Indexing and Browsing Digital Maps with Intelligent Thumbnails

Christoph Schlieder and Thomas Vögele

Centre for Computing Technologies (TZI), University of Bremen, PO Box 33 04 40, 28334 Bremen, Germany,{cs|vogele@tzi.de}

Abstract

With the increasing number of digital maps and other geo-referenced data that are available over the Internet, there is a growing need for access to techniques that allow us to preview the content and to evaluate it relative to the requirements for complex spatial and thematic queries. . Analogous to the digital indices for full-text searches on text documents, we introduce highly condensed, machine-readable indices of digital maps. The purpose of these "intelligent thumbnails" is to support sophisticated queries of the type concept@location. An intelligent thumbnail is based on a projection of the thematic content of a digital map onto a standard reference tessellation. To make the thumbnail exchangeable in a distributed and heterogeneous environment, the underlying Standard Reference Tessellation (SRT) is qualitatively abstracted from a polygonal representation to a connection graph. In combination with a hierarchical place name structure and concept ontology, it is used to evaluate the spatial and thematic relevance of the indexed data sources with respect to spatio-thematic queries.

Keywords: digital maps, metadata, spatial queries, data reduction, spatial reasoning

1 Introduction

Digital maps and other geospatial data are the valuable assets of many private and public organisations. For that reason, more and more geospatial data are made available through online (meta)data catalogues and Internet portals. To a potential user or buyer, metadata catalogues offer browse services of available data sources, and allow an individual to select datasets best suited for a specific task. The selection process should be based on queries of the type concept@location, i.e. an evaluation of which data sets contain information about a specific thematic concept at a specific geographic location.

Typically, metadata catalogues provide tools to select and rank the available data sets based on their semantic relevance with respect to certain thematic concepts. Most state-of-the-art systems still use simple keyword matches. Considerable research effort has been made, however, for the development of more sophisticated terminological queries, for example with the help of formalised domain ontologies and terminological reasoning (Wache et al., 2001).

On the other hand, the mechanisms available to solve spatial queries and to evaluate the spatial relevance of a data source are still fairly simplistic. Typically, they are based on a match between a bounding box representation of the area covered by a data source, and in most cases, a rectangular search area. All data sets for which the corresponding bounding box intersects with the search area are assumed spatially relevant with respect to the query. Even if the general shortcomings of bounding box reasoning are neglected, the fact that a digital map covers an area delimited by its bounding box does not necessarily imply that the map actually contains information about the concept in question at every location within this area.

For example, the digital map of *Federal Land Features of the United States* published by the U.S. Geological Survey (USGS) (USGS, 2000) 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.

For a metadata description of the data set, the USGS uses the Content Standard for Digital Geospatial Metadata published by the Federal Geographic Data Committee (FGDC) (FGDC, 1994), and expresses the spatial coverage of the map in terms of a bounding box. As a result, somebody interested in, for example, the outlines of *National Parks* in *Contra Costa County, California*, will be pointed to the *Federal Land Features* map because a) it contains information about national parks, and b) it covers the whole contiguous United States, which *Contra Costa County* is a part of. After a time-consuming download (the data set exceeds 50 MB), a detailed analysis of the data with the help of a geographic information system (GIS) would reveal that there are in fact no *National Parks* in *Contra Costa County*, making the data set unsuitable for the intended purpose.

The example above shows, that metadata for geospatial data often do not provide enough information to make educated decisions about the usability of data sets for specific tasks. Obviously, there is a need for tools that support more sophisticated screening-level queries, which do not require the download of large data files and the application of complex GISs. Some GISs, like for example ESRIs ArcGIS 8.x, therefore introduced data management tools like ArcCatalog (ESRI, 2001). In ArcCatalog, small raster images of digital maps, the so-called “thumbnails”, can be created to provide a visual preview of the data.

We argue, that these purely visual indices do not suffice to support sophisticated spatial queries. Firstly, they can only represent a single thematic concept at a given location, and secondly they have to be analysed and evaluated “manually” by a human, thus providing no means for effective automatic searches. We therefore propose to create “intelligent thumbnails”, i.e. small, machine-readable indices of the thematic and spatial content of digital maps, analogous to the thematic indices used for full text searches in digital text documents.

2 Intelligent Thumbnails

2.1 Components of an Intelligent Thumbnail

An intelligent thumbnail is a machine-readable and highly condensed index of the thematic and spatial information-contents of a digital map. It is designed to support reasoning about whether the map is relevant with respect to a specific query of the type *concept@location*. The level of relevance assigned to a map depends on how closely the content of the map matches the query. This takes into account direct matches of the specified *concept* and *location*, but also near matches, which result from terminological specialisation or generalisation of the concept, as well as spatial generalisations of the location.
The index references the thematic and spatial contents of a data source (i.e. a digital map) to both thematic domain ontology and a spatial reference model; see Fig. 1. The spatial reference model consists of a standard reference tessellation (SRT) and a place name structure.

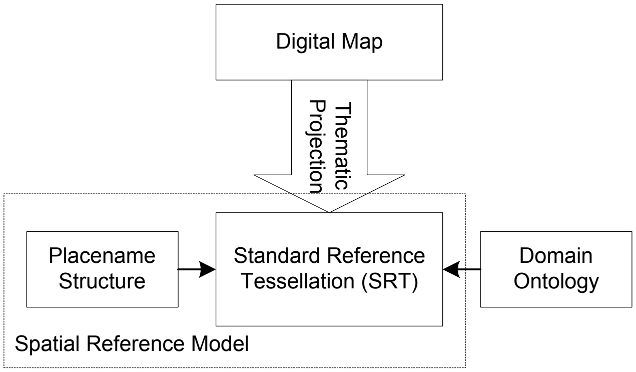

Fig. 1. Components of an Intelligent Thumbnail

2.1.1 Thematic Projection onto a Reference Tessellation

An intelligent thumbnail links concepts to locations. In this context, the term “concept” refers to the thematic attributes of the features represented in the digital map and the term “location” to geographic entities represented in a spatial reference model. The thumbnail is created through a projection of the thematic layers of a digital map onto a spatial reference model (Fig. 1).

The spatial reference model is based on a polygonal tessellation of spatial entities with a coverage equal to or larger than the extent of the digital map to be indexed. To make the indices of different digital maps comparable, the spatial reference model has to be standardised; i.e. it has to use a well-defined set of spatial entities and descriptors. In addition, to allow for complex spatial queries as described above, the spatial reference model has to support reasoning about the topological and partonomic relations between spatial entities.

One option to build a standardised spatial reference model is to use a uniform grid. This approach is applied by some of the more advanced gazetteer services such as the gazetteer integrated in the German Environmental Information Network (GEIN) (Riekert, 1999), (Bilo et al., 2000). Gazetteers are place name-lists that link the names of geographic features to geographic footprint representations (Hill, 2000). The projection of the polygonal footprint of a geographic feature onto a regular, homogeneous grid is depicted in Fig 2. Projecting geographic footprints onto reference grids allows us to apply spatial reasoning in topologically related ways. For example, if the grid cells occupied by object A are also occupied by object B, it can be inferred that object A is contained in object B. Other topological relations like disjunction, connection, and overlap, can also be evaluated using this approach.

Instead of uniform grids, we propose to use polygonal tessellations as the basis for spatial reference models. Maps of postal code areas, administrative units, and census districts are examples of such Standard Reference Tessellations (SRTs). Polygonal SRTs have a number of advantages over uniform grids:

- Many polygonal SRTs represent well-known and officially named spatial entities, which a user can relate to more easily than to arbitrarily created and cryptically named grid cell rasters. For example, it is much easier to refer to the place name *Contra Costa County* than to a grid cell descriptor like *CA1089*.
- From a user perspective, polygonal SRTs are a conceptually more logical way in which to organise spatially distributed data. Many companies, for example, arrange their marketing areas along the lines of postal code areas or other popular reference tessellations. As a result, polygonal SRTs are available in many organisations in digital form, including GIS data formats.
- Administrative units and other SRTs are typically associated with a hierarchical partonomic structure. Each state in the U.S., for example, consists of a number of counties which consists of a number of communities. As we will see in section 2.2.1, the evaluation of such hierarchical partonomies can be part of a metric to compute spatial closeness in an attempt to evaluate the spatial and thematic relevance of a data set.

As we pointed out above, an intelligent thumbnail is created by mapping the thematic layers of a digital map onto the a polygonal SRT; (see Fig. 2) This can be achieved effectively within a 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 SRT (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 SRT polygons.

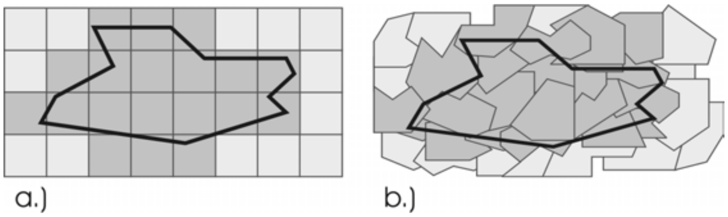

Fig. 2(a) Thematic Mapping onto a Regular Grid, and **(b)** a Polygonal Standard Reference Tessellation (SRT)

2.1.2 Qualitative Spatial Models

In the previous section we argued, that polygonal tessellations representing well-known spatial units, like administrative boundaries and postal code areas, provide better and more intuitive spatial reference models for intelligent thumbnails than regular grids. Polygonal SRTs do, however, have some disadvantages as well. One disadvantage is that different polygonal representations of the same SRT may not be identical in terms of an exact match of vertex co-ordinates. Such mismatches can have a number of reasons, like different levels of accuracy and spatial resolution, or different coordinate systems and projections. As a result, each data provider should or must use the same, “officially approved” polygonal SRT to create a meaningful intelligent thumbnail.

The second disadvantage is, that by using polygonal SRTs in the form of GIS data sets as a basis for the search through intelligent thumbnails, the spatial reasoning needed to resolve spatial queries has to rely heavily on GIS functionality. This requires the handling of potentially large volumes of GIS data and the availability of complex GIS software even for a browsing-level information retrieval. Especially in highly heterogeneous and distributed data-exchange infrastructures, this may hamper the flow of information and limit the number of participants (Vögele et al., 2002).

We argue, that for the spatial reasoning tasks needed in the context of information retrieval, most of the spatial information-content of polygonal GIS data is superfluous. In order to make the information retrieval independent of potential GIS-related inconsistencies and complexities, and to create light-weight, exchangeable intelligent thumbnails, we propose to use qualitative spatial reference models. Such models are based on a combination of qualitative abstractions of polygonal spatial reference tessellations and take the form of connection graphs, and hierarchical place name structures.

2.1.3 Connection Graphs

We use the concept of connection graphs as it was defined in (Schlieder et al., 2001). Connection graphs are an extension of the well-known neighbourhood graphs (Molenaar, 1998) such that the topological neighbourhood relations of polygons in an homogeneous tessellation are encoded, with their ordering, and, if applicable with the connection to an external area. Technically speaking, the connection graph consists of the dual of the tessellation together with the combinatorial embedding of the dual. Connection graphs can be used to encode standard reference tessellations as part of qualitative spatial models. Although in this paper we focus on the encoding of topological (neighbourhood) relations, connection graphs may also be used to represent ordinal and distance relations (Stuckenschmidt et al., 1999).

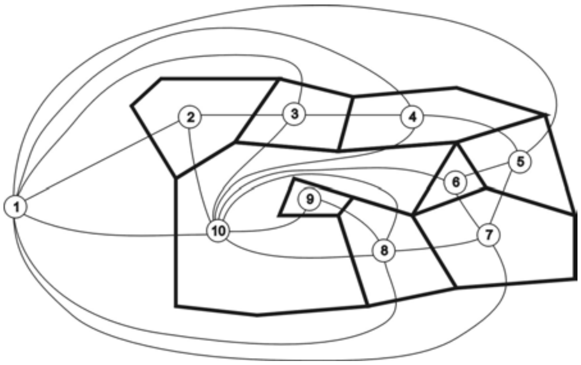

Fig. 3. Connection Graph G_c of Polygonal Tessellation D

Fig. 3 shows the connection graph G_c obtained by homogeneous decomposition of a polygonal tessellation D. Each polygon from D is represented by a vertex from G_c. In addition, there is the node 1 representing the external polygonal region. The edges from G_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 neighbours 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.

2.1 4 Place Name Structures

Place name structures link the entities of standard reference tessellations (SRTs) to place names, which are the basis for intuitive and user-friendly spatial queries. A place name structure can be seen as an hierarchical tree, where the nodes of the tree represent well-known name descriptors for geographic features, and the edges reflect the binary part-of relations between these features. In a qualitative spatial model, the leaves of the tree coincide with the nodes of the connection graph representing the SRT (**Fig. 4**. Hierarchical Tree of a Place name Structure).

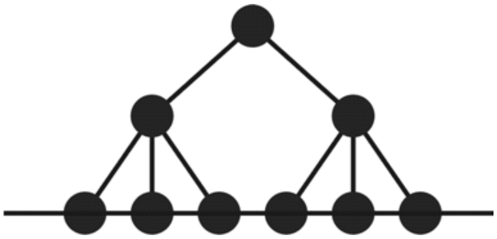

Fig. 4. Hierarchical Tree of a Place name Structure

For many SRTs, place name structures are readily available: Administrative units, for example, are typically represented by hierarchical trees that describe the part-of relations between entities like countries, states, counties, communities etc. Used as a framework for qualitative spatial reference models, such standard place name structures provide a common vocabulary for spatial references. They ensure that all spatial references and spatial queries in an intelligent thumbnail can ultimately be mapped onto the same, unambiguous and consistent set of place names.

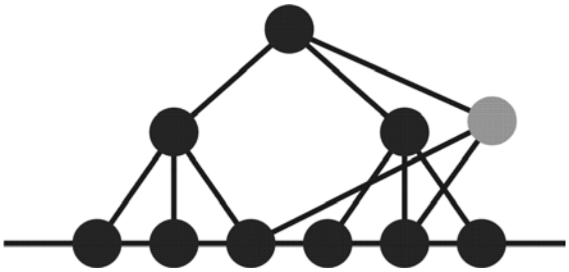

Fig. 5. Poly-Hierarchy of an Extended Place name Structure

However, spatial reference models and standard place name structures have to be extensible in order to allow for the incorporation of “colloquial” place names, i.e. commonly used descriptors for geographic features, as well as user-specific terms. For example, the place name *East-Bay* is a popular term to describe a number of counties situated on the eastern side of the *San Francisco Bay*. To solve a typical user-driven query like “Are there any lands managed by the National Park Service in the *East-Bay*?”, the system has to have access to a place name structure that formalises the spatial semantics of the term *East-Bay*.

A qualitative spatial model of the administrative units of *California* can be extended to “understand” the meaning of the term *East-Bay* by establishing the respective part-of relations based on an existing place name structure. The result is a poly-hierarchic, Directed Acyclic Graph (DAG) representing an extended, user-specific place name structure (Fig. 5). Because qualitative spatial models do not rely on GIS functionality and complex binary data formats, it should be easy to provide simple tools that support the manipulation of place names in a user-friendly way. This is the basis for highly distributed data exchange infrastructures, where users may modify standard place name structures to match their individual needs and use them to specify highly personalised spatial queries (Vögele et al., 2002).

2.1.5 Domain Ontologies

A domain ontology is the basis for the evaluation of the semantic connection between thematic concepts specified in the query and concepts indexed in the intelligent thumbnail. For example, in a query like *NPS-lands @ Contra Costa County*, we could use a domain ontology representing the organisational structure of US federal agencies and federal lands to find out which types of federal lands are managed by the *National Park Service (NPS)* (Fig. 6).

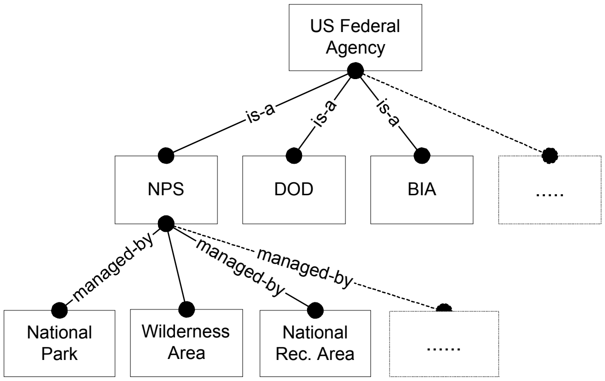

Fig. 6. Ontology of US federal agencies and federal lands (schematic)

This paper is focused on the solution of the *location*-part of a query of type *concept@location* using spatial reasoning. For brevity, however, we will not explore the representation of domain ontologies and terminological reasoning. There are a number of approaches and systems available for this task (Wache et al., 2001). For the prototypical implementation of the intelligent thumbnail, we used description-logics based on the knowledge representation language OIL (Ontology Interchange Language) (Fensel et al., 2000; Stuckenschmidt, 2000), in conjunction with the RACER theorem prover (Haarslev et al., 2001).

2.2 Reasoning about Spatial and Thematic Relevance

With the help of concepts described above, queries of type *concept@location*, have been specified and resolved. In addition, the data sets retrieved have been ranked based on their spatial and thematic relevance with respect to the query. The basis for such evaluations are metrics that express the degree of spatial and terminological distance between locations and concepts represented in the thumbnail, and locations and concepts specified in the query.

2.2.1 Metrics for Spatial Distance

In an intelligent thumbnail, metrics for spatial distance can be computed based on the connection graph representation of the SRT and the DAG representation of the place name structures, respectively. In the simplest case, spatial closeness can be expressed using a metric based on graph-theoretical node distances.

In the connection graph, the Euclidian distance ν between two nodes N1 and N2 is a measure for the actual spatial proximity of the two areas represented by these nodes. In the DAG, the distance δ between the same nodes N1 and N2 indicates their degree of separation with respect to a hierarchical partonomy. The total distance, or spatial relevance measure, D(N1,N2) between N1 and N2 is obtained by a linear combination of ν and δ:

$$D(N1,N2) = \alpha\, \nu(N1,N2) + (1-\alpha)\, \delta(N1,N2),$$

By manipulating the weighting factor α, a spatial query may be fine-tuned to favour either locations that are spatially close ($\alpha = 1$) to the place name of interest, or locations that belong to the same part of a hierarchical partonomy ($\alpha = 0$). The value of α depends mostly on the intent of the query and the semantics of the concepts in question. In the case of federal lands mapped on administrative units, for example, hierarchical closeness is less important than spatial proximity. This is the case, since the indexed concepts (i.e. the federal agencies managing the lands) do not change between the administrative partonomy units. In the case of lands managed by state agencies, however, hierarchical distance could be more important because the same concepts (i.e. state parks) have different properties (i.e. are managed by different agencies) in different units of the partonomy (i.e. different states).

In the prototypical implementation of the intelligent thumbnail, α is set to a default value of 0.5, resulting in an equal weighting of spatial and hierarchical proximity. It is up to the user to change α and fine-tune the query. However, concept-specific default values for α could also be included in the terminological domain ontology, allowing for an automatic fine-tuning of queries.

2.2.2 Metrics for Thematic Distance

To evaluate the thematic or semantic distance between two concepts, both concepts have to be part of a formalised concept hierarchy, or ontology. As described in section 2.1.5, we use the knowledge representation language OIL to encode concept ontologies, and a theorem prover based on description logics to reason about the semantic distance σ between two concepts C_q and C_t. 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 σ 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. However, as we focus on spatial queries in this paper, we will not discuss metrics for thematic distance in more detail.

2.2.3 Combined Metric for Spatial and Thematic Distance

The result of a 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. As a first approach, we use a spatial relevance metric $R_{S,}$ which is a linear combination of the total spatial distance D and the semantic distance σ:

$$R_S = \beta\, D + (1-\beta)\, \sigma$$

Again, the weighting factor β can be used to bias the query towards spatial relevance (β=1), or thematic relevance (β=0).

2.3 Data Reduction

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 thumbnail size and underlying qualitative spatial reference model as small as possible. Further, open, non-proprietary data formats should be used 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 generalisation measures must be applied:

1. The first generalisation 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 generalisation 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 abstracted as *land managed by the NPS*.
3. Maybe the largest generalisation and data reduction takes place during the thematic projection. All selected feature types in a digital map are projected

onto a standard reference tessellations, but if two identical concepts happen to map onto the same spatial entity, only one reference is maintained. The concept *National Historic Site*, for example, would be assigned only once to *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).
5. In addition to data reduction, generalising polygonal standard reference tessellations as connection graphs help to avoid the typical problems of heterogeneous GIS data, like incompatible discretizations, proprietary data formats, and different coordinate-systems and map projections.

3 Summary and Discussion

In this paper, we describe intelligent thumbnails and qualitative spatial models as the basic components to create machine-readable indices of the spatial and thematic contents of digital maps. Such indices are necessary to support queries of the type concept@location, i.e. to select from a large number of data sources the ones that are relevant with respect to the thematic and spatial criteria specified in a query.

Although it is the ultimate goal of our work to combine the evaluation of thematic and spatial relevance, this paper was focused mainly on the representation and reasoning about spatial concepts. The reason for this bias is the fact, that while considerable research was concerned with thematic and terminological knowledge representation and reasoning in the last few years, comparatively little attention was given to the development of qualitative spatial representations and reasoning within the geographical data application domain.

Jones (Jones et al., 2001) recently proposed the use of “parsimonious” spatial models for geographical information retrieval. Similar to the intelligent thumbnail, his system uses metrics for hierarchical, spatial (euclidean), and semantic distance to come up with an overall score for geo-referenced data objects in a database. In addition, in the area of place name structures, Hill (Hill, 2000) is working on a data standard for gazetteer services.

Many aspects of the digital thumbnails presented in this paper are still based on simple assumptions and methods. One example is the metrics for spatial and thematic distance evaluations. Our future work will be concerned with refining these assumptions and incorporating more sophisticated methods into the intelligent thumbnail.

References

Bilo M, Streuff H (2000) Das Umweltinformationsnetz Deutschland - GEIN2000 - Fachliche Anforderungen an ein Forschungs- und Entwicklungsvorhaben. 3rd workshop Hypermedia im Umweltschutz, Ulm

ESRI (2001) System Design Strategies. Environmental Systems Research Institute, Inc.

Fensel D, Horrocks I, Harmelen F V, Decker S, Erdmann M, Klein M (2000) OIL in a Nutshell. In: 12th International Conference on Knowledge Engineering and Knowledge Management EKAW 2000. Juan-les-Pins, France.

FGDC (1994) Content Standards for Digital Geospatial Metadata [online]. Washington, D.C., US Government, Federal Geographic Data Committee. Available from: ftp://fgdc.er.usgs.gov.

Haarslev V, Möller R (2001) RACER System Description. University of Hamburg, Hamburg, Germany

Hill LL (2000) Core elements of digital gazetteers: place names, categories, and footprints [online]. In: ECDL 2000, Lisbon, Portugal. Available from: http://www.alexandria.ucsb.edu/~lhill/paper_drafts/ECDL2000_paperdraft7.pdf.

Jones C, Alani H, Tudlope D (2001) Geographical Information Retrieval with Ontologies of Place. In: COSIT 2001. Morro Bay, California

Molenaar M (1998) An Introduction to the Theory of Spatial Object Modelling. Taylor & Francis, London Bristol

Riekert W-F (1999) Erschließung von Fachinformationen im Internet mit Hilfe von Thesauri und Gazetteers. In: Management von Umweltinformationen in vernetzten Umgebungen, 2nd workshop HMI. Nürnberg

Schlieder C, Vögele T, Visser U (2001) Qualitative Spatial Reasoning for Information Retrieval by Gazetteers. In: COSIT'01. Morro Bay.

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 H, Visser U, Schuster G, Vögele T (1999) Ontologies for Geographic Information Integration. In: Workshop Intelligent Methods in Environmental Protection: Special Aspects of Processing in Space and Time, 13. International Symposium of Computer Science for Environmental Protection (CSEP '99). University of Bremen

USGS (2000) Federal Land Features of the United States [online]. Reston, VA, U.S. Geological Survey. Available from: http://nationalatlas.gov/atlasftp.html.

Vögele T and Schlieder C (2002) The Use of Spatial Metadata for Information Retrieval in Peer-to-Peer Networks. Proceedings of AGILE2002, Palma de Mallorca, Spain (in press)

Wache H, Vögele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, Hübner S (2001) Ontology-Based Integration of Information - A Survey of Existing Approaches. In: IJCAI 2001.