilon(r_i, r_q)$. For the special case of a cumulative distance of 0, i.e. if r_q is equal to r_i, $\sigma(r_i, r_q)$ is set to unity:

$$\sigma(r_i, r_q) = \begin{cases} \varepsilon(r_i, r_q)^{-1} & : \quad r_i \neq r_q \\ 1 & : \quad r_i = r_q \end{cases} \tag{7.7}$$

7.1.4 Application Example: A Tessellation of US Counties

The following example will show the application of the concepts described above to a real-world application. The example uses a qualitative spatial reference model based on a tessellation of US counties and US states (Figure 7.4).

The respective spatial reference model $\mathcal{S}_{\mathrm{US}}$ consists of two connection graphs, C_{states}, encoding the tessellation of US-states, and $C_{counties}$, representing the tessellation of US-counties. The decomposition tree of this reference model has three levels: The top node (L_0) refers to the polygon representing the contiguous United States, while C_{states} and $C_{counties}$ refer to resolution levels L_1 and L_2, respectively. Represented in uncompressed ASCII XML format, the size of the qualitative reference model does not exceed $900KB$. This amounts to a data reduction by a factor of 3.7 in comparison to the US counties encoded in ESRI Shape format ($3140KB$).[3]

In this reference model, we focussed on a geographic area that is located in the central part of the United States, at the junction of the states of *Kansas*, *Missouri*, *Nebraska*, and *Oklahoma* (Figure 7.4).

As location of interest r_q, we chose a unit on the county-level of our spatial model. Starting in r_q, we computed a distance field at level $L = 2$ of the spatial model, i.e. in the connection graph representing the tessellation of counties. Following equations 7.1 and 7.2, we computed the normalized horizontal distance

[3]Note that the XML format used in this example has not been optimized for efficient data storage. Optimization of the representations scheme is likely to result in an even higher factor of data reduction.

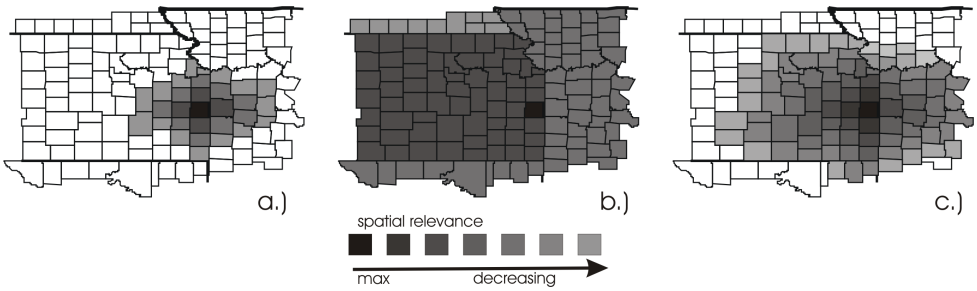

Figure 7.4: Relevance field for US counties for (a) $\alpha = 1$, (b) $\alpha = 0$, and (c) $\alpha = 0.5$

of each reference unit on level $L = 3$. The connection graph in our example is not weighted, i.e. any context-specific heterogeneities in the connectivity of individual counties are ignored. Accordingly, the maximum horizontal distance in the model is equal to the highest order of neighborhood that any node in the graph can have with respect to r_q. The vertical distances with respect to r_q are computed using equations 7.4 and 7.5. Because the hierarchy of $\mathcal{S}_{\text{US}}$ has only 3 levels, the maximum vertical distance with respect to r_q is 2.

Using equations 7.6 and 7.6, horizontal and vertical distances were combined to yield a relevance field on level $L = 2$ of $\mathcal{S}_{\text{US}}$. The shape of this relevance field depends on the weighting of horizontal versus vertical distance: For a weighting factor of $\alpha = 1$ (i.e. only horizontal distance counts), a quasi-circular region of counties of decreasing spatial relevance from r_q was computed (Figure 7.4). This region reflects the mere geographic neighborhood of the counties, without taking into account state boundaries or other hierarchical organization patterns.

For a weighting factor $\alpha = 0$ (i.e. only vertical distance counts), all counties in the state which contains r_q were assigned the same spatial relevance because they all belong to the same administrative super-unit (Figure 7.4-b). All counties in the neighboring states are uniformly assigned a lower spatial relevance.

Finally, for a weighting factor $\alpha = 0.5$ (i.e. both horizontal and vertical distance are counted equally), two semi-circles of decreasing spatial relevance were drawn around the location of interest (Figure 7.4-c). They are separated by the state boundary, with the counties in the neighboring state generally showing a lower spatial relevance than the counties in the "target" state.

7.2 Spatial Relevance in Place Name Structures

7.2.1 Horizontal and Vertical Distance for Place Name Regions

In the previous section we introduced a simple metric that can be used to compute the relative spatial relevance of two *reference units* in a qualitative spatial reference model. In this section, we will extend this metric to compute the relative spatial relevance of *place name regions* in a place name structure.

In a place name structure, a place name region is extensionally defined in terms of reference units of a qualitative spatial reference model. At the same time, each place name region has a defined position within the hierarchical

partonomy of the place name structure. Therefore the metric described in section 7.1 applies in principle to the computation of relative spatial relevancies between place name regions as well. However, there are two differences to consider:

1. The concept of *horizontal distance* refers to a semi-qualitative measure of metric distances in geographic space. The horizontal distance between two place name regions can therefore be computed directly as a function of the horizontal distances of their qualitative spatial footprints.

   However, the approximation of a place name region (i.e., the qualitative spatial footprint) may comprise multiple reference units. Consequently, we have to find a heuristic of how to integrate the relative spatial relevancies computed for these reference units into one meaningful relevance measure that represents the whole place name region.

2. *Vertical distance* is a measure of the distance of two geographic objects within a hierarchical partonomy and is closely related to the intrinsic semantics of the hierarchy. The vertical distance of two place name regions in a place name structure is therefore related to the semantics of the place name structure, not to the semantics of the underlying spatial reference model.

In the following we first describe how horizontal and vertical distances for place names that belong to the same place name structure can be computed. We then outline a method to integrate multiple place name structures, i.e. to compute the relative spatial relevance of two place names that belong to different place name structures. Using this method, we can specify a spatial query in one place name structure and evaluate the respective spatial relevance of place names in a second place name structure.

7.2.2 Horizontal Distance between Place Name Regions

In a place name structure $\mathcal{P}$, two place name regions p and q are represented by their qualitative spatial footprints S_p and S_q. If all elements S_p and S_q belong to the same qualitative spatial reference model $\mathcal{S}$, i.e. if $S_p, S_q \subseteq \mathcal{S}$, the horizontal distance $\delta(p, q)$ between the place name regions can be computed as a function of the horizontal distance between their spatial footprints. Formally we can say:

$$\forall S_p, S_q \subseteq \mathcal{S} \longrightarrow \delta(\mathrm{p}, \mathrm{q}) = \mathrm{f}(\delta(\mathrm{S_p}, \mathrm{S_q})) \tag{7.8}$$

Each qualitative spatial footprint consists of a finite set of elements, or reference units. In section 7.1 we have see that the horizontal distance between two reference units r_p and r_q can be computed on the basis of their neighborhood distance in the respective connection graph. To compute the horizontal distance of two sets of reference units (i.e. two spatial footprints), we integrate the horizontal distances computed for the individual elements of these sets. The horizontal distance $\delta(S_p, S_q)$ of two spatial footprints p and q can thus be computed as a function of all possible pairings $(r_p, r_q), r_p \in S_p \wedge r_q \in S_q$ given by the Cartesian product of S_p and S_q:

$$\delta(S_p, S_q) = f(\delta(S_p \times S_q)) \tag{7.9}$$

Normalization

The horizontal distance of two qualitative spatial footprints S_p and S_q can only be computed if both footprints are *normalized* to the same level of resolution L of the underlying spatial reference model $\mathcal{S}$. Normalization of a spatial footprint S is achieved by the iterative application of a *normalization function* η. The function η recursively decomposes all reference units $r \in S$ until $\forall r_j \in S | D(r_i) = D(r_j)$, where D denotes the depth of the decomposition tree $\mathcal{D}_\mathcal{S}$ of the reference model. In the resulting *normalized spatial footprint* Sn_p, all reference units $r \in Sn_p$ belong to the same connection graph $\mathcal{C}_\mathrm{L}$ at level L.

$$S_p = Sn_p \Longleftrightarrow \forall r | r \in S_p \wedge r \in \mathcal{C}_\mathrm{L} \tag{7.10}$$

To be able to normalize two qualitative spatial footprints S_p and S_q, two preconditions have to be fulfilled:

- All elements of S_p and S_q have to belong to the same qualitative spatial reference model $\mathcal{S}$, and
- there has to exist a common level of resolution L_N in $\mathcal{S}$ to which both S_p and S_q can be normalized. In general, L_N is given by the highest level of resolution that can be found among the reference units of the two spatial footprints.

$$L_N = L_r \Longleftrightarrow \forall r \in S_p, S_q \exists r_x | L_{r_x} \geq L_r \tag{7.11}$$

Distance Matrices

In the general case of a *complex spatial footprint* (see section 6.1), a place name region is extensionally defined through a tuple $S_p = (\bar{S}_p, \underline{S}_p)$, denoting a qualitative spatial footprint that consists of an upper and a lower approximation of the spatial extent of p (Figure 7.5).

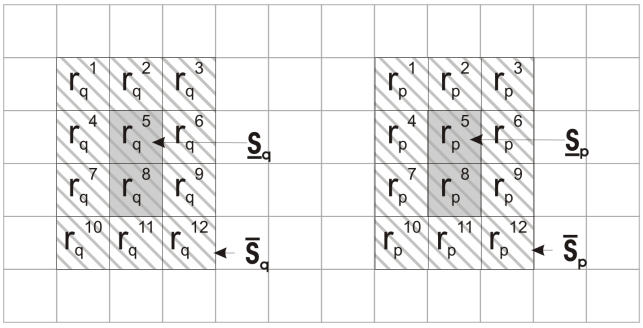

Figure 7.5: Generalized horizontal distance matrix for two complex spatial fooprints

For two place name regions p and q, the horizontal distance $\delta(p, q)$ is computed as a function of the horizontal distances between the pairings of upper and lower approximations given by the Cartesian product of $S_p = (\bar{S}_p, \underline{S}_p)$ and $S_q = (\bar{S}_q, \underline{S}_q)$:

$$\delta(p,q) = E(\delta(S_p \times S_q)) = E\begin{pmatrix}\delta(\bar{S}_p,\bar{S}_q) & \delta(\bar{S}_p,\underline{S}_q)\\ \delta(\underline{S}_p,\bar{S}_q) & \delta(\underline{S}_p,\underline{S}_q)\end{pmatrix} \tag{7.12}$$

The approximation of a region p comprises a finite number of reference units derived from a qualitative spatial reference model $\mathcal{S}$. Therefore the horizontal distance between two qualitative spatial footprints is a function of the horizontal distances between the individual elements of the spatial footprints. Given the upper approximations $\bar{S}_p$ and $\bar{S}_q$ of two regions p and q with n and m elements, respectively, the following distance matrix can be computed:

$$\delta(\bar{S}_p,\bar{S}_q) = F(\delta(\bar{S}_p \times \bar{S}_q)) = F\begin{pmatrix}\delta(r_p^1,r_q^1) & \cdots & \delta(r_p^1,r_q^m)\\ \cdots & \ddots & \cdots\\ \delta(r_p^n,r_q^1) & \cdots & \delta(r_p^n,r_q^m)\end{pmatrix} \tag{7.13}$$

For each field in matrix 7.12, the horizontal distances between all pairings of the elements of the respective approximations (i.e. the matrix 7.13) has to be computed. Function F integrates the resulting set of horizontal distances into one horizontal distance that is representative for the respective pairing of approximations. A straightforward heuristic for the integration function F is to compute the *arithmetic mean* of all element pairings. For a pairing of two upper approximations with n and m elements this would yield:

$$\delta(\bar{S}_p,\bar{S}_q) = F(\delta(\bar{S}_p \times \bar{S}_q)) = \frac{\sum_{i=1}^{n}\sum_{i=1}^{m}\delta(r_p^n,r_q^m)}{n+m} \tag{7.14}$$

Another feasible heuristic F^* is to select the *smallest* (i.e., closest) horizontal distance:

$$F^*(\delta(\bar{S}_p \times \bar{S}_q)) = min(\delta(\bar{S}_p \times \bar{S}_q)) \tag{7.15}$$

Analogous to F, functions E and E^*, respectively, may integrate the four matrix components given by the Cartesian product $S_p \times S_q$ of the spatial footprints of p and q by computing their *arithmetic mean*, or by selecting the *minimum* distance:

$$E(\delta(S_p \times S_q)) = \frac{\delta(\bar{S}_p,\bar{S}_q) + \delta(\bar{S}_p,\underline{S}_q) + \delta(\underline{S}_p,\bar{S}_q) + \delta(\underline{S}_p,\underline{S}_q)}{4} \tag{7.16}$$

and

$$E^*(\delta(S_p \times S_q)) = min(\delta(S_p \times S_q)) \tag{7.17}$$

7.2.3 Computing Vertical Distances in a Place Name Structure

Like the decomposition tree of a standard reference tessellation in a spatial reference model (see section 5.2), the hierarchical partonomy of a place name structure can be represented as a graph. And analogous to the partonomic, or vertical distance of two reference units of a spatial reference model, the vertical distance of two place names in a place name structure can be computed on

basis of this graph. However, because place name regions do not need to form tessellations there are two major differences that have to be considered:

1. In a place name structure it is possible for a place name region to be (both spatial and functional) part of multiple regions. The result is a directed acyclic graph (DAG) where one child node may have multiple parent nodes.

2. Unlike standard reference tessellations, place name structures may form irregular hierarchies. The resulting partonomic hierarchy may therefore be unbalanced, and the levels of the hierarchy may not necessarily reflect a conceptual classification.

Analogous to the computation of vertical distances in a spatial reference model presented in section 5.2, the relative vertical distance between two place names that belong to the same place name structure can be computed as a function of their graph-theoretic distance in the hierarchical DAG. For two place names p and q represented as nodes in a DAG, we compute the vertical distance of p with respect to q by starting in q and traversing the DAG upwards until the first common parent of both place names, P_{pq} is reached. The vertical distance $\nu(p, q)$ of p with respect to q is then defined as the shortest path SP between q and P_{pq}:

$$\nu(p, q) = SP(P_{pq}, q) \tag{7.18}$$

Following equation 7.18, the vertical distance $\nu(p, q)$ for all place names $p \in \mathcal{P}$ with respect to a place name of interest, $q \in \mathcal{P}$ can be computed (Figure 7.6).

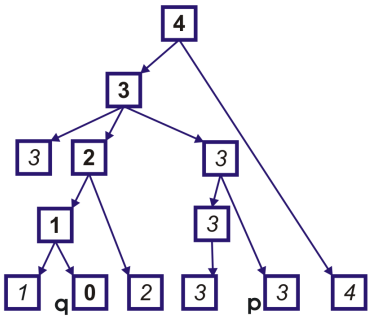

Figure 7.6: Vertical distances relative to place name p. Fat numbers denote common parents.

Multiple Inheritance

As we have seen above, the general procedure to compute the vertical distance of a place name p with respect to a location of interest p_q in a place name structure $\mathcal{P}$ is to start in p_q and recursively traverse the hierarchy upwards until the first common parent of p and p_q is reached. For place name structures that contain

nodes with multiple parents like the one depicted in Figure 7.7, the iterative sequence of this process is important. Figure 7.8 visualizes the computation process given the place name structure shown in Figure 7.7. In this example, place name j is the region of interest:

- Step 0: Node j is initialized with $\nu(j, j) = 0$, all other nodes are initialized with $\nu(x, j) = \infty$.
- Step 1: The parents of j, i.e. d and h, are assigned $\nu(d, j) = \nu(h, j) = 1$. All children x of j and d for which $\nu(x, j) = \infty$ (i.e. which have not been evaluated yet) inherit the vertical distance from their parents. Consequently we get $\nu(f, j) = \nu(i, j) = \nu(k, j) = 1$.
- Step 2: The parents of d and h, i.e. b and g are assigned a vertical distance of 2. Analogous to Step 1, all children of b and g that have so far not been evaluated, inherit a vertical distance of 2. This yields $\nu(c, j) = \nu(e, j) = \nu(l, j) = \nu(l, m) = 2$.
- Step 3: Because e has already been assigned a vertical distance (through inheritance from b), only a as the parent node of b is involved in this step. We assign $\nu(a, j) = 3$, and consequently $\nu(n, j) = \nu(a, j) = 3$.

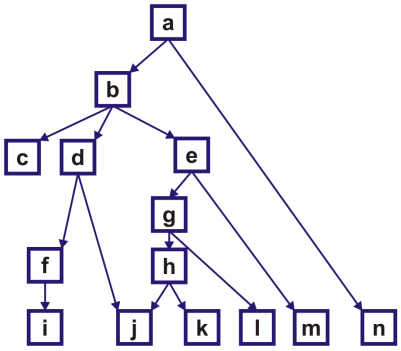

Figure 7.7: Example for an irregular PNS

If we follow this iterative procedure, we can compute a vertical distance with respect to j for each place name in $\mathcal{P}$. Note that the existence of shortcuts in the DAG, i.e. branches where hierarchy levels are missing, has a strong impact on the overall distribution of vertical distances.

To be able to compare them to the respective horizontal distances and to integrate them in a combined measure of cumulative distance, the absolute vertical distances in $\mathcal{P}$ (Figure 7.9a) have to be normalized with the maximum vertical distance in the place name structure (Figure 7.9b).

7.2.4 Spatial Relevance of Place Name Regions

Like the spatial relevance of two units of a qualitative spatial reference model, the relative spatial relevance of two place names in a place name structure is computed as a function of their horizontal and vertical distance. Analogous to the algorithm described in section 7.1, we use a linear combination of horizontal

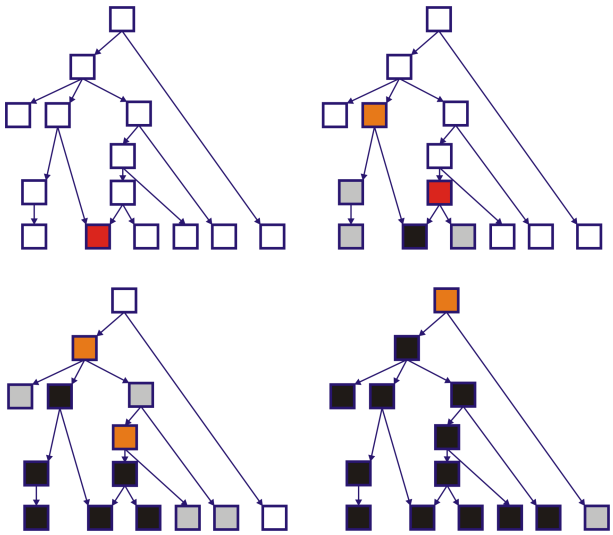

Figure 7.8: Iterative computation of vertical distances in an irregular PNS

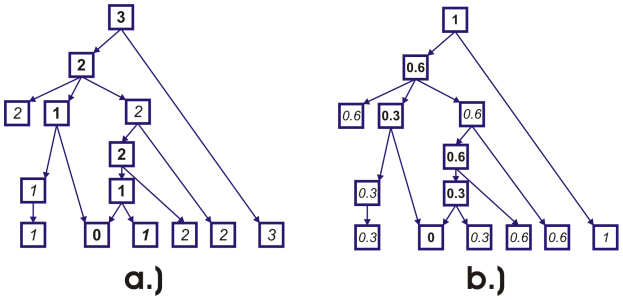

Figure 7.9: Absolute (a) and normalized (b) vertical distances

and vertical distance to compute the cumulative distance and the relative spatial relevance of two place name regions in a place name structure.

The relative cumulative distance of two place name regions $p, q \in \mathcal{P}$ is computed as a linear combination of the normalized horizontal distance $\delta(S_p, S_q)$ of their spatial extensions in a qualitative spatial reference model, and the normalized vertical distance of the two place names $\nu(p, q)$ within the partonomic hierarchy of $\mathcal{P}$.

$$\varepsilon(p, q) = \alpha\delta(S_p, S_q) + (1 - \alpha)\nu(p, q) \tag{7.19}$$

For two place names that do not have the same spatial extension, the relative spatial relevance $\sigma(p, q)$ is given by the inverse of their cumulative distance. Two place names that have an identical spatial extension and location in the $\mathcal{P}$ hierarchy (i.e., that are equal) are assigned a spatial relevance of 1.

$$\sigma(p, q) = \begin{cases} \varepsilon(p, q)^{-1} & : \quad \exists r | r \in S_p \wedge (r \notin S_q) \\ 1 & : \quad S_p = S_q \end{cases} \tag{7.20}$$

7.2.5 Integration of Multiple Place Name Structures

In the previous section we outlined our approach to compute the relative spatial relevance of two place names as a function of the horizontal and the vertical distances of the respective qualitative spatial footprints. So far, we assumed that both place names belong to the same place name structure. In the following, we extend our method to be able to compute the relative spatial relevance of two place names that belong to different place name structures. Our approach is based on two assumptions:

1. To be able to compute the relative spatial relevance of two place names $p \in \mathcal{P}$ and $p^* \in \mathcal{P}^*$, both place name structures $\mathcal{P}$ and $\mathcal{P}^*$ have to be based on the same frame of reference (i.e. the same qualitative spatial reference model $\mathcal{S}$).

2. Given that $p \in \mathcal{P}$ is the location of interest of a spatial query, the vertical distance of p^* with respect to p is computed exclusively on the basis of the partonomic hierarchy of $\mathcal{P}$, while the partonomic hierarchy of $\mathcal{P}^*$ is irrelevant.

The second assumption implies that it is always the hierarchy of the place name structure to which the location of interest belongs that determines vertical distances. To illustrate this implication, we can look at a query like: *"Find a natural region @ Lower Saxony"*. In this query, the place name *"Lower Saxony"* belongs to a place name structure $\mathcal{P}$ of administrative units of Germany. We are interested in finding natural regions that are spatially relevant to *"Lower Saxony"*, to a neighboring *Bundesland* (same level in $\mathcal{P}$), or to the *Landkreise* of *"Lower Saxony"* (one level down in $\mathcal{P}$). It is therefore its relation with respect to the hierarchy of administrative units encoded in $\mathcal{P}$ that determines whether a natural region is relevant to our query or not. The hierarchy of natural regions, i.e., the hierarchy of a place name structure $\mathcal{P}^*$ is irrelevant in this context.

Computing Horizontal Distances in Multiple PNS

In section 7.2.2 we described how the horizontal distance between two place names $p \in \mathcal{P}$ and $p^* \in \mathcal{P}^*$ can be computed. Essentially, for each element of the normalized spatial footprint representation of p, a horizontal distance field is computed within the respective connection graph in the spatial reference model. Based on the computation of neighborhood distances, the distance field assigns a horizontal distance value to each node of the connection graph. For a spatial footprint consisting of more than one element, this distance value is computed through an overlay of the respective distance fields.

Using the compounded distance field as a basis, the horizontal distances of a second place name region p^* with respect to p is computed as a function of its spatial footprint. As long as the place name structures $\mathcal{P}$ and $\mathcal{P}^*$ are based on the same spatial reference model, the computation of the horizontal distance follows the same procedure as the computation of the horizontal distance between two place names that belong to the same place name structure (section 7.2.1).

Computing Vertical Distances in Multiple PNS

Under the assumption that the vertical distance $\nu(p, p^*)$ between two place names $p \in \mathcal{P}$ and $p^* \in \mathcal{P}^*$ depends only on the hierarchical structure of $\mathcal{P}$ (see above), $\nu(p, p^*)$ can be computed using a procedure similar to the one used to compute $\delta(p, p^*)$:

For each pair of place names $p, p_q \in \mathcal{P}$, the vertical distance $\nu(p, p_q)$ can be computed as shown in section 7.2.3. Following this procedure, each node in $\mathcal{P}$ is assigned a vertical distance with relative to p_q. To compute the vertical distance $\nu(p^*, p_q)$ of a place name in another place name structure $\mathcal{P}^*$, we project the vertical distances $\nu(p_i, p_q)$ of the place names $p_i \in \mathcal{P}$ onto the reference units of the qualitative spatial reference model underlying both $\mathcal{P}$ and $\mathcal{P}^*$. The vertical distances are projected onto the connection graph $\mathcal{C}_\mathrm{L}$ at the level L of $\mathcal{S}$ that corresponds to the *normalized spatial footprint* Sn_i of p_i. In case of a complex footprint, the normalized upper approximation $\bar{S}_i$ is used [4]. Each node in $\mathcal{C}_\mathrm{L}$ that belongs to Sn_i is assigned the vertical distance $\nu(p_i, p_q)$ of the corresponding place name. If a node in $\mathcal{C}_\mathrm{L}$ belongs to more than one spatial footprint it may be assigned multiple vertical distances. In this case, the minimum vertical distance assigned is chosen as the representative value for this node. The result is what can be called the *"vertical distance"* field $\mathcal{VD}_\mathrm{q}$ of p_q (Figure 7.10).

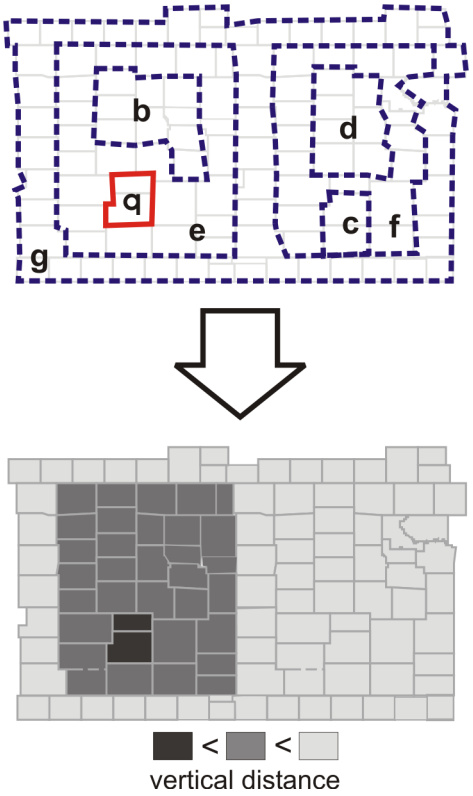

Figure 7.10: Vertical distance field $\mathcal{VD}_\mathrm{q}$ for region q

[4] Note that the vertical distances for the upper and the lower approximation of a place name are identical.

The vertical distance $\nu(p^*, p_q)$ for a place name $p^* \in \mathcal{P}^*$ is computed on basis of the vertical distance field of p_q. It is a function of the vertical distances assigned to the elements of the place name's spatial footprint normalized to $\mathcal{C}_{\mathrm{L}}$. Analogous to the computation of the horizontal distance described above, a straightforward heuristic is to compute $\nu(p^*, p_q)$ as the *arithmetic mean* of the vertical distances assigned to the elements of S_{p^*}.

Because the upper and the lower approximation of p^* may cover a different area in the vertical distance field, we have to make a distinction here between the vertical distance of $\bar{S}_{p^*}$ and the vertical distance for $\underline{S}_{p^*}$. We compute both values separately and use their arithmetic mean as representative for p^*.

A frequent case in practical applications is that the geographic regions covered by two place name structures only partially overlap. For a place name p_x that is completely outside the coverage of $\mathcal{P}$ (i.e., does not overlap any place name regions in $\mathcal{P}$), we use a generic vertical distance of $\nu(p_x, p_q) = MAX(\nu(p_i, p_q) + 1)$[5] This signifies that, independent of which location of interest p_q, the vertical distance of a place name p^* that is outside the area of coverage of $\mathcal{P}$ will always be larger than the vertical distance of any other place name in $\mathcal{P}$.

[5]To define the vertical distance of reference units outside the coverage of $\mathcal{P}$, the DAG representing the partonomic hierarchy of $\mathcal{P}$ is extended by a new top node p_{T^*}. One child of p_{T^*} is the former top-node p_T. The other children are all place names with a spatial extension outside of the region cover ed by $\mathcal{P}$.

Part III

Prototypical Implementation and Evaluation

Chapter 8

Evaluation and Applications

8.1 Retrieval Performance Evaluation

8.1.1 Objective and Methodology

The objective of this analysis is to evaluate the capabilities of the approach outlined in the previous chapters when applied in a real-world scenario. We are particularly interested in answering the question whether our approach can improve the results of the spatial part of an information request of type *concept@location*[1].

In the literature, a number of methods to test and evaluate search algorithms can be found. In this work, we apply some of the standard methods outlined in [Baeza-Yates and Ribero-Neto, 1999]. Generally speaking, we are interested in evaluating the *retrieval performance* of our search algorithm. Retrieval performance relates to the question how precise the answer given by the search algorithm is, i.e., how well the set of retrieved information objects matches the expectations of an information seeker.

The evaluation of retrieval performance involves a reference collection of documents, a set of test information requests, and a set of relevant result sets, one for each information request. The search algorithm, or *retrieval strategy* S is tested against these reference data. Using a number of measures described below, we try to evaluate the similarity between the documents retrieved and the reference set. Following this procedure, we try to make a judgement about the *goodness* of the retrieval strategy.

8.1.2 Evaluation Measures

From the methods for testing the retrieval performance of a search algorithm described in [Baeza-Yates and Ribero-Neto, 1999] we focus on measures for the evaluation of the *precision* and the *recall* of a search algorithm. In a collection that contains a set C of information objects, the set of all relevant information objects with respect to an information request I is called $R_I : R_I \subseteq C$. The set of information objects that are actually retrieved using I is called the *result*

[1]Note that this analysis is focused on the spatial part of the information request. The conceptual part plays a role only when it comes to the context-driven optimization of the spatial query. Likewise, temporal aspects are ignored completely.

set A. Those information objects that are both retrieved by I and relevant to I form a set R_a called the *relevant result set*. The set R_a is given by the intersection of R and A.

Typically, only a fraction of relevant information objects is retrieved. The fraction of the number of all available relevant information objects, $|R_I|$, and the number of relevant information objects $|R_a|$ that appear in the result set A is called the *recall RC* of an information request:

$$RC = |R_a|/|R_I| \tag{8.1}$$

An information retrieval algorithm works optimal if the answer set A contains only relevant information objects. The fraction of retrieved objects that are relevant is called the *precision PR* of an information request:

$$PR = |R_a|/|A| \tag{8.2}$$

Baeza-Yates and Ribero-Neto suggest to measure precision and recall with respect to *ranked* answer sets. The ranking reflects the relevance of an information object as computed by the information retrieval algorithm. Using this approach, precision-recall curves can be computed that reflect the quality of the retrieval algorithm.

A pre-requisite to compute recall and precision measures is knowledge about R_I, i.e. the set of all documents in a collection C that are relevant with respect to an information request I. In very large collections like the WWW it is virtually impossible define R_I properly. Here, a relative value for recall can be calculated by estimating the total number of relevant sites. If different search algorithms are to be compared, one way is to use the highest number of relevant sites found for any of the search algorithms. Another option is to sum up the total number of unique relevant sites that are returned by each of the of the compared search algorithms [Killmer and Koppel, 2002].

Another option to determine R_I is to have a human expert select all information objects that he/she considers as relevant with respect to a specific information request. Obviously, the selection is highly subjective and if the expert does not know all information objects in the collection, R_I is likely to be only a sub-set of the set of all relevant objects. On the other hand, the reference set identified in following this approach contains all information objects the expert (i.e., a human user of the search algorithm) wants to find. Therefore it can be regarded as a good approximation of the complete set of relevant objects.

A number of derived measures and graphical depictions based on recall and precision calculations can be applied to evaluate different aspects of the retrieval performance of a search algorithm:

Precsion-recall curves: For each individual information request a precision-recall curve can be calculated. This curve is derived from a comparison of the ranked answer set A with the reference result set R_I. For example, let us consider a reference result set $R_I = \{O_2, O_5, O_6, O_{45}, O_{48}\}$ and a ranked answer set of $A = \{(1; O_{48}),(2; O_1),(3; O_2),\ (4; O_4),(5; O_{20}),(6; O_{45}),(7; O_{25}),(8; O_5),(9; O_{13}),(10; O_6)\}$, where each tuple $(N; O)$ represents the ranking and the ID of the respective information item. In this case, the precision-recall curve would be calculated as follows:

- The first element of A is O_{48}, which is contained in R_I and therefore considered as relevant. Because O_{48} corresponds to 20% of R_I, we say that there is a precision of 100% at 20% recall, i.e. $PR = 100$ and $RC = 20$.
- The next relevant object in A is O_2 at the third position. Therefore, $PR = 66$ (two out of three objects are relevant) and $RC = 40$ (two of the five relevant objects were examined).

Plotting this example yields the precision-recall curve depicted in Figure 8.1.

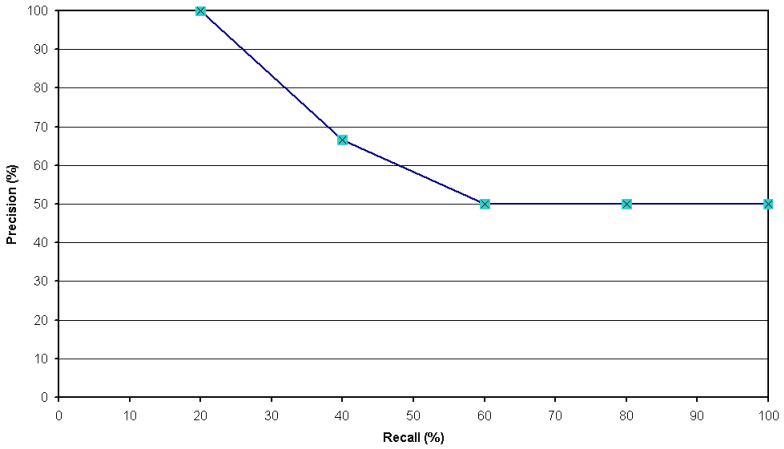

Figure 8.1: Precision-recall curve for example query

Average precision at seen relevant documents: If we want to express the performance of a search algorithm for an individual query as a single value, we can use the average precision at seen relevant documents, $PR_{\varnothing}$. We obtain this measure by averaging the precisions computed for every new relevant information object found in the result set.

R-precison: Like $PR_{\varnothing}$, R-Precision PR_R is a single value parameter to evaluate an individual query. PR_R is given by the precision computed for the information object at the R^{th} position in the ranked result set. R in this context refers to the number of elements in the respective reference set. Given a reference set of $R = 10$ elements, PR_R is $3/10 = 0.3$ if 3 relevant elements are found within the first 10 objects in the result set.

Precision histograms: If we want to compare two algorithms (or query parametrizations A and B) for different search terms and reference sets, a useful tool is the *precision histogram*. For this purpose, we plot the difference $\Delta PR_R = PR_R(A) - PR_R(B)$ of the R-precision values computed for two individual search algorithms with respect to a specific query. If the difference is 0, both algorithms perform equally well. If the difference greater than 0, algorithm A performs better. If it is smaller than 0,

algorithm B shows a better performance for the query in question. A histogram plot of ΔPR_R for multiple queries can provide information about the strengths and weaknesses of the two algorithms compared.

8.1.3 The Test Reference Collection

The UDK Data Set

Because we want to evaluate the efficiency of a *spatial* retrieval algorithm, we need a reference collection that contains information objects that are explicitly geo-referenced, preferably by place names. One information collection that fulfills these requirements is the *Umweltdatenkatalog (UDK)* [Swoboda et al., 2000]. The *UDK* is a metadata information system that is used in Germany and Austria by the federal states and the respective federal environmental agencies for the management of environmental data (see also section 3.2). In the *UDK*, each information object is given a geo-reference through one or multiple place names that correspond to administrative units as defined in the respective *NUTS* classification scheme.

In Germany, the *UDK* integrates the data sets maintained by the Environmental Ministry in each of the 16 federal state, plus a database maintained by the *Umweltbundesamt (UBA)* in Berlin. For our evaluation, we integrated the data holdings of four federal states, namely *Lower Saxony* (Niedersachsen), *Northrhine-Westfalia* (Nordrhein-Westfalen), *Thuringia* (Thüringen), and *Hesse* (Hessen). The integrated database consists of a total of 12639 data objects.

Spatial reference model and place name structure

For the purpose of this evaluation, we concentrated on a region at the intersection of the four German federal states included in the data set (Figure 8.2).

We used a polygonal reference tessellation of administrative units in Germany as the basis for our spatial reference model. The maximum resolution level of this model was the *NUTS 3* level. Based on this spatial model, we built a place name structure that models the *Landkreise*, *Regierungsbezirke* (where applicable), and *Bundesländer* of the four federal states mentioned above.

Conceptual reference model

All information objects in the UDK data collection are annotated with one or multiple thematic *keywords*. The common vocabulary used to assign these keywords is the UDK thesaurus, an index of mainly environmental terms integrated in the UDK database.

Determining reference sets

As mentioned above, to determine which information objects are relevant to a specific query is not trivial and depends to a certain extent on the subjective judgement of the information seeker. To determine the reference sets used in this analysis, a human user was assigned the task to select manually (i.e., using the standard SQL requests on the *UDK* database) from the *UDK* data collection those information objects that seemed to be the most relevant with respect to a

Figure 8.2: Region of interest for evaluation study

specific *concept @ location* query. The idea behind this approach was to simulate a user who, without the help of an information retrieval tool but with access to all available information objects, browses through the database to find what he/she identifies as relevant.

The resulting set of information objects was ranked based on their spatial relevance with respect to the place name specified in the *location* part of the query.

Spatio-Conceptual queries

The test information requests were we performed against the a *UDK* data collection stored in a relational (MS-Access) database. The query procedure had two steps:

Step I - Thematic query: One or multiple thematic keywords were specified to pre-select all information objects that refer to the thematic coverage in question. If multiple keywords were used, they were connected by the logic *OR* operator. In some cases, wildcards were used to extend the query and to ge a more substantial result set.

Step II - Spatial query: The spatial query was applied to the result set A_C of step I. The spatial query consisted of a sequence of place names connected by the logic OR operator. The sequence of place names was sorted based on the relative spatial relevance of the individual place names, i.e. the first element of the sequence was the place name with the highest relevance,

the last element the place name with the lowest relevance. The result of step II was a result set $A_{CS} \subseteq A_C$.

Evaluation of query results

The result set A_{CS} obtained after step II of the information request was evaluated against the respective set of relevant reference objects. For each query, the evaluation was performed for both an unsorted and a sorted result set. To obtain a sorted result set, each element of the unsorted result set (i.e., A_{CS} after step II of the query) was assigned a spatial relevance in accordance to the spatial relevance of the place name it was annotated with. The result set was then sorted based on spatial relevance.

Each query was performed with a number of different parametrizations of the search algorithms. This refers mainly to a variation of the parameter α which determines the bias of the spatial query on horizontal ($\alpha = 1.0$) or vertical ($\alpha = 0.0$) distance.

8.1.4 Evaluation Runs and Results

To evalutate the retrieval performance of our search algorithm we defined four queries (Table 8.1.4). The concept part of each query consists of one or multiple keywords (in German) that were taken from a thesaurus of environmental terms integrated in the *UDK* to ensure consistency with the *UDK* metadata. For the location part of the query, a place name denoting a specific *"Landkreis"* was chosen. Geographically, the locations of interest were chosen to be within the region shown in the rectangle in Figure 8.2.

ID	*Concept*	*Location*
UDK-E1	"Landschaftsveränderung"	"Holzminden"
UDK-E2	"Wasserrahmenrichtlinie"	"Lippe"
UDK-E3	"Abfallwirtschaft*"	"Lippe"
UDK-E4	"Wasserschutz*"	"Göttingen"

For each spatio-conceptual query, a set of relevant documents was defined manually. The *UDK* database was then queried with the information requests listed in Table 8.1.4 using the 2-step procedure described above. The final result set was evaluated against the respective reference set to determine a number of evaluation measures, including recall R, the total number of retrieved information objects $|A|$, precision PR, average precision PR_{AV}, and R-precision PR_R. The results of the analysis are described below.

Total number of retrieved information objects

With the help of the metric described in section 7.1, we define a ranking of all place names in a place name structure according to their spatial relevance with respect to a location of interest. By including the first n place names in this ranked list in a spatial query, we can expand the query and create what could be called a *"qualitative buffer"* around the location of interest. It can be expected

that the more place names are included in this buffer, the more information objects of a collection will be retrieved.

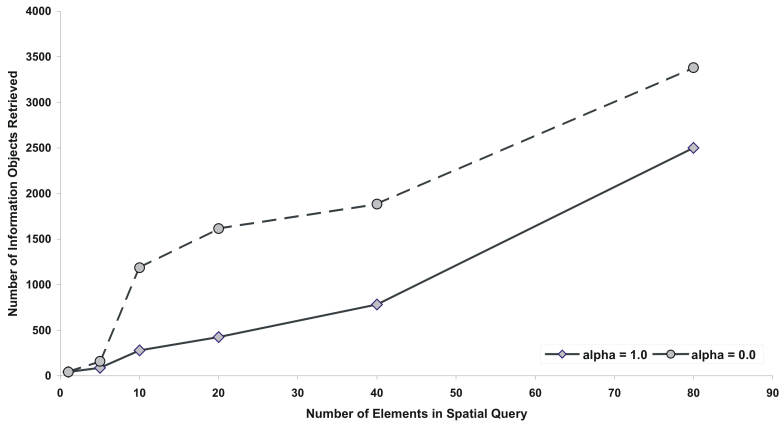

Figure 8.3: Retrieval of information objects as a function of query size nQ

We tested the impact of query size nQ on the number of retrieved information objects on the complete *UDK* reference collection with it's 12639 records . As expected, the number total number of retrieved information objects $|A|$ inceased with the number nQ of place names allowed in the qualitative buffer. (Figure 8.3)[2].

Another interesting result was the dependency of the query results on the choice of the factor α. The selection of α determines the ranking of the place names in the expanded query: For $\alpha = 1.0$, place names that are geographically closer to a location of interest p_q are ranked higher. For $\alpha = 0.0$, place names that represent super-regions of p_q are preferred. We generally say that in a spatial query, $\alpha = 1.0$ favors *horizontal distance*, while $\alpha = 0.0$ puts more weight on *vertical distance*.

However, the fact that the query retrieves more objects when vertical distance is stressed does not indicate that information objects in the *UDK* data set are generally annotated with place names that refer to a higher level of the hierarchy. In fact, only 26% of the records are annotated with direct references to a *"Regierungsbezirk"* or a *"Bundesland"*, while 74% of the data are annotated with a reference to a *"(Land)kreis"*. It rather shows that when we use a bias towards vertical distance, a smaller query yields more results faster because more information objects are annotated with a reference to the *same "Regierungsbezirk"* or *"Bundesland"*. The fast rise in retrieved information objects for queries that have between 5 and 10 place names support this observation.

With respect to the general behavior of our search algorithm this indicates that the *resolution* of a spatial query is higher, i.e., the relevance ranking is more differentiated and the clusters of equally-weighted information objects are smaller when we look only at horizontal distances.

[2]Note that the maximum of $nQ = 80$ is dictated by the SQL database we used, which does not allow for larger queries.

The impact of the factor α on precision

We have seen above that the choice of the factor α has a strong impact on the number of information objects that are retrieved through a spatial query. But does it also affect the *precision* of the query? To answer this question, we evaluated the results of the four test queries described above. The spatial part of each query is performed against the result set of the conceptual part of the query. The final result is compared to the query-specific set R_a of relevant information objects.

We ran each evaluation query using two parametrizations A and B of the search algorithm: In A we set α to 1.0, in B, α was set to 0.0. We determined the respective *R-precision* PR_R for each parametrization and computed the the R-precision difference ΔPR_R of $PR_R(A)$ and $PR_R(B)$ for each of the four evaluation queries.

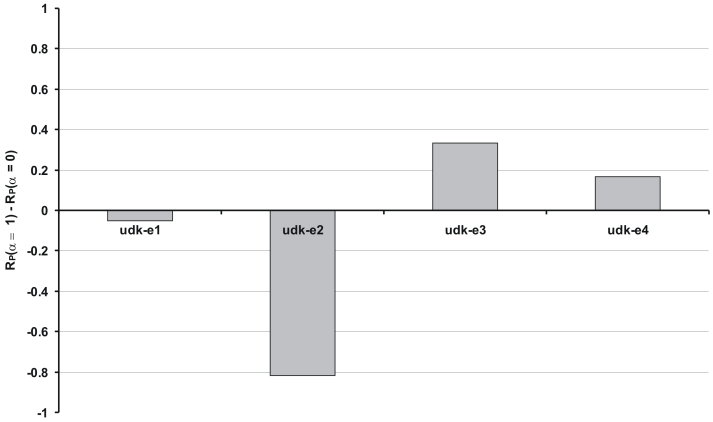

Figure 8.4: R-Precision differences for $\alpha = 1$ and $\alpha = 0$

The precision histogram given in Figure 8.4 shows that there are significant differences for the individual queries: While $\alpha = 0$ works much better for $udk - e2$, the evaluation queries $udk - e3$ and $udk - e3$ show a better performance for $\alpha = 1$. The query $udk - e1$ seems to show no explicit preference with respect to which α was chosen.

This result supports our assumption that the parametrization of the spatial part of an information request has to take into account the thematic context of the whole query. Given a spatial reference model of administrative subdivisions (as used both in our evaluation runs and to organize the *UDK* data), information objects concerned with concepts that have a strong relation to the administrative hierarchy (e.g., *"Wasserrahmenrichtlinie"* in *udk-e2*) are more likely to be retrieved by a vertically-biased query than objects annotated with concepts that have a more spatial connotation (e.g., *"Abfallwirtschaft"* and *"Wasserschutz"* in *udk-e3* and *udk-e4*).

However, not for all concepts there is a significant impact of search results due to which factor α was chosen (e.g., *"Landschaftsveränderung"* in *udk-e1*. In addition, we should always keep in mind that whether an information object is

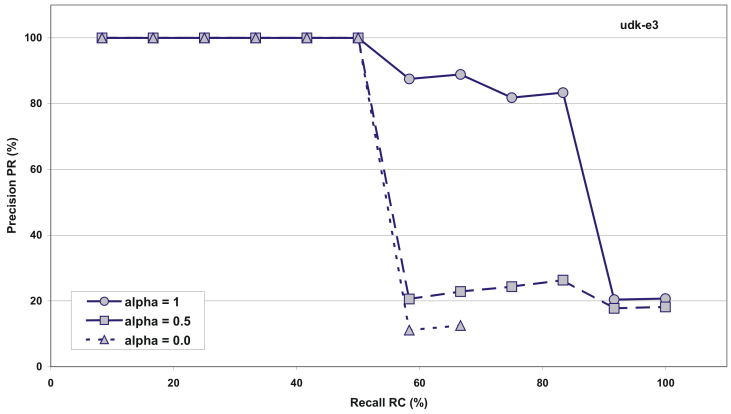

Figure 8.5: Recall-precision curve for *udk-e3* as a function of α

annotated with a place name from a higher or from a lower level of a place name hierarchy depends very much on the annotation guidelines and organizational paradigms implemented in the organization responsible for the information collection. Therefore we cannot develop general rules of which α to choose for specific concepts, but rather advise the user of our search algorithm to find out the best parametrization on a case-to-case basis.

The impact of the factor α on recall

The analysis described above showed that the choice of the factor α has a strong impact on the precision of a spatial query. To evaluate the impact of α on the recall of a query, we computed precision and recall for three parametrizations A, B, and C of *udk-e3*. In A, α was set to 1.0, in B to 0.5, and in C to 0.0. The resulting recall-precision curve is shown in Figure 8.5.

The figure shows that the recall for both $\alpha = 1.0$ and $\alpha = 0.5$ is equal, namely 100%. For $\alpha = 0.0$, the recall drops to only 65%. Precision, on the other hand, is at an equal 100% for all three parametrizations up to a recall of 50%. Then shows a slow decline for $\alpha = 1.0$ and a sharp drop for $\alpha = 0.5$ and 0.0.

In a second analysis, we plotted the precision-recall curves for *udk-e2* and *udk-e3* for $\alpha = 1.0$ and $\alpha = 0.0$, respectively (Figure 8.6). The figure shows that the choice of parameter α has a strong impact on both the precision and the recall of a spatial query.

Improved precision through sorting

In the analysis described above we used ranked, or *sorted* result sets. We sort a result set by ranking its records according to their individual relevance weights. Each record is assigned a relevance weight that corresponds to the spatial relevance computed for the place name the respective information object is annotated with.

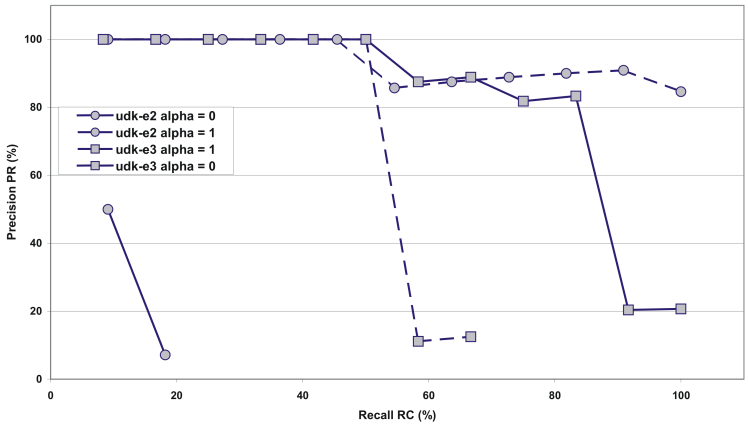

Figure 8.6: Recall-precision curves for *udk-e2* and *udk-e3* as a function of α

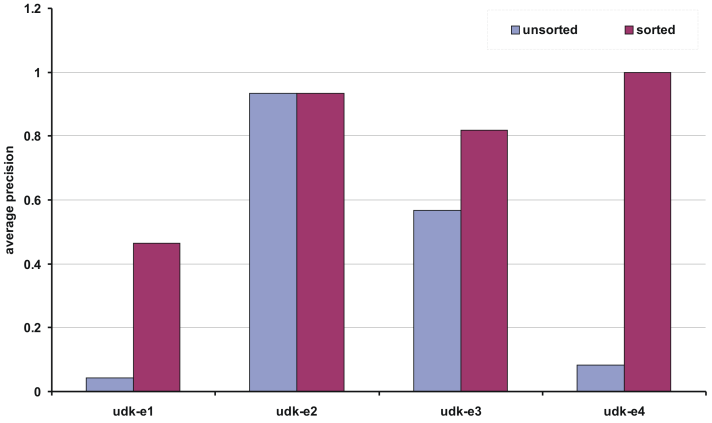

Figure 8.7: Average precision in unsorted and sorted result sets

In a sorted result set, the records with the highest rankings (i.e., the information objects that are spatially most relevant) are on top. Obviously, this should induce an improved precision of the query because more elements of the reference set should be found faster.

In Figure 8.7, we plotted the *average precision* for each of the four evaluation queries for an unsorted and a sorted result set. We chose the best average precision that could be obtained as a function of α, i.e. α was set to 0.0 for $udk - e2$, and to 1.0 for $udk - e1$, $udk - e3$, and $udk - e4$.

We observe that the average precision for the sorted result sets is significantly higher than for the unsorted result sets. As the unsorted result sets correspond to result sets that would be retrieved for instance through a standard database query, we can claim that our approach of sorting the search results according to spatial relevance significantly improves the precision of a query. The only exception is *udk-e2*, where both the unsorted and the sorted result set show the

same precision. This behavior is related to the fact that for vertically biased queries ($\alpha = 0.0$) the resolution of the spatial query, i.e. the number of relevance clusters, is small. The effect of sorting is therefore reduced.

8.2 A Tool for Spatial Information Retrieval

8.2.1 The BUSTER Information Broker

In the previous sections, we presented an approach for the representation of place name regions on the basis of discrete, qualitative spatial reference models. Such representations can be used to compute the relative spatial relevance of place name regions. Using this approach, information objects that are spatially referenced by place names can be ranked on basis of their relative relevance with respect to a spatial query. In information retrieval tasks, such a ranking can be used as a relevance measure for the spatial part of a complex information request.

A spatial reasoning module that implements some of the methods developed in this work has been implemented at the University of Bremen as part of an application called the *Bremen University Semantic Translator for Enhanced Retrieval (BUSTER)* [Schlieder et al., 2001, Neumann et al., 2001, Visser et al., 2002, Vögele et al., 2003a]. BUSTER is a prototypical information broker middleware that was built to tackle, among others, the challenges of *information discovery* and semantic *information integration* in heterogeneous and distributed information collections like the Semantic Web or service-based spatial data infrastructures.

Information integration refers to the second step of information retrieval in which heterogeneous information objects are integrated into one comprehensive view. In BUSTER, information integration is achieved through the resolution of the respective structural, syntactical, and semantic heterogeneities. To achieve semantic translation, BUSTER applies an approach that is based on the use of formal ontologies in the sense of [Grüninger and Uschold, 2002].

The focus of this work is information discovery, i.e. the ability to point an information seeker to those information items that promise to be the best match with respect to a specific task or information request. We will therefore concentrate on the information discovery capabilities of the BUSTER system.

8.2.2 Information Discovery with BUSTER

In tasks that involve information discovery, the BUSTER system is able to evaluate and rank information objects based on their conceptual and spatial relevance with respect to specific information requests. In the latest version of BUSTER, methods for the evaluation of temporal relevance were added. Thus the BUSTER system now supports queries of the type *concept @ location* ***in*** *time*.

A typical example for such an information request would be: *"Find a* ***conference-hotel*** *near the* ***Thüringer Wald*** *region for the period* ***between Christmas and New Year 2002****"*. In the concept part of this query, the term *"conference-hotel"* defines a specific class of accommodation, the place name *"Thüringer Wald"* is used to describe the location of interest, and the

term *"between Christmas and New Year 2002"* signifies the time period of interest. In the following, we will briefly discuss the methodology used to assess the conceptual and temporal relevance of information items, and then discuss the implementation of the reasoner for spatial relevance in more detail.

The Conceptual Search Dimension

Within the last last decade, a number of approaches to tackle the problem of intelligent information retrieval and semantic data integration have been proposed (e.g. Ontobroker [Decker et al., 1999] [Fensel et al., 1998] or PICSEL [Goasdoué et al., 1999]). Most of these systems use formal ontologies to represent the semantics of information sources. In heterogeneous and distributed environments (e.g., the Semantic Web) a major challenge is to provide the means for ontology integration. Currently there are three different types of approaches to this problem, namely the so-called *single ontology* approaches, *multiple ontology* approaches, and *hybrid* approaches (see [Wache et al., 2001] for an overview).

In BUSTER, a hybrid ontology approach is applied. A distinction is made between *application ontologies* and *domain ontologies*. The semantics of an information item is formalized in an application ontology which is attached to the information object. As indicated by it's name, an application ontology reflects the semantics of concepts that are specific to a given application and is therefore limited in scope. Different application ontologies become comparable if they were built on basis of the same *shared vocabulary*. This shared vocabulary is defined in the domain ontology, the scope of which is an entire application domain.

The shared vocabulary consists of basic terms that are described in terms of *properties* and *values*: Properties are used to describe conditions for concept membership, while a common set of values specifies value ranges for the properties. In order to gain expressiveness, an additional comparison predicate is used, so that a property p_i^X of an entity X is defined by:

$$p_i^X \equiv property(X, Value) \wedge comparison(Value, ValueRange) \tag{8.3}$$

Typical comparison relations are *equality*, *range*, *order*, and *type*. A local (application) ontology can be viewed as a (partial) refinement of the global (domain) ontology, i.e. it restricts the value range of some attributes. Thereby all local ontologies remain comparable because they are built upon the vocabulary defined by the global ontology.

Using this approach, two or more application ontologies can be integrated. For example, we can select a concept $X \in A$ that has the properties *has-conference-rooms 3* and *has-specials bowling-alley* (Figure 8.8). We can infer that X subsumes $Y \in B$ because Y has more than 3 *conference rooms* and a *9-pin-bowling-alley*, which according to the shared vocabulary is a sub-concept of *bowling-alley*.

In the current BUSTER prototype, both application and domain ontologies are encoded in *DAML+OIL* [Berners-Lee et al., 2001a], an ontology interchange language based on description logics. The logic theorem provers *RACER* [Haarslev and Möller, 2001] and *FaCT* [Horrocks, 1998] are applied to classify concepts.

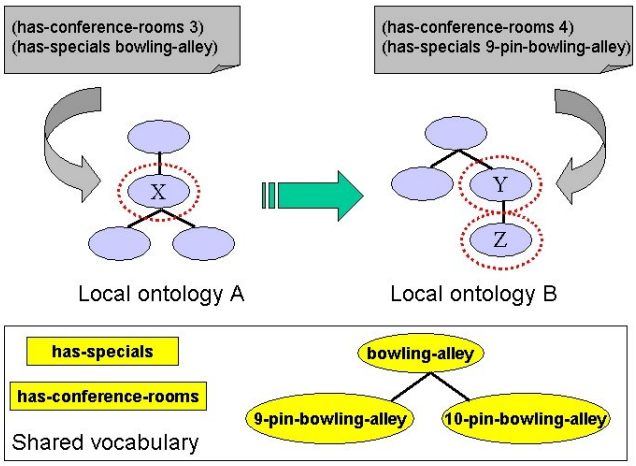

Figure 8.8: Integration of two (local) application ontologies

The Temporal Search Dimension

For most data and information objects, temporal relations are an important part of the respective metadata description. Typical examples are the creation and modification dates of an information object. Meaningful temporal queries, however, should not only be concerned with points in time, but also with *time intervals*. A typical temporal query might be the question whether a vacation home is available for a specific time period, say *"between July 1st and 22nd, 2003"*. Other queries may ask for an event that happened *"yesterday"* (denoting a time interval relative to another point/interval), or for a new law in effect *"since January 1st 2003"* (i.e. a time interval with a fixed starting point and an indefinite ending). In BUSTER, this time interval is called the *temporal relevance interval* of an information object.

The Dublin Core Metadata Standard (DC) [DCMI, 2000b] is the basis for the internal metadata representation in the BUSTER prototype. In the DC, two elements are available to annotate temporal metadata [DCMI, 2000a]: The first one is *Date* with its qualifiers *Created, Valid, Available, Issued*, and *Modified*. *Date* is more related to the *"technical"* aspects of temporal annotation and is typically used to specify the instantiation or version of a resource. Another (qualified) element is often better suited to match the demands of the Semantic Web: *Coverage*. *Temporal coverage* represents the *"temporal characteristics of the intellectual content of the resource"* ([DCMI, 2000a]). This is a fairly vague definition, but obviously the element can be used to contain the temporal relevance interval mentioned above.

The Dublin Core defines two encoding schemes to be used with *Date* and *Coverage*, namely the *W3C Date Time Format* [Wolf and Wicksteed, 1998] and the *DCMI Period format* [DCMI, 2000a]. While these formats are well suited to define exact time points or time intervals with crisp boundaries, they lack the ability to refer to expressions like *"this year's summer vacation"* or *"during the middle ages"*. However, such vague qualitative expressions are more intuitive

and therefore often more useful in information retrieval than exact representations of time. To represent vague temporal concepts, the notion of *period names* was introduced by [Hübner, 2003]. Period names are a qualitative abstraction of temporal intervals in the form of intuitive name descriptors like *"The Middle Ages"* or *"Easter 2003"*.

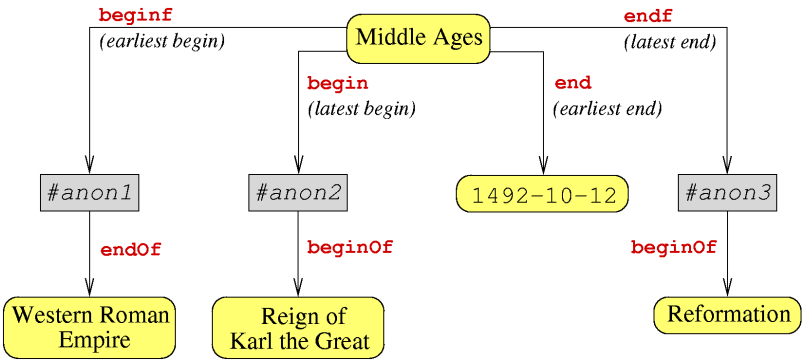

Figure 8.9: Example of a period name definition (from [Vögele et al., 2003a])

Vague temporal concepts represented by period names are used in conjunction with a set of purely qualitative temporal relations, like *"younger"*, *"older"*, *"survives"*, *"survivedBy"*, *"contemporary"* (as introduced by [Freksa, 1992a] 1992)). An algorithm based on Allen's time logic [Allen, 1983] and the concept of coarse knowledge introduced by [Freksa, 1992a] is used to reason with period names. This algorithm allows to draw inferences based on incomplete temporal information by referring to implicit knowledge.

For example, it is possible to identify that a time period T_A is older than period T_B, even if the start of T_A is unknown or persistent, or if its end point is earlier than, or equal to, that of T_B. By creating a graph of all valid and invalid relations between the known period names, inconsistencies are identified. This includes contradictions between a qualitative statement and a quantitative fact, between two or more qualitative statements, and between a set of statements and facts. While testing the temporal data for consistency, the graph of relations is extended to its maximum, so all possible new statements about the known periods are gathered and new knowledge is created.

Analogous to the conceptual query, the user is supported by the BUSTER GUI in specifying the temporal part of a query. From the set of *temporal models* that are registered with the system the user may select a time period of interest. BUSTER expands this time period into a sorted list of all relevant time periods that are part of the registered temporal models.

The Spatial Search Dimension

The spatial reasoning module integrated in the BUSTER prototype implements the approach for the representation and relevance reasoning with place names described in the chapters 5 to 7. Place names in BUSTER are organized in place name structures (PNS), where each PNS defines a unique name space. Within this name space, place names are specified through

- a name (does not have to be unique),

- a unique identifier,
- an extension in terms of units of a qualitative spatial reference model, and
- $part-of$ relations to other place names in the same place name structure.

Technically, a place name structure in BUSTER is represented using a specific XML schema (Figure B.1). This schema incorporates the following general sections:

- A header in which the URL of the underlying qualitative spatial reference model is specified,
- an initialization part, in with all place names in the PNS are listed, and
- an instantiation part, in which the extension-definitions and the mereologic relations of all place names in the PNS are given.

While place name structures are used analogous to the application ontologies described above, the underlying qualitative spatial reference model can be compared to the domain ontology. It defines the (geographic) domain of interest as well as the shared vocabulary in terms of units of a reference tessellation. Spatial reference models are encoded in an XML-based format that is very similar to that of place name structures (see Figure A.1). In general, an XML file encoding a spatial reference model has the following parts:

- A header, defining the name, the namespace, and the XSD schema of the reference model,
- the names of the different (type) levels that are represented in the model,
- the names, types, and (unambiguous) identifiers of the reference units contained in the reference model, and
- the hierarchical structure of the reference model, expressed as an XML sub-concept hierarchy.

8.2.3 Query Specification and Reasoning

The BUSTER prototype provides a graphical user interface that supports the specification of combined conceptual, spatial and temporal queries. The first step to specify a BUSTER query is the selection of appropriate *application domains* (Figure 8.10). This is necessary to narrow the search space of the query and to specify which common *domain ontologies* are used. From a list of all domain ontologies registered with the system, the user may select one domain ontology for each search dimension.

In the next step, the user is asked to select an appropriate application ontology (Figure 8.11)[3]. From the conceptual, spatial, and temporal application

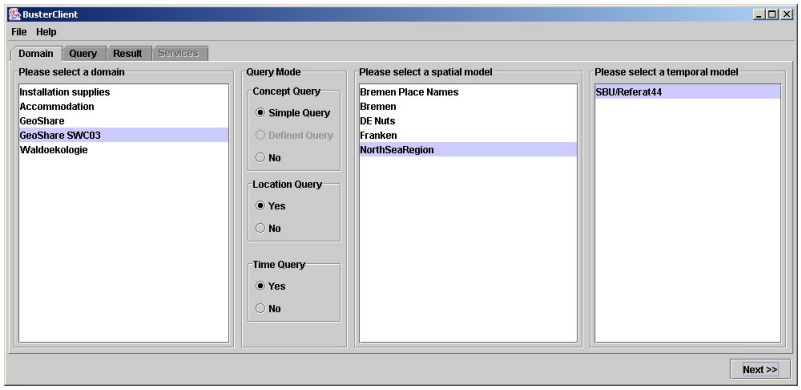

Figure 8.10: BUSTER GUI for the selection of application domains

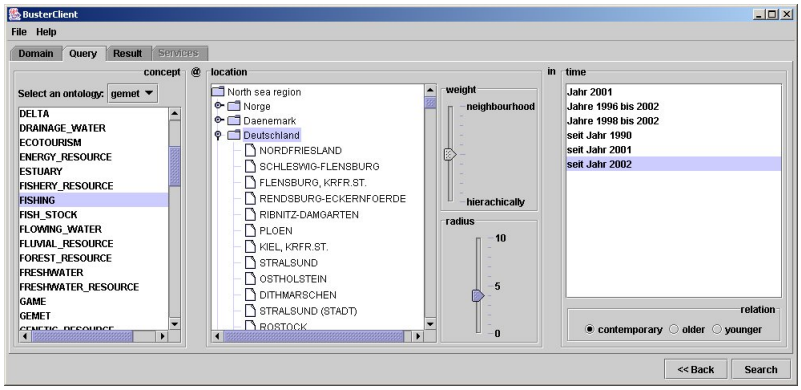

Figure 8.11: BUSTER GUI for query specification

ontologies chosen, the user may select a thematic concept, a place name, and time period, respectively[4].

For the spatial part of the query this means that the user first selects a place name structure $\mathcal{P}_q$ that covers the (geographic) region of interest for the intended query. He/she then then selects the place name of interest q from that place name structure.

To parametrize the spatial part of the query correctly, the user may adjust the parameter α that is used to define the bias on horizontal or vertical distance evaluation (section 7.1). The user may also adjust the cutoff value r_{max}. This parameter determines the extent of the horizontal search radius and affects the normalization of the computed relevance values.

Once the query is specified and submitted, the BUSTER system expands the query in all three search dimensions. To do so it uses the conceptual, spatial and

[3]Note that in the current prototype of BUSTER (version 1.0), no distinction is made between domain ontologies and application ontologies for the spatial and the temporal dimension.

[4]In case of the conceptual query, the user has the choice between a a *"simple query"*, where the user may select a keyword directly from a selected application ontology as described above, or a *"complex query"*, where concepts can be described on the basis of *necessary conditions* as defined in the domain ontology.

temporal reasoning mechanisms described above. The result of this operation are three sets of search terms that contain the initial keyword as well as relevant concepts, place names, and period names.

The output from the BUSTER spatial module is a sorted list of tuples Q_S, where each tuple consists of a reference to a place name p and the relative spatial relevance of p with respect to q, $\sigma(p, q)$ (Figure 8.12). Q_S is a subset of all place names p that belong to the involved place name structures and that are within the horizontal search radius specified by r_{max}.

Figure 8.12: Input and output of the BUSTER/Q spatial module (schematic)

8.2.4 Information Filtering

In BUSTER, complex queries of the type *concept@location in time* are implemented as a sequential process involving an conceptual, a spatial, and a temporal information filter. The sequencing of these filters is variable. However, in the current BUSTER prototype it is implemented in the following way:

Filter 1: Conceptual From the set O of all information objects that are referenced in the BUSTER metadata repository, the subset C containing those information objects that match the conceptual query is selected. An information object $o \in O$ is said to match the set of keyword Q_C specified in the conceptual query if at least one of the conceptual keywords that describe o are in Q_S.

Because subsumption is a binary relation, i.e. a concept c_1 is either subsumed by a concept c_2, or it is not, the set C is not ordered. Consequently, the conceptual relevance that is assigned to all elements $c \in C$ is set to unity.[5]

Filter 2: Spatial In the second process step of the combined query, the information objects in C are compared to the ordered list Q_S, which is the output of the BUSTER spatial module (see above).

For each element in C, the place name descriptors in the respective metadata set are matched against the place name descriptors in Q_S. Only those elements in C that match at least one place name descriptor in Q_S are selected. The resulting set of selected information objects, S, is a subset of C. Each element in S is assigned the spatial relevance of the respective place name descriptor. In the case of several matching place name descriptors, the highest spatial relevance is selected.

[5]The assumption of a binary conceptual relevance is a very crude approximation. Work is under way to integrate more sophisticated evaluation of conceptual relevance in future implementations of the BUSTER/Q module

Filter 3: Temporal In the third and final process step of the combined query, the temporal metadata (i.e. period name descriptors) of the information objects in S are evaluated against the expanded list of period names. The resulting set of information objects, T, is a subset of both S and C with $T \subseteq S \subseteq C$.

The final ranking σ^* of all information sources that were found to be relevant is a function of their conceptual relevance σ_C, their spatial relevance σ_S, and their temporal relevance σ_T. The current BUSTER prototype uses a simple formula to compute σ^* [6] :

$$\sigma^* = (\sigma_C + \sigma_S + \sigma_T)/3 \tag{8.4}$$

The result of the information filtering is a list of relevant information objects, ranked according to relevance (Figure 8.13).

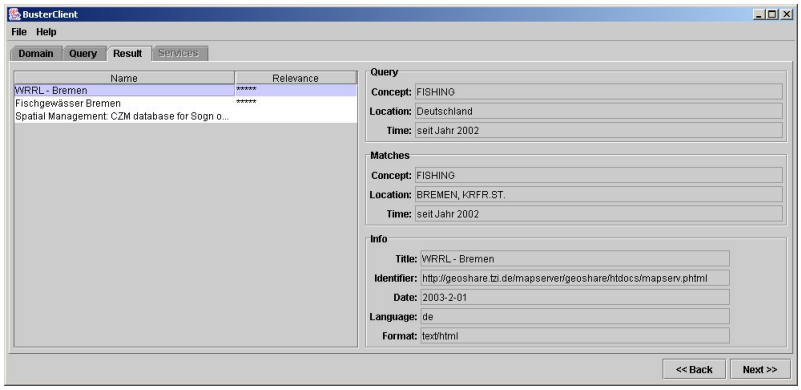

Figure 8.13: Display of search results in the BUSTER GUI

8.2.5 The Comprehensive Source Description (CSD)

The information filtering described above is performed against metadata descriptions of the data and services registered with the system. In BUSTER, each information object is described through a *Comprehensive Source Description (CSD)*. It is important to note that the CSD of an information object resides with the object, i.e. on the server of the information provider. The BUSTER system maintains only a database of references to the information objects and their CSDs. This architecture ensures that the CSDs remain up-to-date and that data and services that are (temporarily or permanently) not available do not appear in the result set of a BUSTER query.

The *Dublin Core (DC) Element Set, Version 1.1* [DCMI, 1999] was used as a basis for the metadata schema for the CSD. Since February 2003, this version of the DC is an official ISO standard, namely *ISO 15836-2003* [ISO, 2003a].

[6] In the current version of the BUSTER system, σ_C is set to unity for all information objects that pass the first (conceptual) filter. The overall relevance ranking of an information object is therefore a function of the second (spatial) and third (temporal) filter.

```
<dc:subject>
<csd:topic-area rdf:resource="geoshare:leisure-activity" />
<csd:topic-area rdf:resource="geoshare:bathing" />
<csd:topic-area rdf:resource="geoshare:bathing-water-directive" />
</dc:subject>
```

Figure 8.14: Examples for the *CSD.topic-area* element

The choice of the DC to become the basis for the CSD metadata schema was based on the simplicity of the DC and its good international reputation and acceptance. The purposely small set of metadata elements that represent the DC is compatible to most other metadata standards. Therefore, the DC is a generic, domain independent standard and mappings between the DC and sector-specific metadata standards (e.g., *ISO 19115* for geospatial data) is easy.

However, some of the metadata elements defined by the Dublin Core proved to be not sophisticated enough for the conceptual, spatial and temporal reasoning with regard to their expressiveness (e.g. the *DC.relation* element) or lack of formal semantics (e.g. the *DC.description* element). There was a need for additional qualifiers for these elements that can be described in a language providing formal semantics (e.g., DAML, OIL, SHIQ, or OWL). For this purpose, some of the DC elements were extended to be able to encode additional features. When possible, we used the RDF(S) syntax for these extensions to ensure a wide acceptance with respect to accessibility and usability. Please note that the expressiveness of RDF(S) was sometimes too limited. In those cases, we referred to explicit ontologies available on the Web. The following DC elements were extended in the BUSTER CSD:

***DC.subject*:** The best practice guidelines issued by the DCMI recommend the use of keywords from controlled vocabularies as fillers for the *DC.subject* metadata element. In this context, the DCMI recommendations name a number of thesauri and classification schemes (e.g. the *Library of Congress Subject Headings*, the *Medical Subject Headings*, and the *Universal Decimal Classification*).

To be able to use formal ontologies as a basis for the subject specifications in a CSD, the *DC.subject* element was refined by a *CSD.topic-area* element (Figure 8.14). This element allows to specify a conceptual *"keyword"* on basis of a formal ontology, including a reference to the respective name space. Within a *DC.subject* element, it is possible to specify multiple *CSD.topic-area* elements, i.e. multiple keywords.

***DC.coverage*:** Since in the DC there is no further distinction between spatial and temporal coverage, the DC element *coverage* had to be extended. Two new sub-elements of *DC.coverage*, *CSD.spatial-coverage* and *CSD.temporal-coverage* were introduced.

CSD.spatial-coverage: The best practice recommended by the Dublin Core Metadata Initiative (DCMI) is to select the filler of *DC.coverage* from a controlled vocabulary. Where appropriate, place names or time periods should be used in preference to numeric identifiers such as sets of coordinates or date ranges. As controlled vocabularies for place names, the DCMI recommends the usage of place name lists,

```
<dc:coverage>
<csd:spatial_coverage rdf:resource="eunuts:BREMEN, KRFR.ST." />
<csd:temporal_coverage rdf:resource="temporal:seit_Jahr_2002" />
</dc:coverage>
```

Figure 8.15: Examples for *CSD.spatial-coverage* and *CSD.temporal-coverage*

```
<dc:relation>
<csd:reference alias="eunuts" source="http://www.tzi.de/.../NorthseaRegion-ru.xml"/>
<csd:reference alias="geoshare" source="http://www.tzi.de/.../geoshare.racer"/>
<csd:reference alias="temporal" source="http://www.tzi.de/.../SBU-Referat44.xml"/>
</dc:relation>
```

Figure 8.16: Examples for the *CSD.reference* element

or gazetteers, such as the GETTY Thesaurus of Geographic Names (TGN) [TGN, 2002].

The *CSD.spatial-coverage* element was extended to allow for the specification of place names that are managed in place name structures. This includes the specification of both the place name and the respective name space, i.e. place name structure (Figure 8.15)

***CSD.temporal-coverage*:** In DC, the recommend best practice for temporal metadata to use either the *DCMI Period* or the *W3C-DTF* encoding scheme. *DCMI Period* was developed by the *DCMI* and allows for the definition of time intervals through the specification of the respective start and end points as time stamps. *W3C-DTF* represents the *W3C* encoding rules for dates and times and is a profile based on *ISO 8601* [ISO, 2000].

The new element *CSD.temporal-coverage* allows for the specification of time points and time intervals in terms of qualitative expression, the so-called *period names*. Analogous to *CSD.spatial-coverage*, the *CSD.temporal-coverage* element encodes both the period names and the respective name space, i.e. the respective ontology of period names (Figure 8.15).

***DC.relation*:** The DCMI recommendations contain only a limited set of qualifiers for the *DC.relation* element. Therefore a *CSD.reference* element was introduced that extends *DC.relation* to be able to include explicit references to thesauri and gazetteers, but also to formal ontologies, place name structures, period name lists and other common vocabularies.

In the *CSD.reference* element, the reference to a specific common vocabulary is encoded as an *XML name space*. This description consists of the URI of the resource and an alias which is used internally as a prefix to mark terms that belong to the respective name space (Figure 8.16).

Figure 8.17 shows an extract from a typical CSD metadata description. The information object described in this CSD is a set of land-use data for the region of Lower-Saxony, a federal state in the northern part of Germany. The sample shows only the elements that are different from the standard DC schema and that were described above.

The *DC.subject* element contains multiple entries for *CSD.topic-area* that

```
 <?xml version="1.0" encoding="ISO-8859-1" ?>
- <rdf:RDF
  …
  <dc:title>Database for land use in southern Lower Saxony, Germany</dc:title>
  <dc:type>dataset</dc:type>
  <dc:source>http://www.tzi.de/buster/data/csd/clc.dbf</dc:source>
- <dc:subject>
    <csd:topic-area rdf:resource="gemet:land-use" />
    <csd:topic-area rdf:resource="gemet:land-use-classification" />
    <csd:topic-area rdf:resource="gemet:landscape-utilisation" />
  </dc:subject>
- <dc:description>
    <geodesy:reference   rdf:resource="geodesy:Bessel-ellipsoid-1841" />
    <csd:table-name      rdf:resource="corine:clc" />
    <csd:table-attribute rdf:resource="csd:row_number" />
    <csd:table-attribute rdf:resource="csd:identifier" />
    <csd:table-attribute rdf:resource="corine:nomenclature" />
    <csd:table-attribute rdf:resource="geodesy:Bessel-ellipsoid-1841" />
    <csd:table-attribute rdf:resource="csd:tknr" />
  </dc:description>
- <dc:relation>
    <csd:reference alias="csd"     source="csd.rdfs" />
    <csd:reference alias="gemet"   source="gemet.rdfs" />
    <csd:reference alias="corine"  source="corine.rdfs" />
    <csd:reference alias="geo"     source="germany.xml" />
    <csd:reference alias="tgn"     source="tgn.daml" />
    <csd:reference alias="geodesy" source="geodesy-ontology.rdfs" />
  </dc:relation>
- <dc:coverage>
    <geo:state  rdf:resource="tgn:Niedersachsen" />
    <geo:region rdf:resource="geo:Northwest-Germany" />
  </dc:coverage>
- <dc:creator>
+ <dc:rights>
…
  </rdf:RDF>
```

Figure 8.17: Example for an Comprehensive Source Descriptor (CSD)

specify concepts taken from the GEMET thesaurus, an multi-language European thesaurus for the environmental domain [Nax and Jensen, 1999].

8.3 Enhanced Metadata: Intelligent Thumbnails

8.3.1 Spatial Metadata and Catalog Services

The main purpose of spatial data infrastructures (SDIs) is to improve the distribution of and the access to digital maps and other geospatial data. The currently prevailing technical architecture paradigm for SDIs is that of interoperable (geo)web-services. The *OpenGIS Consortium (OGC)*, together with the International Organization for Standardization (ISO), has taken a lead in the development of specifications and standards for a number of such services (e.g., [OGC, 2001b], [OGC, 2003]). Catalog services [OGC, 2002a] play a central role in this framework. They use metadata descriptions to manage and to provide access to all resources (i.e., data as well as services) that are registered in an SDI.

To describe the properties of data sources, the *OGC* suggests to use the *ISO 19115* metadata scheme. This standard defines a data model that provides a comprehensive description of technical, administrative, and content related parameters of a data source (see section 2.2). Nevertheless, as most other metadata standards, *ISO 19115* does not describe the contents of a data source in detail, which proves to be a drawback with respect to intelligent information retrieval.

For example, a digital map may be described to cover a certain (rectangular)

geographic area (given as the bounding box coordinates of the map) and to contain information about a number of thematic topics (described through multiple keyword entries). This information suffices to select the map as potentially relevant with respect to an information request. However, it is impossible to deduce from this information whether the digital map actually holds information about a specific topic at a specific sub-section of the region covered. Usually the only way to find out is to download (often this means to purchase) the map and load it into a GIS.

The digital map of Federal Land Features of the United States published by the U.S. Geological Survey (USGS) [USGS, 2003], for example, covers the contiguous and non-contiguous United States. The data set contains the polygon features of all federally administered lands larger than 640 acres, their name identifiers, and information about the federal agencies in charge. The metadata description of this map (using the *Content Standard for Digital Geospatial Metadata published by the FGDC* [FGDC, 1994]) contains information about the spatial coverage of the map in terms of a bounding box, as well as keywords describing it's thematic content. As a result, an information request like "*National Parks @ Contra Costa County*" will rate the Federal Land Features map as relevant because firstly it contains information about national parks, and secondly it covers the whole contiguous United States, which Contra Costa County is part of. However, after a time-consuming download (the data set exceeds 50 MB), a detailed analysis of the data with the help of a GIS would reveal that there are in fact no National Parks in Contra Costa County, making the data set unsuitable for the intended purpose. To avoid this tedious and costly process, some sort of map-preview is needed that shows which topics are covered at which locations within the map.

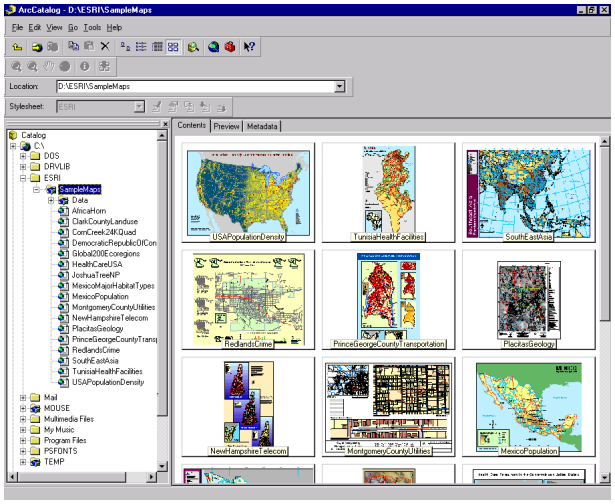

Figure 8.18: Visual thumbnails in ArcCatalog

In some of the more advanced GIS tools (like ESRI's ArcGIS 8.x desktop GIS), such previews are offered as part of the integrated data catalog facilities (Figure 8.18). They are often called *"thumbnails"* because they consist of small renderings of the digital map in question. Although these small bitmaps

server their purpose of providing the user with a quick overview of the available resources and their content, they have two major shortcomings:

- Only a selection of the thematic layers in a digital map can be shown in a thumbnail image. This and the small format of the previews severely constrains their information content.
- Bitmap images of digital maps have to be analyzed by human users. They are not a sound basis for automatic information retrieval algorithms as needed in an efficient catalog service.

The latter problem could be addressed by advanced image processing methods [Miene et al., 2003] which are able to automatically extract information from bitmaps. However, the approach would be counter-productive, as information that was initially available in computer-accessible form (i.e., as a vector map) would first be generalized in form of a bitmap, only to be tediously extracted from the bitmap again.

In the approach outlined below and first described in [Schlieder and Vögele, 2002], we propose to create machine-readable indices of digital maps. Analogous to the digital indices used for full-text searches on text documents, such non-visual but "intelligent" thumbnails provide computer-readable indices of the thematic and spatial contents of digital maps.

8.3.2 Thematic Projection

An *"intelligent"* thumbnail is created through a projection of the thematic content of a digital map onto a polygonal tessellation which is part of a qualitative spatial reference model (see chapter 5). In that sense, the representation of thematic regions in an *"intelligent"* thumbnail follows the same principles as the representation of place name regions in a place name structure (see chapter 6). The result is an index that links the thematic concepts that are covered by a digital map to the locations within the map at which they are found. Depending on the granularity of the underlying reference tessellation, the spatial references in the index are more or less crude approximations of the true locations.

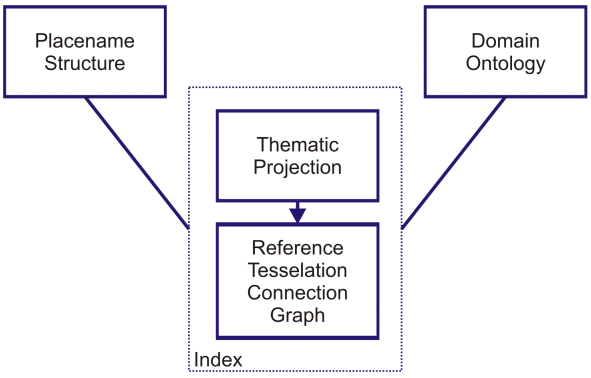

Figure 8.19: Components of an *"intelligent"* thumbnail

As we pointed out above, an "intelligent" thumbnail is created by mapping the thematic layers of a digital map onto the a polygonal standard reference

tessellation (pSRT). Given a digital map where the thematic concepts are represented as vector data (i.e., as point, line, or polygonal objects) this mapping can be done very efficiently from within a standard GIS, using standard GIS functionality. Good results were achieved with a prototypical extension for ESRI's ArcView desktop GIS. The extension uses a GIS-specific polygonal representation of the pSRT (i.e. ESRI shape files) for the mapping task. The result of the process is an XML-encoded list of thematic concepts, assigned to the name descriptors of the pSRT polygons.

8.3.3 Relevance Measures

The purpose of the "intelligent" thumbnail of a digital map is to support reasoning about whether the map is relevant with respect to a specific information request. The level of relevance assigned to a map depends on how closely the content of the map matches the query. This has to take into account direct matches of the specified concept and location, but also near matches which result from terminological specialization or generalization of the concept, as well as spatial generalizations of the location.

Our approach for the approximation of locations and spatial regions, and the relevance reasoning based on these approximations, was described in detail in chapters 5 and 6. Because the *"intelligent"* thumbnail is build on the basis of the qualitative spatial reference models described there, the same algorithms to compute horizontal distance, vertical distance, cumulative distance and ultimately the spatial relevance of a given region of interest and the locations included in the index do apply. The only pre-requisite for this approach is that both the information request and the *"intelligent"* thumbnail use the same spatial frame of reference (i.e., the same spatial reference model).

Thematic relevance is more difficult to evaluate. For the representation of thematic concepts, we use an approach very similar to that applied in the BUSTER system (see section 8.2), i.e. we describe concept in terms of formalized ontologies. The thematic concepts in a digital map are expressed as terms from that belong to a domain-specific ontology. Subsumption reasoning is used to find sub-concepts that are related to the query expression. For example to index the map of US federal lands mentioned above we would use an ontology representing the organizational structure of US federal agencies and federal lands.

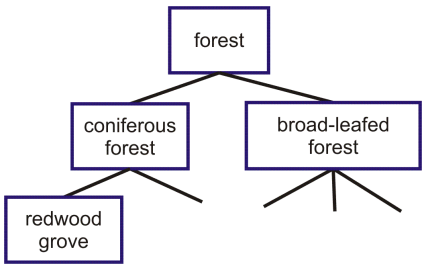

Figure 8.20: A simple terminological ontology (schematic)

If two concepts C_q and C_t are part of the same formal ontology, a theorem prover can be used to reason about the semantic distance s between them. In

the simplest case, a binary metric based on subsumption could be defined: C_t is either a sub-concept of C_q, or it is not, making s to be either 1 or 0. Of course, such a simplistic approach does not suffice and in the long run has to be replaced by a more powerful approach. Jones [Jones et al., 2001], for example, proposes a weighted shortest path procedure to evaluate the semantic distance between two concepts C_q and C_t in a semantic net.

In this work we focus on the solution of the location-part of spatio-thematic information requests. For that reason, we will not go into further detail with respect to the representation of formal ontologies and terminological reasoning. There are a number of approaches and systems available for this task (a summary can be found in [Wache et al., 2001]). As in the BUSTER system described in the previous section, we propose to use formal ontologies like the Ontology Interchange Language (OIL) [Fensel et al., 2000] [Stuckenschmidt, 2000] in connection with appropriate theorem provers (e.g., the RACER theorem prover [Haarslev and Möller, 2001]).

Independently from the method that is used to evaluate semantic distances, the final result of a spatio-thematic query on a set of data sources indexed by an intelligent thumbnail is the ranking of these data sources based on their spatial and thematic relevance with respect to the query. The basis for such a ranking is an integrated measure RS, which can be computed as a linear combination of the cumulative spatial distance C and the semantic distance s:

$$RS = ßC + (1 - ß)s \tag{8.5}$$

Analogous to the computation of the cumulative spatial distance, a weighting factor ß can be used to bias the query towards spatial relevance (ß = 1), or thematic relevance (ß = 0).

In our example, an information request like "Give me all digital maps that describe *NPS-lands @ Contra Costa County*" would yield a result set that comprises all digital maps in the SDI where the "intelligent" thumbnail contains a reference to any federal lands managed by the National Park Service (NPS) and located in or in the vicinity of *Contra Costa County*. The map with the highest ranking would be the one with the lowest RS measure for the respective concept/location tuple in the digital thumbnail.

8.3.4 Data Reduction and Interoperability

The main objective for creating intelligent thumbnails is to be able to index and preview the spatial and thematic contents of digital maps without having to access the data set in full. Especially in distributed and heterogeneous data exchange infrastructures, it is very important to keep the size of the thumbnails and the underlying qualitative spatial reference models as small as possible, and to use open, non-proprietory data formats for their representation. At the same time, enough information has to be retained to support complex queries of the type *concept@location*, and to rank digital maps based on the thematic and spatial relevance metrics described above. To achieve this objective, a number of data reduction and generalization measures have to be applied:

1. The first generalization step is the selection of a subset of all the thematic layers in a digital map. This selection reflects the data provider's choice

of which thematic contents of the data source is important, and should therefore be included in the intelligent thumbnail.

2. In a second generalization step, thematic concepts are abstracted from objects and lumped into a number of feature types. A specific site (i.e. *"John Muir National Historic Site"*) is lumped together with other sites and represented as one feature type (i.e. *"National Historic Site"*). Depending on the level of thematic detail wanted, the index can be condensed even further by using concepts that are higher up in the concept hierarchy. The feature type *"National Historic Site"*, for example, can be further abstracted as "*land managed by the NPS"*.

3. Maybe the largest generalization and data reduction takes place during the thematic projection: All selected feature types in a digital map are projected onto a standard reference tessellation, but if two identical concepts happen to map onto the same spatial entity, only one reference is maintained. The thematic concept *"National Historic Site"*, for example, would be assigned only once to the location *"Contra Costa County"*, even if several such sites exist in the same spatial unit. Applied to the *USGS Federal Land Features* map, the initial 53 MB of map data (ESRI shape format) can be reduced to an index of about 500 KB (XML ASCII format). This amounts to a data reduction factor of approximately 1 : 100.

4. Considerable data reduction can also be achieved by using a qualitative, graph-based representation of polygonal standard reference tessellations. For example, the 9.8 MB ESRI shape file holding the polygonal tessellation of all counties within the contiguous United States could be reduced by a factor of 6 to a connection graph of 1.6 MB (XML ASCII format).

Chapter 9

Summary and Outlook

9.1 Summary

Given the growing number of information objects that become accessible through data collections like the WWW, digital libraries, and specialized information networks (e.g., spatial data infrastructures for the distribution of geospatial data), the need for tools that support efficient information discovery and information retrieval is obvious. The evolution of the so-called *"Semantic Web"* proves that the scientific community is well aware of this fact and that efforts are under way to push for a new quality of information discovery and retrieval.

This work tries to contribute to these efforts by concentrating on information discovery and information retrieval driven by the *spatial relevance* of information objects. The spatial dimension of information retrieval has (with the exception of a few notable exceptions) been neglected so far by the IR community, in spite of the fact that a growing number of *"spatially-aware"* tools are needed by new technologies like mobile and location-based services and service-based spatial data infrastructures.

In the approach to spatial information retrieval outlined in this work, a low-level notion of spatial relevance is applied. Spatial relevance is closely linked to the geographic location an information object can be associated with, i.e., to its *geo-reference*. Common-sense knowledge (supported by cognitive studies) tells us that most geographic locations, or places in geographic space, are given names. These *place names* are used to describe geographic space, but also to geo-reference information objects. An analysis of some of the most relevant metadata standards showed that *indirect geo-referencing* by place names is not only important for unstructured documents, but also for structured (meta)data. There are indications that a significant fraction of the information objects available in large information collections (e.g., the WWW) are indirectly geo-referenced with place names, rather than directly through geographic coordinates.

For each place name there is a spatial footprint that describes its regional extent in geographic space. By evaluating some of the basic spatial relations between such footprints, conclusions about the spatial relevance of the respective place names can be drawn. In most state-of-the-art spatial information systems

(e.g., GIS, spatial databases, and gazetteers), spatial footprints are expressed in terms of exact, coordinate-based representations. We showed that spatial reasoning based on such quantitative representations can have serious limitations because quantitative footprints often abstract too much from the true regional extent of a place name. The reason is the intrinsically vague nature of many place names, but also the unavailability of suitable digital representations.

To overcome some of the limitations of coordinate-based spatial footprints, a representation for spatial regions on the basis of discrete partitionings of geographic space was developed. To be able to create tools that are intuitive and easy to use for a human user, partitionings based on polygonal standard reference tessellations, like administrative subdivisions or postal code zones, were applied. Such tessellations have the advantage that they represent standardized and well-known place names which are organized in a meaningful hierarchical structure.

To reduce data volumes and to support efficient reasoning algorithms we abstracted from the polygonal representation of a reference tessellations and used graph-based representations instead. In such a representation, neighborhood relations between polygons are encoded as *connection graphs*, while their hierarchical *part-of* relations are formalized as a *decomposition tree*. Compared to polygonal representations, such *qualitative spatial reference models* are lightweight and highly interoperable. Even with an encoding that was not optimized for efficient data storage, a reduction of data volumes by a factor of 4 to 6 could be achieved. Data interoperability could be improved by using an open XML-based format.

Based on qualitative spatial reference models as common frames of reference, a representation of regions in terms of sets of spatial indices was developed. It could be shown that with such *qualitative spatial footprints*, both crisp and vague (place name) regions can be approximated with a precision that is sufficient for the spatial reasoning needed by most information retrieval applications. An architecture of *place name structures* which can be used to organize place names in a hierarchical structure similar to that of a qualitative spatial reference model was outlined.

To evaluate the relative spatial relevance between units of a spatial reference model, we developed a simple metric that is based on the evaluation of graph-theoretical distances in the qualitative spatial reference model. This approach was extended to evaluate the spatial relevance of place names based on their qualitative spatial footprints and their position within the place name structure. Because this metric combines both the *"horizontal distance"* (i.e., the neighborhood distance in a connection graph) and the *"vertical distance"* (i.e., the node distance in the graph representing the hierarchy of a place name structure), it is able to evaluate the relative spatial relevance of two place names both in terms of their geographic relevance, which is related to their distance in Euclidean space, and their partonomic distance, which represents their distance within the organization-hierarchy of a place name structure.

In addition to a mere evaluation of spatial distance, this approach applies a basic notion of the semantics of a place name structure (i.e., the hierarchical part-of relations of place names that represent their organizational context) to improve the results of an information retrieval process. Two place names that are vertically close are assumed to have a strong relationship in terms of the organizational scheme of the place name structure, while they need not be

spatially close. By manipulation of a single parameter α, a spatial query can be biased either to focus on horizontal, or on vertical distance.

9.2 Outlook

The methods outlined in this paper have been implemented in the *BUSTER* information broker middleware, where they are used in support of queries of the type *concept location in time* (see section 8.2). Within the framework of the *GeoShare* project[1], the *BUSTER* system will be integrated in a spatial data infrastructure based on current *OGC* standards [Vögele et al., 2003c, Vögele and Spittel, 2004]. As part of this integration, it is planned to implement the spatial reasoner module as an *OGC*-compatible (geo)service. As part of a spatial data infrastructure, this tool will be accessible to any *OGC*-compatible catalog service, providing specialized spatial information retrieval capabilities.

Because our approach uses a light-weight and interoperable representation, as well as a simple but effective reasoning algorithm, it may be applied to a number of application areas. For some of them, potential use-cases were already outlined: One example is the use of thematic projections based on qualitative spatial reference models as machine-readable indices of digital maps [Schlieder and Vögele, 2002]. Another potential application is the implementation of spatial metadata and light-weight reasoning modules in highly distributed ad-hoc (*P2P*) networks [Vögele and Schlieder, 2002]. Finally, it could be shown that the methodology can be used to retrieve (geospatial) data in *Semantic Web* applications [Hübner et al., 2004].

For some of the use-cases mentioned above, prototypical systems could already be implemented successfully. However, for a more general application in large-scale real-world scenarios, an optimization of these implementations would be necessary. This refers to an optimization of the encoding of the qualitative spatial reference models and place name structures, as well as the optimization of the spatial reasoning module (with respect to run-time performance).

On the other hand, the evaluation described in section 8.1 showed that some of the initial assumptions of this work had to be revised when moving from an experimental environment to large-scale real-world data collections. For example, the initial assumption about the optimal choice of the query-bias factor α was that it is a function of the intrinsic linguistic semantics of the conceptual part of a spatio-thematic query. According to this assumption, a horizontally-biased query should perform better for concepts that are important in a purely spatial context, while a vertically-biased query should mainly retrieve information objects that are associated with concepts that are relevant in an organizational context.

This was only partly supported by the evaluation runs: Although the results showed that both the precision and the recall of an information request do depend on the choice of the factor α, it turned out that this dependence does not necessarily reflect the linguistic semantics of a concept, but rather the concept semantics assigned by the respective *information community* (i.e., the human experts responsible for the metadata-annotation of an information collection). In case of the *UDK* metadata catalog used in the evaluation, this

[1] GeoShare is a project co-funded by the EU *Interreg IIIB North Sea Region Programme* and involved 5 partners from 4 countries in the north sea region.

means that the criteria applied to geo-reference the same conceptual type of information object may be different in each of the 16 German federal states that contribute to the collection. Accordingly, no general recommendation for the optimal parametrization of the query can be made, but has to be found on a case-to-case basis.

Another result of the evaluation was that in a standard real-world data collection like the *UDK* database, many information objects are assigned multiple geo-references that refer to units on the lower levels of a place name hierarchy (e.g., *"Gemeinden"*), while they are in fact relevant for the parent-region(s) on the next higher level (e.g., the whole *"Landkreis"* or *"Regierungsbezirk"*). Multiple and redundant geo-references are often used by the creators of metadata to ensure that the data set will be found by standard spatial retrieval procedures, even if the location of interest is on a low level of the hierarchy. Through our approach to place name structures and spatial reasoning such redundancies can be avoided. This could lead to simpler (but not less expressive) metadata, as well as to a standardization of the respective annotation criteria.

In summary, the further development of the methodology developed in this work should focus on the application in large-scale, real-world information retrieval systems. The author is confident that with the help of qualitative spatial reference models, place name structures, and the respective spatial reasoning algorithms, a number of existing and emerging applications can be improved with respect to spatial information retrieval.

Bibliography

[ADL, 2002] ADL (2002). Glossary of terms - alexandria digital library. http://fat-albert.alexandria.ucsb.edu:8827/glossary.html.

[ADL, 2004] ADL (2004). Web site of the alexandria digital library. http://fat-albert.alexandria.ucsb.edu:8827/.

[Alani, 2001] Alani, H. (2001). *Spatial and Thematic Ontology in Cultural Heritage Information Systems*. PhD thesis, University of Glamorgan/Prifysgol Morgannwg.

[Alani et al., 2001] Alani, H., Jones, C. B., and Tudhope, D. (2001). Voronoi-based region approximation for geographical information retrieval with gazetteers. *Internation Journal of Geographical Information Science*, 15(4):287–306.

[Allen, 1983] Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

[Allen, 1984] Allen, J. F. (1984). Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154.

[Allen, 1991] Allen, J. F. (1991). Time and time again: The many ways to represent time. *International Journal of Intelligent Systems*, 6(4):341–355.

[Altmann, 1994] Altmann, D. (1994). Set theoretical approaches for handling imprecision in spatial analysis. *International Journal of Geographical Information Systems*, 8(3):271–289.

[Angrick et al., 2002] Angrick, M., Bös, R., and Bandholtz, T. (2002). Semantic network services (sns). In Pillmann, W. and Tochtermann, K., editors, *Proceedings of the 16th conference "Envrionmental Informatics 2002" (EnviroInfo'2002)*, Environmental Communication in the Information Society, pages 78–85, Vienna.

[ANZLIC, 2004] ANZLIC (2004). Web page of ANZLIC. web page, ANZLIC - the Spatial Information Council, 8 January. http://www.anzlic.org.au/.

[Arms, 2000] Arms, W. Y. (2000). *Digital Libraries*. MIT Press, Cambridge, Mass.

[Asher and Vieu, 1995] Asher, N. and Vieu, L. (1995). Toward a geometry of common sense: A semantics and a complete axiomatization of mereotopology. In *IJCAI'95*, pages 846–852. Morgan Kaufmann.

[Atkinson, 2001] Atkinson, R. (2001). Draft gazetteer service specification. OpenGIS Draft Specification OGC Document 01-036, OpenGIS Consortium, March.

[Atkinson and Fitzke, 2002] Atkinson, R. and Fitzke, J. (2002). Gazetteer service profile of the web feature service implementation specification. OpenGIS Discussion Paper OGC Document 02-076r3, OpenGIS Consortium, September.

[Baeza-Yates and Ribero-Neto, 1999] Baeza-Yates and Ribero-Neto (1999). *Modern Information Retrieval*. ACM Press, New York.

[Beinstein and Sievers, 2002] Beinstein, B. E. and Sievers, J. (2002). Digital database of geographical names of Germany (GN 205 and GN 1000). In *Proceedings of the Eighth United Nations Conference on the Standardization of Geographical Names*, Berlin.

[Bennett, 1994] Bennett, B. (1994). Spatial reasoning with propositional logics. In Doyle, J., Sandewall, E., and Torasso, P., editors, *4th International Conference on the Principles of Knowledge Representation and Reasoning (KR94)*, San Francisco. Morgan Kaufmann.

[Berners-Lee et al., 2001a] Berners-Lee, T., Brickley, D., Connolly, D., Dean, M., Decker, S., Fensel, D., Fikes, R., Hayes, P., Heflin, J., Hendler, J., Lassila, O., McGuinness, D., and Stein, L. A. (2001a). DAML+OIL language specification. http://www.daml.org/2001/03/daml+oil-index.html, verified on July, 1st, 2003.

[Berners-Lee et al., 2001b] Berners-Lee, T., Hendler, J., and Lassila, O. (2001b). The semantic web. *Scientific American.com*. http://www.scientificamerican.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

[Bilo and Streuff, 2000] Bilo, M. and Streuff, H. (2000). Das Umweltinformationsnetz Deutschland - GEIN2000 - Fachliche Anforderungen an ein Forschungs- und Entwicklungsvorhaben. In Tochetermann, K. and Riekert, W.-F., editors, *Proceedings of the 3rd workshop Hypermedia im Umweltschutz*, Ulm.

[Bishr and Kuhn, 1999] Bishr, Y. and Kuhn, W. (1999). *The Role of Ontology in Modelling Geospatial Features*, volume 5 of *IFGI prints*. Institut für Geoinformatik, Universität Münster, Münster.

[Bittner, 1999] Bittner, T. (1999). An ontology and epistemology of rough location. In Freksa, C. and Mark, D., editors, *Conference on Spatial Information Theory (COSIT'99)*, volume LNCS 1661 of *Spatial Information Theory - Cognitive and Computational Foundations of Geographic Information Science*, pages 433–448, Stade, Germany. Springer.

[Bittner, 2000] Bittner, T. (2000). Rough sets in spatio-temporal data mining. In Roddick, J. F. and Hornsby, K., editors, *International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, TSDM2000*, volume 2007 of *Lecture Notes in Artificial Intelligence*, Lyon, France. Springer.

[Bittner and Stell, 1998] Bittner, T. and Stell, J. G. (1998). A boundary-sensitive approach to qualitative location. *Annals of Mathematics and Artificial Intelligence*, 24(1-4):93–114.

[Bittner and Stell, 2002] Bittner, T. and Stell, J. G. (2002). Vagueness and rough location. *GeoInformatica*, 6(2):99–121.

[BKG, 1994] BKG (1994). Bundesrepublik Deutschland: Landschaften - Namen und Abgrenzungen. Topographic map, Bundesamt für Kartographie und Geodäsie - Insitut für Angewandte Geodäsie. http://www.ifag.de/.

[BKG, 2002] BKG (2002). Ständiger Ausschuss für Geographische Namen (Stagn). Web page, Bundesamt für Karthographie und Geodäsie (BKG), 10 September. http://www.ifag.de/kartographie/Stagn.

[Borgida et al., 1989] Borgida, A., Brachman, R. J., McGuiness, D. L., and Resnick, L. A. (1989). CLASSIC: A structural data model for objects. In *in proceedings of ACM SIGMOID - International Conference on Management of Data*, Portland, Oregon, USA.

[Bretherton and Singley, 1994] Bretherton, F. P. and Singley, P. T. (1994). Metadata: A user's view. In French, J. C. and Hinterberger, H., editors, *In Proceedings of the 7th International Working Conference on Scientific and Statistical Database Management*. IEEE Computer Society Press.

[Brown, 1998] Brown, D. G. (1998). Classification and boundary vagueness in mapping pre-settlement forest types. *International Journal of Geographical Information Science*, 12(2):105–129.

[Burrough, 1996] Burrough, P. (1996). Natural objects with indeterminate boundaries. In Burrough, P. and Frank, A., editors, *Geographic Objects with Indeterminate Boundaries*, volume 2 of *GISDATA*, pages 3–28. Taylor and Francis.

[Burrough and Frank, 1996] Burrough, P. and Frank, A., editors (1996). *Geographic Objects with Indeterminate Boundaries*, volume 2 of *GISDATA*. Taylor and Francis.

[Burrough and McDonnell, 1998] Burrough, P. and McDonnell, R. (1998). *Principles of Geographical Information Systems*. Spatial Information Systems and Geostatistics. Oxford University Press, New York.

[Casati and Varzi, 1995] Casati, R. and Varzi, A. (1995). The structure of spatial localization. *Philosophical Studies*, 82(2):205–239.

[CEN, 1998] CEN (1998). ENV 12657: Geoinformationen, Datenbeschreibung, Metadaten. proposed european standard, Comite European de Normalization (CEN). http://www.cenorm.be/.

[CEN, 2004] CEN (2004). Web site of cen. Web page, Comite European de Normalization (CEN), 12 January. http://www.cenorm.be/.

[Clarke, 1981] Clarke, B. (1981). A calculus of individuals based on 'connection'. *Notre Dame Journal of Formal Logic*, 22(3):204–218.

[Clarke, 1985] Clarke, B. (1985). Individuals and points. *Notre Dame Journal of Formal Logic*, 26(1):61–75.

[Clementini and di Felice, 1996] Clementini, E. and di Felice, P. (1996). An algebraic model for spatial objects with indeterminate boundaries. In Burrough, P. and Frank, A., editors, *Geographic Objects with Indeterminate Boundaries*, volume 3 of *GISDATA Series*, pages 153–169. Taylor and Francis.

[Clementini and di Felice, 1997] Clementini, E. and di Felice, P. (1997). Approximate topological relations. *International Journal of Approximate Reasoning*.

[Cohn, 1997] Cohn, A. G. (1997). Qualitative spatial representation and reasoning techniques. In Nebel, G. B., Habel, C., and Bernhard, editors, *KI-97, Advances in Artificial Intelligence*, pages 1–30. Springer Verlag, Berlin.

[Cohn, 1999] Cohn, A. G. (1999). Qualitative spatial representations. In *Proceedings of the Workshop on Adaptive Spatial Representations in Dynamic Environments, IJCAI-99*, San Francisco. Morgan Kaufmann Publishers.

[Cohn and Gotts, 1996a] Cohn, A. G. and Gotts, N. (1996a). The 'egg-yolk' representation of regions with indeterminate boundaries. In Burrough, P. and Frank, A., editors, *Geographic Objects with Indeterminate Boundaries*, pages pp. 45–55. Taylor and Francis, London.

[Cohn and Gotts, 1996b] Cohn, A. G. and Gotts, N. (1996b). A mereological approach to representing spatial vagueness. In Doyle, J., Aiello, L., and Shapiro, S., editors, *5th Intl. Conf. on Principles of Knowledge Representation (KR'96)*, pages 230–241. Morgan Kaufmann.

[Cohn and Hazarika, 2001] Cohn, A. G. and Hazarika, S. M. (2001). Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 42:2–32.

[Couclelis, 1992] Couclelis, H. (1992). People manipulate objects (but cultivate fields): Beyond the raster-vector debade in gis. In Frank, A., Campari, I., and Formentini, U., editors, *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space International Conference GIS - From Space to Territory*, volume 639 of *Lecture Notes in Computer Science*, pages 65–77, Pisa, Italy. Springer-Verlag New-York.

[Cox, 2000a] Cox, S. (2000a). DCMI box encoding scheme: Specification of the spatial limits of a place, and methods for encoding this in a text string. electronic document, Dublin Core Metadata Initiative, 28 July. http://dublincore.org/documents/2000/07/28/dcmi-box/.

[Cox, 2000b] Cox, S. (2000b). DCMI point encoding scheme: A point location in space, and methods for encoding this in a text string. electronic document, Dublin Core Metadata Initiative (DCMI), 28 July. http://dublincore.org/documents/2000/07/28/dcmi-point/.

[DCMI, 1999] DCMI (1999). Dublin core metadata element set (dcmes), version 1.1: Reference description (recommendation). Technical report, Dublin Core Metadata Initiative (DCMI), July. http://dublincore.org/documents/1999/07/02/dces/#rfc2413.

[DCMI, 2000a] DCMI (2000a). DCMI period encoding scheme: Specification of the limits of a time interval, and methods for encoding this in a text string. Technical report, Dublin Core Metadata Initiative (DCMI), July. http://dublincore.org/documents/2000/07/28/dcmi-period/.

[DCMI, 2000b] DCMI (2000b). The dublin core: A simple content description model for electronic resources. electronic document, Dublin Core Metadata Initiative (DCMI), 28 March. http://dublincore.org/.

[de Berg et al., 2000] de Berg, M., Schwarzkopf, O., van Kreveld, M., and Overmars, M. (2000). *Computational Geometry: Algorithms and Applications*. Springer-Verlag, second edition.

[Decker et al., 1999] Decker, S., Erdmann, M., Fensel, D., and Studer, R. (1999). Ontobroker: Ontology based access to distributed and semi-structured information. In et al., R. M., editor, *Proceedings of DS-8 - Semantic Issues in Multimedia Systems*, pages 351–369. Kluwer Academic Press, Boston.

[DeLaguna, 1922] DeLaguna, T. (1922). Point, line and surface as sets of solids. *The Journal of Philosophy*, 19:449–461.

[Dijkstra, 1959] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

[Dirichlet, 1850] Dirichlet, G. (1850). Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. *Journal für die Reine und Angewandte Mathematik*, 40:209–227.

[Dutta, 1989] Dutta, S. (1989). Qualitative spatial reasoning: A semi-quantitative approach using fuzzy logic. In *Proceedings of the 1st Int. Symp. on the Design and Implementation of Large Spatial Databases*, LNCS 409, pages 345–364. Springer-Verlag.

[Egenhofer and Franzosa, 1991] Egenhofer, M. and Franzosa, R. (1991). Point-set topological relations. *International Journal of Geographical Information Systems*, 5(2):161–174.

[Egenhofer and Kuhn, 1998] Egenhofer, M. and Kuhn, W. (1998). Beyond desktop GIS. In *Proceedings of the International Conference and Exhibition on Geographic Information - GIS PlaNET '98*, pages 281–290, Lisbon, Portugal.

[Egenhofer, 1989] Egenhofer, M. J. (1989). A formal definition of binary topological relationships. In Litwin, W. and Schek, H.-J., editors, *3rd Intl. Conf. on Foundations of Data Organization and Algorithms*, pages 457–472.

[Egenhofer, 1995] Egenhofer, M. J. (1995). Modeling conceptual neighborhoods of topological line-region relations. *Internation Journal of Geographical Information Systems*, 9(5):555–565.

[Egenhofer and Franzosa, 1995] Egenhofer, M. J. and Franzosa, R. (1995). On the equivalence of topological relations. *Internation Journal of Geographical Information Systems*, 9(2):133–152.

[Egenhofer and Mark, 1995] Egenhofer, M. J. and Mark, D. M. (1995). Naive geography. In Frank, A. U. and Kuhn, W., editors, *Spatial Information Theory, A Theoretical Basis for GIS - COSIT'95*, pages 1–16. Springer.

[Egenhofer et al., 1993] Egenhofer, M. J., Sharma, J., and Mark, D. M. (1993). A critical comparison of the 4-intersection and 9-intersection models for spatial relations: Formal analysis. In *Auto-Carto 11*, volume 1, pages 1–11, Minneapolis. ACSM-ASPRS.

[Erwig and Schneider, 1997] Erwig, M. and Schneider, M. (1997). Vague regions. In Scholl, M. and Voisard, A., editors, *Advances in Spatial Databases*. Springer, Berlin.

[Eschenbach and Heydrich, 1995] Eschenbach, C. and Heydrich, W. (1995). Classical mereology and restricted domains. *International Journal of Human-Computer Studies*, 43:723–730.

[Escrig and Toledo, 1998] Escrig, M. T. and Toledo, F. (1998). *Qualitative Spatial Reasoning: Theory and Practice - Application to Robot Navigation*, volume 47 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam.

[EuroGeographics, 2003] EuroGeographics (2003). Seamless Administrative Boundaries of Europe (SABE). electronic document, EuroGeographics, April. http://www.eurogeographics.org.

[EuroGeographics, 2004] EuroGeographics (2004). Web site of eurogeographics. electronic document, EuroGeographics, 9 January. http://www.eurogeographics.org.

[European-Commission, 2003] European-Commission (2003). Regulation (Ec) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the Establishment of a Common Classification of Territorial Units for Statistics (NUTS). *Official Journal of the European Union*, 154:1–41.

[Euzenat, 1995] Euzenat, J. (1995). An algebraic approach to granularity in qualitative time and space representations. In *Proceedings of the International Joint Conference on AI - IJCAI'95*, pages 894–900, Montreal, CA. ACM Publications.

[FAO, 2004] FAO (2004). Web page of the United Nations Food and Agriculture Organization (FAO). electronic document, Food and Agriculture Organization of the United Nations (FAO), 20 January. http://www.fao.org.

[Fensel, 1999] Fensel, D. (1999). Adding semantics to the web. In *Proceedings of the VIII Conferencia de la Asociación Española para la Inteligencia Artificial*, Muncia, Spain.

[Fensel et al., 1998] Fensel, D., Angele, J., Decker, S., Erdmann, M., Schnurr, H.-P., Staab, S., Studer, R., and Witt, A. (1998). On2broker: Lessons learned from applying AI to the web. Technical report, Institute AIFB.

[Fensel et al., 2003] Fensel, D., Hendler, J., Lieberman, H., and W., W., editors (2003). *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential*. MIT Press, Boston.

[Fensel et al., 2000] Fensel, D., Horrocks, I., Harmelen, F. V., Decker, S., Erdmann, M., and Klein, M. (2000). OIL in a nutshell. In *Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management EKAW 2000*, Juan-les-Pins, France.

[FGDC, 1994] FGDC (1994). Content Standards for Digital Geospatial Metadata. Technical report, US Government, Federal Geographic Data Committee (FGDC). ftp://fgdc.er.usgs.gov.

[Fotheringham and Curtis, 1992] Fotheringham, A. S. and Curtis, A. (1992). Encoding spatial information: The evidence for hierarchical processing. *Lecture Notes in Computer Science*, 639.

[Frank, 1992] Frank, A. (1992). Qualitative spatial reasoning about distance and directions in geographic space. *Journal of Visual Languages and Computing*, 3:343–373.

[Frank, 1996] Frank, A. (1996). Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Systems*, 10(3):269–290.

[Freksa, 1991] Freksa, C. (1991). Qualitative spatial reasoning. In Mark, D. and Frank, A., editors, *Cognitive and Linguistic Aspects of Geographic Space*, volume 63, pages 361–372. Kluwer Academic Press.

[Freksa, 1992a] Freksa, C. (1992a). Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54(1):199–227.

[Freksa, 1992b] Freksa, C. (1992b). Using orientation information for qualitative spatial reasoning. In Frank, A. U., Campari, I., and Formentini, U., editors, *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, pages 162 – 178. Springer, Berlin.

[Freksa and Roehrig, 1993] Freksa, C. and Roehrig, R. (1993). Dimensions of qualitative spatial reasoning. In Carrete, N. P. and Singh, M., editors, *IMACS Workshop on Qualitative Reasoning and Decision Technologies, QUARDET '93*, pages 483 – 492, Barcelona. CIMNE.

[Freundschuh and Egenhofer, 1997] Freundschuh, S. M. and Egenhofer, M. J. (1997). Human conceptions of space: Implications for geographic information systems. *Transactions in GIS*, 2(4):361–375.

[Furrie, 2000] Furrie, B. (2000). Understanding MARC bibliographic: Machine-readable cataloging. Technical report, Follet Software Co., Library of Congress.

[Gahegan, 1995] Gahegan, M. (1995). Proximity operators for qualitative spatial reasoning. *Lecture Notes in Computer Science*, 988.

[Galton, 1999] Galton, A. (1999). The mereotopology of discrete space. In Freksa, C. and Mark, D., editors, *Proceedings of the International Conference on Spatial Information Theory COSIT'99*, Spatial Information Theory: Cognitive and Computational Foundations of Geographic Science, pages 251–266, Stade, Germany. Springer, Berlin.

[GINIE, 2003] GINIE (2003). GI in the wider europe. Project report, GINIE: Geographic Information Network in Europe, 23 October. http://www.ec-gis.org/ginie/.

[Günther, 1998] Günther, O. (1998). *Environmental Information Systems*. Springer, Berlin.

[Goasdoué et al., 1999] Goasdoué, F., Lattes, V., and Rousset, M.-C. (1999). The use of carin language and algorithms for information integration: The PICSEL project,. *International Journal of Cooperative Information Systems (IJCIS)*, 9(4):383 – 401.

[Gonzalez and Wintz, 1987] Gonzalez, R. and Wintz, P. (1987). *Digital Image Processing*. Addison-Wesley Publishing Company, Reading, 2 edition.

[Goodchild, 1999] Goodchild, M. (1999). The future of the gazetteer. report transcribed and edited from audiotape, University of California, Santa Barbara, October 12-14. http://www.alexandria.ucsb.edu/ lhill/dgie/.

[Goodchild, 1993] Goodchild, M. F. (1993). Data models and data quality: Problems and prospects. In Goodchild, M. F., Parks, B., and Steyaert, L., editors, *Visualization in Geographical Information Systems*, pages 141–149. John Wiley, New York.

[GOOGLE, 2004] GOOGLE (2004). German web page of the GOOGLE search engine. web page, Google Germany GmbH., 15 January. http://www.google.de.

[Gotts, 1996] Gotts, N. (1996). Formalizing commonsense topology: The inch calculus. In *Proceedings of the 4th Intl. Symposium on Artificial Intelligence and Mathematics*.

[Grüninger and Uschold, 2002] Grüninger, M. and Uschold, M. (2002). Ontologies and semantic integration, software agents for the warfighter. technical report, Institute for Human and Machine Cognition (IHMC), University of West Florida.

[Gruber, 1991] Gruber, T. (1991). Ontolingua: A mechanism to support portable ontologies. KSL Report KSL-91-66, Stanford University, 1991.

[Gruber, 1993] Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

[Guarino, 1998] Guarino, N. (1998). Formal ontology and information systems. In Guarino, N., editor, *FOIS 98*. IOS Press, Trento, Italy.

[Guarino and Giaretta, 1995] Guarino, N. and Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. In Mars, N., editor, *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25–32, Amsterdam.

[Guesgen and Albrecht, 2000] Guesgen, H. W. and Albrecht, J. (2000). Imprecise reasoning in geographic information systems. *Fuzzy Sets and Systems*, 113:121–131.

[Haarslev and Möller, 2001] Haarslev, V. and Möller, R. (2001). Racer system description. In *Proceedings of the International Joint Conference on Automated Reasoning, IJCAR'2001*, Siena, Italy. Springer-Verlag, Berlin.

[Harary, 1971] Harary, F. (1971). *Graph Theory*. Addison-Weseley, Reading, 2 edition.

[Harper-Collins, 2001] Harper-Collins (2001). Collins Bartholomew. digital map, Harper-Collins Publishers. http://www.bartholomewmaps.com.

[Hübner, 2003] Hübner, S. (2003). *Qualitative Abstraktion von Zeit für Annotation und Retrieval im Semantic Web*. Masters thesis, Technologie-Zentrum Informatik (TZI), Universität Bremen.

[Hübner et al., 2004] Hübner, S., Spittel, R., Visser, U., and Vögele, T. (2004). Ontology-based search for interactive digital maps. *to appear in: Intelligent Systems, special issue on Semantic Web Challenge*, ISSI-0013-0204.

[Helleland, 2002] Helleland, B. (2002). The social and cultural value of place names. In *Proceedings of the 8th United Nations Conference on the Standardization of Geographical Names*, Toponymic education and practice and international cooperation: existing education and practice, Berlin.

[Hernandez, 1994] Hernandez, D. (1994). Qualitative representations of spatial knowledge. *Lecture Notes in Artificial Intelligence*, 804.

[Hill, 2003a] Hill, L. (2003a). ADL Gazetteer Content Standard. technical report, University of Santa Barbara. http://www.alexandria.ucsb.edu/gazetteer/gaz/_content/_standard.html.

[Hill, 2000] Hill, L. L. (2000). Core elements of digital gazetteers: Placenames, categories, and footprints. In Borbinha, J. and Baker, T., editors, *ECDL 2000*, Research and Advanced Technology for Digital Libraries, pages 280–290, Lisbon, Portugal.

[Hill, 2003b] Hill, L. L. (2003b). Guide to the ADL Gazetteer Content Standard version 3.1. http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3-1/GCS3-1-guide.htm.

[Hill and Zheng, 1999] Hill, L. L. and Zheng, Q. (1999). Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with developing and implementing gazetteers. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, Knowledge: Creation, Organization and Use, pages 57–69, Washington, D.C. Information Today, Medford, NJ.

[Hillmann, 2001] Hillmann, D. (2001). Using Dublin Core. electronic document, Dublin Core Metadata Initiative (DCMI), 4 December. http://dublincore.org/documents/2001/04/12/usageguide/.

[Hirtle, 2000] Hirtle, S. C. (2000). The use of maps, images and "gestures" for navigation. In Freksa, C., editor, *Spatial Cognition II*, LNAI 1849, pages 31–40. Springer, Berlin, Heidelberg.

[Hirtle and Jonides, 1985] Hirtle, S. C. and Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory & Cognition*, 13(3):208–217.

[Holyoak and Mah, 1982] Holyoak, K. J. and Mah, W. A. (1982). Cognitive reference points in judgements of symbolic magnitude. *Cognitive Psychology*, 14:328–352.

[Horrocks, 1998] Horrocks, I. (1998). The FaCT system. In de Swart, H., editor, *Proceedings of TABLEAUX'98*, volume 1, pages 307–312. Springer-Verlag.

[Hunter and Goodchild, 1993] Hunter, G. J. and Goodchild, M. F. (1993). Managing uncertainty in spatial databases: Putting theory into practice. In *Proceedings of URISA*, volume 1, pages 1–14, Atlanta, Georgia.

[ISO, 1997] ISO (1997). Codes for the representation of names of countries and their subdivisions - part 1: Country codes. international standard 01.140.30, International Organization for Standarization (ISO), TC 46.

[ISO, 2000] ISO (2000). ISO 8601 data elements and interchange formats - information interchange - representation of dates and times. Technical report, International Standardization Organization (ISO), December.

[ISO, 2003a] ISO (2003a). Information and documentation - the Dublin Core metadata element set. international standard ICS 35.240.30, International Organization for Standardization (ISO), TC 46/SC 4.

[ISO, 2003b] ISO (2003b). ISO 19112 - Geographic Information - spatial referencing by geographic identifiers. international standard ICS 35.240.70, International Standardization Organization (ISO), TC211, October.

[ISO, 2003c] ISO (2003c). ISO 19115 Geographic Information - Metadata. international standard ICS 35.240.70, International Organization for Standardization (ISO), 5 August.

[ISO, 2004] ISO (2004). Web site of the International Organization for Standardization. web page, International Organization for Standardization (ISO), 25 January. http://www.iso.ch.

[Jones et al., 2001] Jones, C., Alani, H., and Tudlope, D. (2001). Geographical information retrieval with ontologies of place. In Mortello, D. R., editor, *COSIT 2001*, pages 322–335, Morro Bay, California.

[Kaufman, 1991] Kaufman, S. (1991). A formal theory of spatial reasoning. In *International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 347–356.

[Kettani and Moulin, 1999] Kettani, D. and Moulin, B. (1999). A spatial model based on the notions of spatial conceptual map and object's influence areas. In Mark, C. F. and D.M., editors, *COSIT 99: Spatial information theory: cognitive and computational foundations of geographic information science*, pages 401–417. Springer.

[Killmer and Koppel, 2002] Killmer, K. A. and Koppel, N. B. (2002). So much information, so little time - evaluating web resources with search engines. *T.H.E Journal - Technological Horizons in Education*, 30(1). http://www.thejournal.com/magazine/.

[Knauff et al., 1998] Knauff, M., Rauh, R., Schlieder, C., and Strube, G. (1998). Mental models in spatial reasoning. In C. Freksa, C. H. and Wender, K., editors, *Spatial Cognition*, pages 267– 291. Springer, Berlin.

[Kraitchik, 1942] Kraitchik, M. (1942). *Mathematical Recreations*. W. W. Norton, New York.

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

[Kuhn et al., 2000] Kuhn, W., Basedow, S., Brox, C., Riedemann, C., Rossol, H., Senkler, K., and Zens, K. (2000). Geospatial Data Infrastructure (GDI) North-Rhine Westfalia - reference model 3.0. Technical report, Universität Münster, Institut für Geoinformatik (IfGI), 15 December.

[Kulik, 2003] Kulik, L. (2003). Spatial vagueness and second-order vagueness. *Spatial Cognition and Computation*, 3(2 and 3):157–183.

[Laurini and Thompson, 1992] Laurini, R. and Thompson, D. (1992). *Fundamentals of Spatial Information Systems*, volume 37 of *The A.P.I.C. Series*. Academic Press, New York.

[Legat et al., 1999] Legat, R., Hashemi-Kepp, Kruse, Nicolai, Nyhuis, Pultz, Stallbaumer, Swoboda, and Zirn (1999). Der Umweltdatenkatalog UDK in Österreich - 5 Jahre Erfahrungen. In *Proceedings of Workshop on "Umweltdatenbanken im Web"*, Karlsruhe. Forschungszentrum Informatik (FZI), Universität Karlsruhe.

[Lenz, 1994] Lenz, H. J. (1994). The conceptual schema and external schemata of metadatabases. In French, J. C. and Hinterberger, H., editors, *Seventh International Working Conference on Scientific and Statistical Database Management*. IEEE Computer Society Press.

[Lesniewski, 1916] Lesniewski, S. (1916). Foundations of the general theory of sets. *Polish Scientific Circle*.

[Lessing et al., 1995] Lessing, H., Günther, O., and Swoboda, W. (1995). An object-oriented class model for the environmental data catalogue. In Kremers, H. and Pillmann, W., editors, *Space and Time in Environmental Information Systems, 9th International Symposium on Computer Science for Environmental Applications*, Berlin. Metropolis Verlag. In German.

[Levine, 1995] Levine, M. M. (1995). A brief history of information brokering. *Bulletin of the American Society for Information Science*, 21(3).

[LOC, 2002] LOC (2002). MARC code-list for countries. web page, Library of Congress (LOC), Network Development and MARC Standards Office, 8 February. http://lcweb.loc.gov/marc/countries/.

[Maguire et al., 1991] Maguire, D. J., Goodchild, M. F., and Rhind, D., editors (1991). *Geographical Information Systems: Principles and Applications*. Longman, London, UK.

[Masolo and Vieu, 1999] Masolo, C. and Vieu, L. (1999). Atomicity vs. infinite divisibility of space. *Lecture Notes in Computer Science*, 1661:235 – 251.

[Mavrovouniotis and Stephanopoulos, 1987] Mavrovouniotis, M. L. and Stephanopoulos, G. (1987). Reasoning with orders of magnitude and approximate reations. In *National Conference on Artificial Intelligence (AAAI-87)*, pages 626–630, San Mateo, CA. Morgan Kaufmann.

[McCurley, 2001] McCurley, K. S. (2001). Geospatial mapping and navigation of the web. In *Proceedings of the tenth international conference on World Wide Web*, pages 221–229, Hong Kong, Hong Kong. ACM Press.

[Miene et al., 2003] Miene, A., Hermes, T., and Ioannidis, G. (2003). Graphical image retrieval with picturefinder. In *DELOS Workshop on Multimedia Contents in Digital Libraries*, Chania, Crete, Greece.

[Molenaar, 1998] Molenaar, M. (1998). *An Introduction to the Theory of Spatial Object Modelling*. Research Monographs in Geographical Information Systems. Taylor and Francis, London, Bristol.

[Montello, 1993] Montello, D. (1993). Scale and multiple psychologies of space. In Campari, I. and Frank, A., editors, *Spatial Information Theory: A Theoretical Basis for GIS*, volume 716 of *Lecture Notes in Computer Science*, pages 312–321. Springer, Berlin.

[Montello et al., 2003] Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P. (2003). Where's downtown ?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition and Computation*, 3(2 and 3):185–204.

[Nax and Jensen, 1999] Nax, K. and Jensen, S. (1999). ETC/CDS GEneral Multilingual Environmental Thesaurus (GEMET) - general information. electronic document, European Topic Center on Catalog of Data Sources (ETC/CDS), Niedersächsisches Umweltministerium. http://www.mu.niedersachsen.de/cds/etc-cds_neu/library/Gemet.pdf.

[Nebert and Fullton, 1995] Nebert, D. D. and Fullton, J. (1995). Use of the ISite Z39.50 software to search and retrieve spatially-referenced data. In *Proceedings of the 2nd Annual Conference on the Theory and Practice of Digital Libraries (Digital Libraries '95)*, volume online proceedings at http://csdl.tamu.edu/DL95/, Austin, Texas. Hypermedia Research Lab, Computer Science Department, Texas A&M University.

[Neumann et al., 2001] Neumann, H., Schuster, G., Stuckenschmidt, H., Visser, U., and Vögele, T. (2001). Intelligent brokering of environmental information with the BUSTER system. In Hilty, L. M. and Gilgen, P. W., editors, *International Symposium Informatics for Environmental Protection*, volume 30 of *Umwelt-Informatik Aktuell*, pages 505–512, Zürich, Switzerland. Metropolis.

[NIMA, 2004] NIMA (2004). Geonames server. online database, National Imagery and Mapping Agency (NIMA). http://www.nima.mil/gns/html/.

[OGC, 2001a] OGC (2001a). The OpenGIS Abstract Specification - Topic 11: OpenGIS Metadata (ISO/TC211 DIS 19115) - version 5. OGC Abstract Specification 01-111, OpenGIS Consortium (OGC), 8 June. http://www.opengis.org/docs/01-111.pdf.

[OGC, 2001b] OGC (2001b). The OpenGIS Abstract Specification - Topic 12: OpenGIS Service Architecture version 4.3. OGC Abstract Specification 02-112, Open GIS Consortium (OGC), 14 September. http://www.opengis.org/docs/02-112.pdf.

[OGC, 2001c] OGC (2001c). OpenGIS Implementation Specification: Coordinate Transformation Services, version 1.00. OpenGIS Implementation Specification OpenGIS Project Document Release 01-009, OpenGIS Consortium (OGC), 12 January. http://www/opengis.org/docs/01-009.pdf.

[OGC, 2002a] OGC (2002a). OpenGIS Catalog Services Specification, version 1.1.1. OpenGIS Implementation Specification 02-087r3, OpenGIS Consortium, 13 December. http://www.opengis.org/docs/02-087r3.pdf.

[OGC, 2002b] OGC (2002b). Web Feature Server (WFS) interface implementation specification, version 1.0.0. OGC Implementation Specification 02-058, OpenGIS Consortium (OGC), 17 May. http//www.opengis.org/docs/02-058.pdf.

[OGC, 2003] OGC (2003). OpenGIS Reference Model (ORM), version 0.1.2. OGC Approved Technical Baseline 03-040, OpenGIS Consortium (OGC), 4 March. http://www.opengis.org/docs/04-040.pdf.

[OGC, 2004] OGC (2004). Web site of the OpenGIS Consortium (OGC). web page, OpenGIS Consortium (OGC), 8 January. http://www.opengis.org.

[Okabe et al., 1992] Okabe, A., Boots, B., and Sugiahara, K. (1992). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley, Chichester.

[Pawlak, 1982] Pawlak, Z. (1982). Rough sets. *International Journal of Information and Computer Science*, 11(5):341–356.

[Pawlak, 1984] Pawlak, Z. (1984). Rough sets. *International Journal of Man-Machine Studies*, 21:127–134.

[Pawlak, 1991] Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Norwell, MA.

[Pawlak, 1993] Pawlak, Z. (1993). Hard and soft sets. In Alagar, V., Bergler, S., and Dong, F., editors, *Proceedings of the 3rd International Conference on Rough Sets and Soft Computing*, San Jose, CA.

[Pratt and Lemon, 1997] Pratt, I. and Lemon, O. (1997). Ontologies for plane, polygonal mereotopology. *Notre Dame Journal of Formal Logic*, 38(2):225–245.

[Pullar and Egenhofer, 1988] Pullar, D. V. and Egenhofer, M. J. (1988). Toward formal definitions of topological relations among spatial objects. In *Proceedings of the 3rd International Symposium on Spatial Data Handling (SDH'88)*, pages 225–241, Sydney, Australia.

[Randell and Cohn, 1989] Randell, D. and Cohn, A. G. (1989). Modelling topological and metrical properties of physical processes. In Brachman, R. J., Levesque, H., and Reiter, R., editors, *First International Conference on Principles of Knowledge Representation and Reasoning (KR'89)*, pages 55–66. Morgan Kaufmann.

[Randell et al., 1992] Randell, D., Cui, Z., and Cohn, A. G. (1992). A spatial logic based on regions and connection. In *3rd International Conference on Knowledge Representation and Reasoning*, pages 165–176, San Francisco. Morgan Kaufmann.

[Rauh et al., 2000] Rauh, R., Hagen, C., Schlieder, C., Strube, G., and Knauff, M. (2000). Searching for alternatives in spatial reasoning: Local transformations and beyond. In *Proceedings of the TwentySecond Annual Conference of the Cognitive Science Society*, pages 871–876, Mahwah, NJ. Lawrence Erlbaum Associates.

[Riekert, 1999] Riekert, W.-F. (1999). Erschließung von Fachinformationen im Internet mit Hilfe von Thesauri und Gazetteers. In Dade, C. and Schulz, B., editors, *Proceedings of the 2nd workshop Hypermedia im Umweltschutz - Management von Umweltinformationen in vernetzten Umgebungen*, Nürnberg.

[Samet, 1984] Samet, H. (1984). The quadtree and related hierarchical data structures. *ACM Computing Surveys*, 16:187–260.

[Samet, 1989] Samet, H. (1989). *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, MA.

[Schlieder, 1993] Schlieder, C. (1993). Representing visible locations for qualitative navigation. In Carrete, N. P. and Singh, M., editors, *3rd IMACS International Workshop on Qualitative Reasoning and Decision Technologies QUADRET'93*, pages 523–532, Barcelona. CIMNE.

[Schlieder, 1995] Schlieder, C. (1995). Reasoning about ordering. In Frank, A. and Kuhn, W., editors, *Conference on Spatial Information Theory COSIT'95*, volume 988 of *LNCS*, pages 341–349. Springer.

[Schlieder, 1996] Schlieder, C. (1996). Qualitative shape representations. In Burrough, P. and Frank, A., editors, *GISDATA Symposium on Geographical Objects with Undetermined Boundaries*. Francis Taylor.

[Schlieder, 1999] Schlieder, C. (1999). The construction of preferred mental models in reasoning with allen's relations. In Habel, C. and Rickheit, G., editors, *Mental models in discourse processing and reasoning*. Elsevier Science, Oxford.

[Schlieder and Vögele, 2002] Schlieder, C. and Vögele, T. (2002). Indexing and browsing digital maps with intelligent thumbnails. In Richardson, D. E. and Van Oosterom, P., editors, *Advances in Spatial Data Handling: Proceedings*

of the 10th International Symposium on Spatial Data Handling (SDH'02), Ottawa, Canada. Springer Verlag.

[Schlieder et al., 2001] Schlieder, C., Vögele, T., and Visser, U. (2001). Qualitative spatial representation for information retrieval by gazetteers. In Mortello, D., editor, *Spatial Information Theory: Foundations of Geographic Information Science International Conference, COSIT 2001*, volume 2205 of *LNCS*, pages 336 – 352, Morro Bay, CA, USA. Springer-Verlag.

[Schneider, 1996] Schneider, M. (1996). Modelling spatial objects with undetermined boundaries using the realm/ rose approach. In Burrough, P. and Frank, A., editors, *Geographic Objects with Indeterminate Boundaries*, volume 3 of *GISDATA Series*, pages 141–152. Taylor and Francis.

[Schneider, 1999] Schneider, M. (1999). Uncertainty management for spatial data in databases: Fuzzy spatial data types. In Güting, R., Papadias, D., and Lochovsky, F., editors, *Advances in Spatial Databases: 6th International Symposium, SSD'99*, volume 1651, pages 330–351, Hong Kong, China. Springer-Verlag.

[Schneider, 2000] Schneider, M. (2000). Finite resolution crisp and fuzzy spatial objects. In *International Symposium on Spatial Data Handling (SDH'00)*, pages 5a.3–17.

[Schneider, 2001] Schneider, M. (2001). Fuzzy topological predicates, their properties, and their integration into query languages. In Walid, G. A., editor, *Proceedings of the 9th International Symposium on Advances in Geographic Information Systems (ACM-GIS'01)*, pages 9–14, Atlanta, Georgia.

[Schneider, 2003] Schneider, M. (2003). Design and implementation of finite resolution crisp and fuzzy spatial objects. *Data & Knowledge Engineering*, 44:81–108.

[Shibasaki, 1993] Shibasaki, R. (1993). A framework for handling geometric data with positional uncertainty in a GIS environment. *GIS: Technology and Applications*, pages 21–35.

[Simons, 1987] Simons, P. (1987). *Parts - A Study in Ontology.* Clarendon Press, Oxford.

[Smith, 1995] Smith, B. (1995). On drawing lines on a map. In Frank, A. U. and Kuhn, W., editors, *Spatial Information Theory. A Theoretical basis for GIS*, pages 475–484. Springer, Berlin/Heidelberg/New York.

[Smith, 1996] Smith, B. (1996). Mereotopology: A theory of parts and boundaries. *Data and Knowledge Engineering*, 20:287–303.

[Smith, 1998] Smith, B. (1998). Basic concepts of formal ontology. In Guarino, N., editor, *Formal Ontology in Information Systems (FOIS'98)*, pages 19–28, Amsterdam. IOS Press.

[Smith and Mark, 1998] Smith, B. and Mark, D. M. (1998). Ontology and geographic kinds. In Poiker, T. and Christman, N., editors, *8th International Symposium on Spatial Data Handling (SDH'98)*, pages 308–320, Vancouver. International Geographical Union.

[Smithsonean, 2004] Smithsonean (2004). Web site of the Smithsonean Institution. web page, Smithsonean Institution, 20 January. http://www.si.edu.

[Stefanakis et al., 1996] Stefanakis, E., Vazirgiannis, M., and Sellis, T. (1996). Incorporating fuzzy logic methodologies into gis operations. In *Proceedings of the XVIII Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS)*, Vienna, Austria.

[Stell, 2000] Stell, J. G. (2000). The representation of discrete multi-resolution spatial knowledge. In *Principles of Knowledge Representation and Reasoning*, pages 38–49. Morgan Kaufmann.

[Stell, 2004] Stell, J. G. (2004). Part and complement: Fundamental concepts in spatial relations. *to appear in: Annals of Artificial Intelligence and Mathematics*.

[Stell and Worboys, 1998] Stell, J. G. and Worboys, M. F. (1998). Stratified map spaces: A formal basis for multi-resolution spatial databases. In Poiker, T., editor, *Symposium on Spatial Data Handling (SDH'98)*, Vancouver, Canada. IGU.

[Stell and Worboys, 1999] Stell, J. G. and Worboys, M. F. (1999). Generalizing graphs using amalgamation and selection. In Güting, R., Papadias, D., and Lochovsky, F., editors, *6th International Symposium on Advances in Spatial Databases (SSD'99)*, volume 1651 of *Lecture Notes in Computer Science*, pages 19–32. Springer.

[Stuckenschmidt, 2000] Stuckenschmidt, H. (2000). Using OIL for intelligent information integration. In *Workshop on Applications of Ontologies and Problem-Solving Methods at the European Conference on Artificial Intelligence ECAI 2000*, Berlin.

[Stuckenschmidt et al., 2000] Stuckenschmidt, H., Wache, H., Vögele, T., and Visser, U. (2000). Enabling technologies for interoperability. In Visser, U. and Pundt, H., editors, *Workshop on the 14th International Symposium of Computer Science for Environmental Protection*, pages 35–46, Bonn, Germany. TZI, University of Bremen.

[Swoboda et al., 2000] Swoboda, W., Kruse, F., Legat, R., Nikolai, R., and Behrens, S. (2000). Harmonisierter Zugang zu Umweltinformationen für Öffentlichkeit, Politik und Planung: Der Umweltdatenkatalog UDK im Einsatz. In Cremers, A. B. and Greve, K., editors, *14. Internationalen Symposium "Informatik für den Umweltschutz", Umweltinformatik '00*, volume 26 of *Umweltinformatik aktuell*, Bonn, Germany. Metropolis Verlag.

[TGN, 2002] TGN (2002). Getty Thesaurus of Geographic Names (TGN). online database, The Getty Research Institute. http://www.getty.edu/research/tools/vocabulary/tgn/.

[Timpf, 1999] Timpf, S. (1999). Abstraction, levels of detail, and hierarchies in map series. In Freksa, C. and Mark, D. M., editors, *Proceedings of Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science - International Conference COSIT'99*, volume 1661 of *LNCS*, pages 125–140, Stade, Germany. Springer Verlag.

[Timpf and Frank, 1997] Timpf, S. and Frank, A. (1997). Using hierarchical spatial data structures for hierarchical spatial reasoning. In Hirtle, S. C. and Frank, A., editors, *Proceedings of Spatial Information Theory - A Theoretical Basis for GIS International Conference COSIT '97*, volume 1329 of *LNCS*, pages 69–83, Laurel Highlands, Pennsylvania, USA. Springer Verlag.

[Tobler, 1970] Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(4):360–371.

[Tschangho, 1999] Tschangho (1999). Metadata for geo-spatial data sharing: A comparative analysis. *Annals of Regional Science*, Ann Reg Sci (1999)(33):171 – 181.

[Turau, 1996] Turau, V. (1996). *Alogrithmische Graphentheorie*. Addison-Weseley.

[UNGEGN, 2001] UNGEGN (2001). Consistent use of place names. electronic document, United Nations Group of Experts on Geographical Names (UNGEGN). http://www.un.org/Depts/Cartographic/english/ungegn.pdf.

[Uschold and Grueniger, 1996] Uschold, M. and Grueniger, M. (1996). Ontologies: Principles, methods and applications. *Knowledge Engineering Review*, 11(2):93–155.

[USGS, 2003] USGS (2003). Federal lands and indian reservations. digital map, U.S. Geological Survey (USGS), October. http://nationalatlas.gov/atlasftp.html.

[USGS, 2004] USGS (2004). Geographic Names Information System (GNIS). online database, U.S. Geological Survey (USGS). http://wwwnmd.usgs.gov/www/gnis/.

[van der Weide, 2001] van der Weide, T. P. (2001). Information discovery. Lecture notes on information retrieval and hypertext - part 1, Katholieke Univesiteit Nijmegen, Nijmegen Information Retrieval Group, 14. April 2001.

[van Leeuwen, 1990] van Leeuwen, J. (1990). *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity*. Elsevier and MIT Press.

[Varzi, 1994] Varzi, A. (1994). On the boundary between mereology and topology. In Casati, R., Smith, B., and White, R., editors, *Philosophy and the Cognitive Sciences: Proc. 16th Intl. Wittgenstein Symposium*, Vienna. Hölder-Pichler-Tempsky.

[Varzi, 1996] Varzi, A. (1996). Parts, wholes and part-whole relations: the prospects of mereotopology. *Data and Knowledge Engineering*, 20(3):259–286.

[Vckovski, 1998] Vckovski, A. (1998). *Interoperable and Distributed Processing in GIS*. Taylor and Francis, London.

[Vögele et al., 2003a] Vögele, T., Hübner, S., and Schuster, G. (2003a). BUSTER - an information broker for the Semantic Web. *KI - Künstliche Intelligenz*, 3:31–34.

[Vögele and Schlieder, 2002] Vögele, T. and Schlieder, C. (2002). The use of spatial metadata for information retrieval in peer-to-peer networks. In Ruiz, M., Gould, M., and Ramon, J., editors, *Proceedings of the 5th AGILE Conference on Geographic Information Science AGILE2002*, pages 279–289, Palma de Mallorca, Spain.

[Vögele and Schlieder, 2003] Vögele, T. and Schlieder, C. (2003). Spatially-aware information retrieval with place names. In Russell, I. and Haller, S. M., editors, *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference FLAIRS'03*, pages 470–474, St. Augustine, FL, USA. AAAI Press.

[Vögele et al., 2003b] Vögele, T., Schlieder, C., and Visser, U. (2003b). Intuitive modelling of place name regions for spatial information retrieval. In Kuhn, W., Worboys, M. F., and Timpf, S., editors, *Spatial Information Theory. Foundations of Geographic Information Science International Conference, COSIT 2003*, volume 2825 of *LNCS*, pages 239 – 252, Ittingen, Switzerland. Springer.

[Vögele and Spittel, 2004] Vögele, T. and Spittel, R. (2004). Enhancing spatial data infrastructures with semantic web technologies. In *to appear in: Proceedings of AGILE'04*, Heraklion, Crete.

[Vögele et al., 2003c] Vögele, T., Spittel, R., Visser, U., and Hübner, S. (2003c). Geoshare - building a transnational geodata infrastructure for the north sea region. In Bernard, L., Sliwinski, A., and Senkler, K., editors, *Proceedings of the 2nd Münsteraner GI Tage - Geodaten und Geodienste Infrastrukturen - von der Forschung zur praktischen Anwendung*, volume 18, pages 1–15, Münster, Germany. IfGI prints, Institut für Geoinformatik der Universität Münster.

[Visser et al., 2001] Visser, U., Stuckenschmidt, H., Wache, H., and Vögele, T. (2001). Using environmental information efficiently: Sharing data and knowledge from heterogeneous sources. In Rautenstrauch, C. and Patig, S., editors, *Environmental Information Systems in Industry and Public Administration*, pages 41–73. IDEA Group, Hershey, USA & London, UK.

[Visser et al., 2002] Visser, U., Vögele, T., and Schlieder, C. (2002). Spatio-terminological information retrieval using the BUSTER system. In Pillmann, W. and Tochtermann, K., editors, *Proceedings of the 16th conference "Environmental Informatics 2002" (EnviroInfo'2002)*, Environmental Communication in the Information Society, pages 93–100, Vienna, Austria. Berger Druck, Horn, Austria.

[Voronoi, 1908] Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques, deuxième memoire, recherches sur les parallelloèdres primitifs. *Journal für die Reine und Angewandte Mathematik*, 134:198–287.

[W3C, 2000] W3C (2000). Extensible Markup Language (XML) 1.0 (second edition). W3c recommendation, World Wide Web Consortium (W3C), 6 October. http://www.w3.org/TR/2000/REC-xml-20001006.

[Wache et al., 2001] Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In *IJCAI 2001 Workshop on Ontologies and Information Sharing*, pages 108–117, Seattle, WA.

[Webster, 2001] Webster (2001). *New International Webster's Pocket Computer Dictionary of the English Language*. Trident Press International.

[Whitehead, 1978] Whitehead, A. (1978). *Process and Reality: Corrected Edition*. The Free Press, MacMillan, New York.

[Wolf and Wicksteed, 1998] Wolf, M. and Wicksteed, C. (1998). W3C-DTF World Wide Web Consortium: Date and time formats. W3c note, World Wide Web Consortium (W3C), 27 August. http://www.w3.org/TR/1998/NOTE-datetime-19980827.

[Worboys, 1995] Worboys, M. F. (1995). *GIS - A Computing Perspective*. Taylor and Francis, London, Philadelphia.

[Worboys, 1998] Worboys, M. F. (1998). Imprecision in finite resolution spatial data. *GeoInformatica*, 2:257–279.

[Worboys and Bofakos, 1993] Worboys, M. F. and Bofakos, P. (1993). A canonical model for a class of real spatial objects. In Abel, D. and Ooi, B., editors, *Proceedings of the 3rd International Symposium on Advances in Spatial Databases (SSD'93)*, pages 36–52. Springer Verlag.

[Zadeh, 1965] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

[Zeiler, 1999] Zeiler, M. (1999). *Modeling Our World - The ESRI Guide to Geodatabase Design*. Environmental Systems Research Institute (ESRI), Redlands, CA.

[Zimmermann, 1995] Zimmermann, K. (1995). Measuring without distances: The delta calculus. In Frank, A. and Kuhn, W., editors, *Spatial Information Theory: A theoretical basis for GIS, COSIT'95*, volume 988 of *LNCS*, pages 59–68. Springer.

[Zimmermann and Freksa, 1996] Zimmermann, K. and Freksa, C. (1996). Qualitative spatial reasoning using orientation, distance, and path knowledge. *Applied Intelligence*, 6:49–58.

Part IV

Appendices

Appendix A

XML Encoding of a Qualitative Spatial Reference Model

```
<?xml version="1.0" encoding="UTF-8"?> <refUnits
    xmlns='http://www.tzi.de/buster'
    xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
    xsi:schemaLocation='http://www.tzi.de/buster http://www-agki.tzi.de/buster/data/xsd/spatial_model.xsd'
    signature="http://www.tzi.de/buster/data/ontologies/spat/DEnuts-ru.xml"
    name="DE Nuts">
    <taxonomy cover="L3">
        <level name="L0"/>
        <level name="L1"/>
        <level name="L2"/>
        <level name="L3"/>
    </taxonomy>
    <reference-units>
        <refUnit id="RDE" name="Deutschland" type="L0"/>
        <refUnit id="RDE2" name="Bayern" type="L1"/>
        <refUnit id="RDE1" name="Baden-Wuerttemberg" type="L1"/>
        ...
        <refUnit id="RDE13" name="Freiburg" type="L2"/>
        <refUnit id="RDE12" name="Karlsruhe" type="L2"/>
        <refUnit id="RDE11" name="Stuttgart" type="L2"/>
        ...
        <refUnit id="RDE11A" name="SCHWAEBISCH HALL" type="L3"/>
        <refUnit id="RDE118" name="HEILBRONN" type="L3"/>
    </reference-units>
    <relations>
        <part-of>
            <refUnit id="RDE">
                <refUnit id="RDE1">
                    <refUnit id="RDE13">
                        <refUnit id="RDE134"/>
                        ...
                    </refUnit>
                    <refUnit id="RDE11">
                        <refUnit id="RDE11A"/>
                        <refUnit id="RDE118"/>
                        <refUnit id="RDE117"/>
                        ...
                    </refUnit>
                    ...
                </refUnit>
                ...
            </refUnit>
            ...
        </part-of>

        <neighbors>
            <refUnit id="RDE1">
                <refUnit id="RDE2"/>
                ...
                </refUnit>
            <refUnit id="RDE13">
                <refUnit id="RDE12"/>
                ...
            </refUnit>
            <refUnit id="RDE134">
                <refUnit id="RDE133"/>
                <refUnit id="RDE136"/>
                ...
            </refUnit>
        </neighbors>
    </relations>
</refUnits>
```

Figure A.1: Example for spatial reference model encoded in *XML*

Appendix B

XML Encoding of a Place Name Structure

```
<?xml version="1.0"?> <placeNames
   xmlns="http://www.tzi.de/buster"
   xmlns:xi="http://www.w3.org/2001/XInclude"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.tzi.de/buster http://www-agki.tzi.de/
   buster/data/xsd/spatial_model.xsd"
   signature="http://www.tzi.de/buster/data/ontologies/spat/Weserbergland-pn.xml"
   name="Weserbergland">
   <dependencies>
       <xi:include href="DEnuts-ru.xml"/>
   </dependencies>
   <regions>
       <placeName id="PDL3" name="Weserbergland"/>
       <placeName id="PDL31" name="Porta Westfalica"/>
       <placeName id="PDL32" name="Wesergebirge"/>
       ...
   </regions>
   <relations>
       <partonomy>
           <placeName id="PDL3">
               <placeName id="PDL31"/>
               <placeName id="PDL32"/>
               ...
           </placeName>
           ...
       </partonomy>
       <spatial_footprint>
           <placeName id="PDL3">
               <refUnit id="RDE928"/>
               <refUnit id="RDE923"/>
               ...
           </placeName>
           <placeName id="PDL32">
               <refUnit id="RDEA46"/>
               ...
           </placeName>
           ...
       </spatial_footprint>
   </relations>
</placeNames>
```

Figure B.1: Example for an XML-encoded place name structure in BUSTER