The Role of Spatial Relations in Automating the Semantic Annotation of Geodata

Eva Klien and Michael Lutz

Institut für Geoinformatik (IfGI)
Westfälische Wilhelms-Universität Münster
Robert-Koch-Str. 26-28, 48149 Münster, Germany
{klien|m.lutz}@uni-muenster.de

Abstract. How can the usability of distributed and heterogeneous geographic data sets be enhanced? Semantic interoperability is a prerequisite for effectively finding and accessing relevant data in different application contexts. By using geospatial domain ontologies and semantic annotations of geodata based on these ontologies semantic interoperability can be achieved. However, since no automated methods for the semantic annotation of geodata exist this remains a laborious task, which data providers are neither willing nor capable to perform. In this paper we propose a method for automating the annotation process based on spatial relations. At the domain level, spatial relations play an important role for defining and identifying geospatial concepts. At the data level, spatial relations may be expressed through spatial processing methods, as we can calculate relations like topology, direction or distance between two spatial entities. We show how this potential can be exploited for automating the semantic annotation of geodata. The approach is illustrated by introducing a case study for annotating data containing representations of floodplains.

1 Introduction

Distributed and heterogeneous geographic data sets have a great potential for applications ranging from environmental planning to emergency management or e-commerce. However, even though syntactical standards for Spatial Data Infrastructures (SDIs) already enable the retrieval and multiple exploitation of geodata (cf. [1]), still many problems impede efficient usability. Being able to assess semantic interoperability is a precondition for effectively finding and accessing relevant data in different application contexts. One of the shortcomings of current SDIs is the missing support for this assessment.

An important means for achieving semantic interoperability are *ontologies*, which capture consensual knowledge and formalize this knowledge in a machine-interpretable way [2]. In SDIs, ontologies can be employed for making the semantics of the information content of geospatial web services explicit. In [3, 4] we have shown how ontologies can be used to realise semantic matchmaking during service discovery and retrieval. The backbone of our approach is an infrastructure of geospatial domain ontologies and semantic annotations of the geodata. Domain ontologies represent the basic concepts and relations to which all members of an information

community commit. They provide the foundation on which the geodata is semantically annotated. The common commitment ensures semantic interoperability [5].

So far, no automated method for the semantic annotation of geodata exists. Manual annotation is difficult, time consuming, and expensive and data providers who are no ontology engineering specialists will be neither willing nor capable to perform it. We propose a method for automating the annotation process that relies on the specific characteristics of geographic information.

Spatial relations between entities are characteristic for geographic information and they are often as important as the entities themselves [6]. In geospatial domain ontologies, taxonomic and non-taxonomic relations are used to define concepts of the physical world and to differentiate between them. At this level, spatial relations play an important role for defining and identifying spatial concepts, but when reasoning about these concepts, spatial relations are not treated differently from other non-taxonomic relations. At the data level though, the spatial relations may be expressed through spatial processing methods, as we can calculate e.g. the topology, direction or distance between two spatial entities. In this paper we illustrate how this potential can be exploited for the semi-automatic semantic annotation of geodata.

In this paper we focus on spatial *relations*. In our future work, the approach will be extended to include spatial *attributes* and non-spatial characteristics. The approach is illustrated by introducing an example ("annotating data containing floodplains") and consists of the following steps:

- Extract all concept definitions from the geospatial domain ontology that contain spatial relations.
- Translate spatial relations (e.g. *adjacent to*) into a corresponding spatial analysis method, which is implemented as a sequence of GIS operations.
- Apply these spatial analyses on the geodataset to be annotated.
- Identify sets of spatial entities that share a characteristic set of relations and can then be referenced to the corresponding geospatial concept.

The remainder of the paper is structured as follows. We first introduce a motivating example to clarify in what context we refer to semantic annotation and why we think a method for automating the process is needed (section 2). In section 3, spatial relations are discussed with respect to their role in the semantic annotation process. In section 4, we illustrate the general idea by conducting a walk-through the annotation process and continue with explaining the proposed method in detail. Section 5 provides an overview on related work. Finally, we discuss the approach and identify some future work (section 6).

2 Motivation: Ontology-Based Discovery and Retrieval of GI

Ontologies can be applied for making the semantics of the information content of geospatial web services explicit in order to enhance geographic information (GI) discovery and retrieval in geospatial web service environments [4].

In the following we introduce an example to illustrate our approach for solving semantic heterogeneity problems in SDIs and to describe the semantic matchmaking mechanism which underlies the ontology-based discovery and retrieval.

2.1 Example "Floodplain"

Floodplains are crucial elements for the task of flood management. They serve as a natural water retention area after a river broke its banks during a flooding event. If sufficient area along the river banks has the function of a floodplain, some of the river's water load will "naturally" be absorbed and the flooding event will be less critical for populated areas that lie further downstream.

Floodplains can be looked at from several different perspectives: "To define a floodplain depends somewhat on the goals in mind. As a topographic category it is quite flat and lies adjacent to a stream; geomorphologically, it is a landform composed primarily of unconsolidated depositional material derived from sediments being transported by the related stream; hydrologically, it is best defined as a landform subject to periodic flooding by a parent stream. A combination of these [characteristics] perhaps comprises the essential criteria for defining the floodplain"[7].

We will use this definition for formalizing the concept of a *floodplain* in our geospatial domain ontology. It is important to note, that the relation *lies adjacent to* is interpreted in the sense of "near or close to but not necessarily touching" (WordNet 2.0[1]).

2.2 Semantic Heterogeneity Problems

In order to avoid future flooding disasters, the planning department of a city council has decided to identify potential areas in the district that may be re-designated as floodplains. The task of John, the planner in charge, is a) first to find data that contains the relevant information (*discovery*) and b) to access this data and retrieve the information (*retrieval*).

In current standards-based catalogues users can formulate queries using keywords and/or spatial filters. The metadata fields that can be included in the query depend on the metadata schema used (e.g. ISO 19115) and on the query functionality of the service that is used for accessing the metadata. Even though natural language processing techniques (e.g. [8]) can increase the semantic relevance of search results with respect to the search request, keyword-based techniques are inherently restricted by the ambiguities of natural language. As a result, keyword-based search can have low recall if different terminology is used and/or low precision if terms are homonymous or because of their limited possibilities to express complex queries [9].

For example, if John uses "floodplains" as a keyword he may fail to find existing Web Feature Services (WFS) that offer information on floodplains, because their metadata description uses a different terminology. Furthermore, he might also discover data sources that are annotated with this keyword but not appropriate for answering his purposes, e.g. a service providing areas that are officially appointed and protected for having the function of a floodplain according to national legislation. Another obstacle often encountered is missing metadata entries. In that case a successful search will not be possible at all.

[1] http://wordnet.princeton.edu/cgi-bin/webwn/

Once John has discovered a dataset and wants to access it via its WFS interface, he faces yet another major difficulty. The *DescribeFeatureType* request [10] returns the application schema for the feature type, which is essential for formulating a query filter. John now runs into trouble if the property names are not intuitively interpretable or if the feature type "floodplain" is not explicitly stored in the schema. In our example, it might be sufficient to offer John a natural language description for each property. However, our work is aiming at automating the process of discovery and retrieval and this makes a machine-interpretable description of the properties indispensable.

2.3 Semantic Matchmaking

Fig. 1 illustrates the matchmaking which underlies ontology-based discovery and retrieval. The geospatial ontology contains the basic terms of a domain (e.g. geomorphology). It is assumed that all actors within a domain share a common understanding of the concepts and relations provided at the domain level [5]. The information sources, i.e. the geodata are annotated based on the concepts and relations provided in the geospatial ontologies. In our example the information source is a geodataset that contains polygons with land use attributes. John, the user of geospatial web services, is looking for information sources that will answer his question. His query for "lowlands adjacent to a river that are subject to flooding" is formulated based on a geospatial ontology.

The semantic annotations of the geodata available are created in the same way as John's query and stored in a catalogue. Thus, John's query concept becomes machine-comparable to all geodata descriptions in this catalogue.

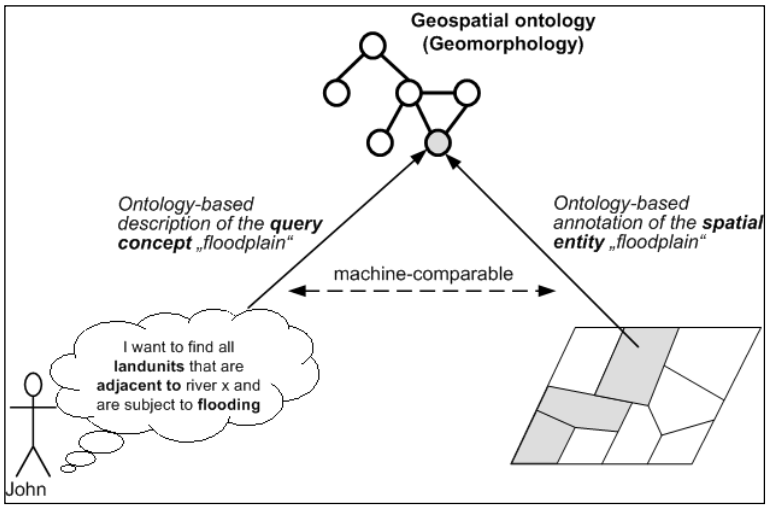

Fig. 1. Ontology-based discovery of geospatial data

As shown in [3] the integration of the matchmaking capability into SDIs overcomes some of the semantic heterogeneity problems in service discovery and thus leads to increased recall and precision.

However, in order for this approach to become widely accepted in the GI community it is essential to provide methods and tools that support the user in creating semantic descriptions. So far, no automated method for the semantic annotation of geodata exists. It remains a laborious task and data providers who are no ontology engineering specialists will neither be willing nor capable to perform it.

3 Spatial Relations

In the geospatial domain, relations among spatial entities are often as important as the entities themselves [6]. For example, for a farmer it is crucial to know that a planned plantation is on lowland *adjacent to* a river. This implies that the plantation will probably be covered by rising water once in a while and the farmer is well advised to choose plants that may cope with these conditions. This makes the representation and processing of spatial relations crucial in geographical applications.

Spatial relations have been classified in all possible various ways. We refer to a classification provided in [11], where spatial relations are classified according to their characteristic behavior in space.. *Topological relations* refer to properties like connectivity, adjacency and intersection among geospatial entities. They stay invariant under consistent topological transformation, such as rotation, translation, and scaling. *Direction relations* deal with order in space (e.g. north, east, south, and west). They are based on the existence of a vector space and, therefore, are subject to change under rotation, while invariant under translation and scaling of the reference frame. The third major type of spatial relations is *distance relation*. They refer to the geographical distances among geospatial objects (e.g. *A is close to B*, *X is very far from Y*). They reflect the concept of metric, thus change under scaling but stay invariant under translation and rotation.

Inherent to all these relations is the vagueness and imprecision in natural language expressions. Moreover, terminology and semantics of the relations varies across application domains [12]. Consider for example the spatial relation *adjacent* taken from our floodplain example. According to WordNet 2.0, *adjacent* has three slightly different senses:

(1) nearest in space or position; immediately adjoining without intervening space;
(2) having a common boundary or edge; touching;
(3) near or close to but not necessarily touching;

For the interpretation of a floodplain as defined in section 2.1, the first two senses are not applicable. Floodplains do not have to touch a river directly as long as the intervening space does not prevent the rising water from flooding the area.

Attempts to capture the semantics of spatial relations have been undertaken from both, the cognitive [13-15] and the mathematical viewpoint [16-19]. Still, there have only been few attempts to link the formal models of spatial relations developed for GIS with people's intuitive understanding of spatial relations as expressed in natural language [11]. Nevertheless, well-defined primitive operators like the Egenhofer

operators for topological relations [20] may be used as the backbone for the definition of terms used in GIS and spatial query languages (cf. [21]).

Spatial relations are usually not explicitly stored together with geographic objects but have to be inferred from the objects' geometry [12]. By extracting them, hidden information in geospatial data becomes explicit. Depending on the application domain, some spatial relations may be more significant than others for identifying relevant implicit information. For the enterprise of using spatial relations for identifying concept characteristics in datasets, it will eventually be necessary to decide on a core set of relations for the geospatial domain of discourse.

In the next section we describe how we want to exploit the potential of spatial relations in order to identify characteristic concept information for the semi-automated annotation of geodata.

4 A Method for Automating Semantic Annotation

In this section we present a method for automating the semantic annotation of geographic datasets within a specific application domain. We first introduce the general idea of using spatial analysis methods that are associated with spatial relations in ontologies to derive annotations for datasets (section 4.1). We then use the case study of annotating data containing floodplains to illustrate the different steps that eventually lead to the semi-automated creation of semantic annotations (section 4.2). In the remaining subsections each of the building blocks of the suggested methodology is presented in detail. These building blocks are:

- a geospatial ontology that defines spatial concepts based on their spatial relations and attributes (section 4.3);
- a method for associating the characteristic spatial relations in this ontology with spatial analysis methods (section 4.4); and
- a reference dataset that is needed to calculate relationships between its well-known reference entities and the unknown ones in the dataset to be annotated (section 4.5).

In this paper we concentrate on the role of spatial *relations* to introduce the fundamental idea of our approach. We are aware, that in order to arrive at reasonable results, the approach will have to be extended to include other (non-spatial) relations and (spatial and non-spatial) *attributes*, e.g. geometry, shape or extent, in the analysis. This will be part of our future work.

4.1 Using Spatial Analyses for Creating Annotations

When reasoning about concept definitions that are (partly) based on spatial relations (e.g. to infer subsumption relationships between them), the spatial relations are not treated differently from other non-taxonomic relations. They simply represent implicit domain knowledge about what it means to be an instance of that concept. This is illustrated in Tab. 1, where the non-taxonomic relations a) "adjacentTo" and b) "owner" produce the same behaviour for subsumption reasoning.

Tab. 1. Two DL inferences, using spatial (a) and non-spatial relations (b)

(a)	Boathouse ≡ House ⊓ ∃ adjacentTo.Waterbody	
	RiverBoathouse ≡ House ⊓ ∃ adjacentTo.River	⇒ RiverBoathouse ⊑ Boathouse
	River ⊑ Waterbody	
(b)	Palace ≡ House ⊓ ∃ owner.Nobleman	
	RoyalPalace ≡ House ⊓ ∃ owner.King	⇒ RoyalPalace ⊑ Palace
	King ⊑ Nobleman	

When dealing with concrete datasets, however, this implicit knowledge can be compared with the inferred characteristics from the objects' geometry, and the results of this comparison can be used for annotation. This requires that each type of spatial relation that has been identified on the domain level is associated with a spatial analysis method (see Fig. 2 for an example). This method provides a formal definition of the semantics of the spatial relation. Note, that this definition is particular for the chosen domain, because the interpretation of the relation can differ significantly depending on the viewpoint (e.g. *adjacent* in section 3). A more detailed description on how spatial relations are associated with spatial analysis methods is given in section 4.4.

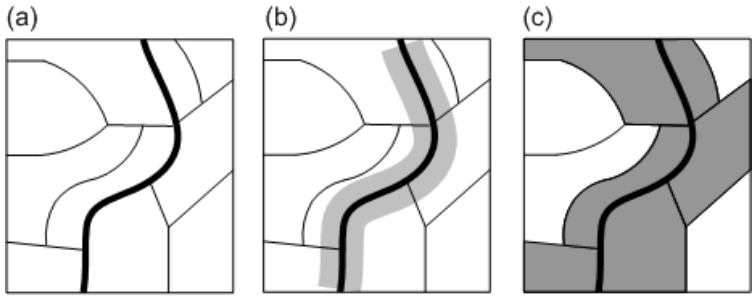

Fig. 2. Using a spatial analysis method associated with the spatial relation adjacent to (a river) to annotate the dataset shown in (a): In (b) a buffer is generated, in (c) the features intersecting the buffer are selected as being "adjacent to the river"

4.2 Walk Through for the "Floodplain" Example

Jane is working at a company that produces thematic datasets for all kind of geographic issues. The company owns a large database of geographic information and wants to make this commercially available for more customers via a geospatial web services environment. The semantic annotation for a specific domain view will consist of the following steps. Before the automated process starts, Jane has to select the domain of discourse. In our example, Jane wants to annotate her data for the geomor-

phology domain. The annotation procedure then consists of the following steps (Fig. 3):

1. All concept definitions that contain spatial relations are identified in the geomorphology ontology. From each of these concept definitions, the characteristic spatial relations are extracted. This extraction (and the subsequent analysis) is "controlled", in the sense that the system will analyse the dataset by looking explicitly for the concepts defined in the ontology (rather than performing an "uncontrolled" search for arbitrary patterns in the dataset). The process can be depicted as a decision tree, e.g. the system identifies a land unit *L* as a floodplain if *L* fulfils the following criteria:
   - *L* is adjacent to a river
   - *L* is flat
   - *L* is at most 2 m higher than the adjacent river
2. For each spatial relation, the corresponding spatial analysis method will be extracted. For example, *adjacent* is implemented as a sequence of GIS operations (section 4.4).
3. The GIS operations are applied to the geodataset to be annotated (*AnnoDS*). In order to be able to calculate the relation *R* between two entities *x* and *y* (e.g. "*x* is adjacent to some river") a reference dataset (*RefDS*) with the well-known geometry of *y* (e.g. all rivers) is required.
4. The spatial entities that meet the characteristic spatial relations of floodplains are stored as the result of this analysis step.
5. Steps 2-4 are repeated for other characteristics that define the analysed concept. In our example, this means that the flatness and difference in altitude compared to the adjacent river also have to be tested.
6. The final result set is created by intersecting the result sets of all analysis steps. If this result set contains a significant number of entities (which is greater than a certain user-defined threshold value), the geodata set will be annotated with "floodplains" in the description.
7. The result of the matchmaking process is finally presented to Jane for verification. She is also asked for further information if necessary. The ontological description is then automatically created.

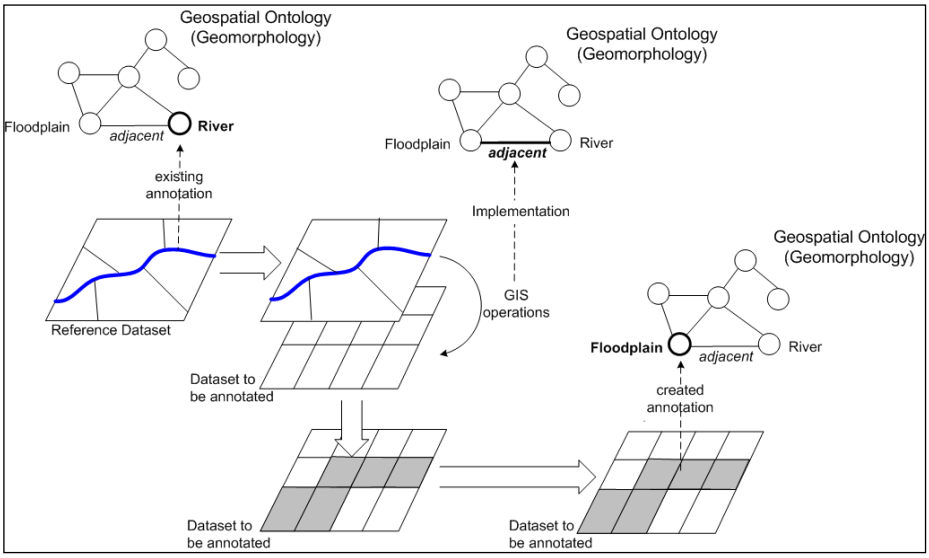

Fig. 3. Procedure for (semi-) automated annotation of geodata

4.3 Defining Geospatial Concepts Based on Characteristic Spatial Relations

What is special about geospatial concepts on the domain level? They describe geographic entities, i.e. entities that are associated with a geometry and a location relative to the earth. In consequence, geospatial concepts stand in high complex relationships to underlying physical reality and these relations may serve to define concepts and distinguish between sub-concepts on the domain level. For example, *floodplain* might be seen as a subconcept of *meadow*. One of the characteristics that distinguish *floodplains* from a *meadow* relies in the spatial relation of lying *adjacent* to a river. Such characteristic relations have to be identified by a domain expert during the ontology modelling process. For extracting the spatial characteristics of floodplains, we adopt the definition of a floodplain introduced in section 2.1. In the following, we illustrate how these can be used for formally defining the *floodplain* concept (1).

$$\forall x\,(\text{Floodplain}(x) \Leftrightarrow \text{Landunit}(x) \wedge \text{HasSlope}(x, \text{Flat}) \wedge \exists y\,[\text{River}(y) \wedge \text{Adjacent}(x, y) \wedge \text{lessThan}(\text{Difference}(\text{Altitude}(x), \text{Altitude}(y)), 2)]) \quad (1)$$

Formula (1) states that all floodplains are landunits that are flat and adjacent to a river, and whose altitude does not differ by more than 2 meters from that of the adjacent river (and that all such landunits are floodplains). For the presented method it is important that all relations and attributes used in this definition can be inferred from the dataset using some kind of spatial analysis method. Therefore, only the defining *spatial* features of floodplains but not their *non-spatial characteristics* (e.g. being subject to periodic flooding, being composed of sediments) are taken into account at this stage. However, the non-spatial characteristics are an important part of the con-

cept definitions as well. They will be subject to different kinds of analyses in future work in order to enhance the annotation results.

Some representational difficulties arise from the definition given in (1). The floodplain's "lowness" can only be described with respect to the adjacent river. Such concept interdependencies are often crucial for describing the semantics of a concept. For example, we distinguish between different spatial entities due to physical processes that lead to observable distinctions in the landscape (we can observe that some areas adjacent to a river are flooded and some are not – depending on their altitude compared to the river). For our approach, this means that the representation language chosen must be expressive enough for describing these kinds of concept interdependencies.

Formal concept definitions like (1) constitute the ontological knowledge at the domain level. Each relation could now be implemented with an analysis method that may be applied on the geodata. In the scope of this paper and to illustrate the idea we concentrate on the spatial relation *adjacent*. Nevertheless, an exhaustive classification algorithm for the concept *floodplain* would require the implementation of all relations used in its definition.

4.4 Associating Spatial Analysis Methods with Spatial Relations

The association of relations in the ontology with spatial analysis methods can be done in different ways. An analysis method can be associated as a "black box" containing the spatial relation it implements. This has the advantage that the description of the method and thus the association with the spatial relation in the ontology is simple. However, this also means that the implementation of the spatial relation is not transparent to a service provider like Jane. It can be assumed that there are always a number of different possible implementations for one spatial relation, especially if one takes into account more "fuzzy" relations like *side by side* instead of only precisely defined ones like *meet* [20].

We therefore reject the possibility to represent the implementation as a "black box" in favour of representing the analysis methods based on primitive operations defined in the ISO 19100 series of standards and specifications of the Open Geospatial Consortium (OGC). This strategy provides flexibility for adjusting the semantics of spatial relations in different application domains by changing the underlying implementation (e.g. *adjacent* could also be implemented using other primitives). Moreover, this implementation, and thus the semantic interpretation of a spatial relation remain transparent to the user.

Tab. 2 shows an informal representation of an algorithm containing the spatial analysis steps involved in implementing *adjacent*, which is based on the following specifications and standards:

- The *Web Feature Service (WFS) Implementation Specification* [10] defines the *GetFeature* operation for selecting features of a particular feature type.
- In *ISO 19109 "Rules for Application Schema"* [22] states that the geometric characteristics of a feature are described by one or more *spatial attributes* whose values are given by a geometric object (GM_Object) or a topological object

(TP_Object). We introduce the *geometry* attribute to refer to the geometric representation of a feature.

- *ISO 19107 "Spatial Schema"* [21] defines the notions of GM_Object and TP_Object and a number of operations that can be applied to them. In our example, we use the *buffer* operation, which returns a buffer polygon, and the *intersects* operation, which returns a boolean value to indicate whether two geometries intersect.

Tab. 2. Example for representing the implementation of a spatial analysis method (RefDS represents the reference dataset, and AnnoDS the dataset to be annotated)

Algorithm for implementing "adjacent to some X"	**Reference Standards and Specifications**
select all features from *RefDS* where featureType = X	Web Feature Service
create empty set *A*	
for each selected feature *f*	
A.add(f.geometry.buffer(d : Distance))	ISO 19107, ISO 19109
create empty set *B*	
for each feature *g* in *A*	
for each feature *h* in *AnnoDS*	
if (*g*.geometry.intersects(*h*.geometry))	ISO 19107, ISO 19109
B.add(*h*)	
return *B*	

4.5 Annotating a Reference Dataset

We have argued that spatial relations are especially useful for extracting implicit information from geodatasets and that their characteristics make them a perfect candidate for the extraction of implicit information from geodata. However, if a concept definition is based on a spatial relation to another geographic feature *of a particular type,* it is necessary to first identify these related features. For example, for calculating a relation like "adjacent to a *river*" the river features in the dataset (or in a different dataset covering the same spatial extent) have to be known. This can be achieved by providing *reference datasets* that have already been annotated. That is, the *river features* in the reference dataset would already be associated with the *river concept* of the domain ontology. One interesting question in this context is to what extent the provision of reference datasets could be substituted by a recursive process (e.g. for calculating the floodplain's characteristic spatial relation "is adjacent to river", the system would first have to identify rivers in a dataset by calculating *their* spatial characteristics and so on).

We propose to introduce a reference dataset for each geospatial domain ontology. The role of a reference dataset in our example can be fulfilled by the national topographic map, e.g. ATKIS (Amtliches Topographisch-Kartographisches Informa-

tionssystem)[2] in Germany. Other reference datasets needed may include a Digital Terrain Model (DTM) for calculating e.g. slope and altitude of unknown spatial entities.

5 Related Work

In this section, we relate the presented approach for semi-automatic annotation of geodata to existing work in the area of spatial data mining and to other approaches for automatic annotation.

5.1 Spatial Data Mining

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases [23-25]. We incorporate a similar strategy for automatically extracting relevant information from a geospatial database. But, instead of "mining" the dataset for potentially interesting patterns we define the spatial constraints a priori in the geospatial concepts of our domain ontology. These spatial constraints are then implemented into a supervised analysis process that aims to identify a specific concept (and not some previously unknown, potentially useful pattern). Evidently our approach requires far less complex techniques than those applied in Spatial Data Mining.

5.2 Automatic Annotation

With the emergence of the Semantic Web, the creation of semantic metadata by annotating documents has become a major concern in the community [26]. Several approaches are concerned with automating the process of semantically annotating information for the Semantic Web [27, 28].

Some research in this area has focused on the idea that spatial characteristics play a central role for effectively supporting information retrieval and annotation. In [29], Manov *et al.* demonstrate how their annotation platform can be extended by using spatial knowledge in conjunction with information extraction. In their approach, integrated gazetteers (like the Alexandria Digital Library Gazetteer) provide the additional spatial knowledge. At the MINDSWAP group, Hiramatsu and Reitsma [30] have worked on a geographic ontology to circulate geographically referenced information on the Semantic Web. Their idea is to associate georeferenced data (instead of using some gazetteer) to any other non-spatial information related to the geographic feature, i.e. they want to make use of the inherent characteristic spatial relations in order to add semantics to hypertext coded information. Similar work is done in the SPIRIT project, where knowledge stored in geographical data is made usable in an internet search engine [31]. While these approaches aim at annotating hypertext or enable spatially enhanced internet search, our work provides a method for the seman-

[2] http://www.adv-online.de

tic annotation of geodata in order to enable its ontology-based discovery and retrieval through web services.

Also related to our work is the automatic extraction of "classical" metadata (like ISO 19115) from geographic data [32]. We believe that the method presented in this paper will not only be useful for semantically annotating but also for populating missing entries in the standard metadata documents for geographic information.

6 Discussion and Future Work

We propose a method for automating the semantic annotation of geodata based on spatial relations and suggest to apply spatial analysis methods in order to extract information on spatial relations useful for annotation. Compared to knowledge extraction techniques like string-based attribute analysis, the calculation of spatial relations remains independent from the textual description of geographic features and their properties. This has the advantage that semantic heterogeneity problems inherent in the processing of natural language descriptions are avoided. In this paper we have concentrated on the role of spatial relations. Taking into account the different types of spatial relations, topological relations seem to be especially useful as they stay invariant under transformations. This is a valuable characteristic since we have to deal with heterogeneous data sources in a variety of formats, scales, and projections.

A crucial issue not yet decided on is the choice of representation language for the ontological knowledge. As has been illustrated in section 4.3, the representation language must be expressive enough for describing concept interdependencies, i.e. a floodplain's "lowness" can only be described with respect to the adjacent river. In Description logic (DL), which has been used in our previous work on ontology-based GI discovery and retrieval [3, 4], "it is impossible to describe classes whose instances are related to another anonymous individual via different property paths" [33]. Thus, current DL-based ontology languages like OWL [34] are not applicable. First-order-logic (FOL) provides the expressivity needed in our approach. However, while reasoning in DL is decidable and therefore guaranteed to terminate, proving entailment in FOL is only semi-decidable [35]. Therefore, the final decision on a representation language will have to take into account the tradeoff between expressivity of the languages and the complexity of their reasoning problems.

In section 4.4 we have described our strategy on how to define a spatial analysis method that implements a single spatial relation by combining well-defined primitive operators. The question of how to explicitly associate a spatial analysis method to a spatial relation remains. A possible approach is outlined in [36], suggesting to integrate the invocation of executable programs into static ontological knowledge. Likewise, the strategy on how to generate and formalise the analysis algorithm for the entire concept definition remains an open issue. For this task a workflow description is needed, that is applicable for representing a decision tree as outlined in section 4.2. For this, we will examine current workflow description languages like BPEL (Business Process Execution Language) and PSL (Process Specification Language).

Apart from the benefits for automating the process of annotating geodata, our approach might also contribute to enhance retrieval capabilities in geospatial web ser-

vice environments. For example: A user wants to retrieve "*all motorbike roads*", with *motorbike roads* being a concept in the domain ontology. The geometric characteristic of a *motorbike road*, i.e. its high twisting grade, is associated with a spatial analysis method. Thus, the system can apply this spatial analysis method for the on-the-fly retrieval of motorbike roads from a street network. There is no need for finding a dataset that explicitly stores *motorbike roads*.

Another application area for which the approach might be beneficial is the automation of metadata population for standard metadata like ISO 19115. In the ISO case, fields that might be filled with information extracted by the semantic annotation process are the following: descriptiveKeywords, topicCategory, geographicBox, geographicDescription [37].

In our future work we will extend our approach by providing implementation strategies for spatial attributes, like geometry, extent, and shape. Spatial attributes probably have an equally high potential for identifying characteristic information in geodata as spatial relations. For example, the analysis of the "straightness" of a water course might identify an entity as *channel* rather than *river*. However, the applicability of such an analysis highly depends on the resolution of the geodata. If all watercourse geometries are generalized in straight lines, the straightness attribute is of no value for information extraction. These dependencies on representation, resolution, and scale of the data have to be taken into account.

At this stage, we assume that the analysis of spatial characteristics will be the core methodology for identifying characteristic concept information in geodata. However, the analysis of spatial characteristics will probably not suffice for describing the semantics at the conceptual level in many cases (or not be applicable at all). Consequently, besides taking spatial attributes into account, we will also refine the presented approach by combining spatial analyses with the analysis of non-spatial attributes. This combination will eventually lead to reasonable results in the annotation process.

The vagueness of the specification of spatial characteristics is also a crucial issue. Consider the formalisation of the floodplain concept we provided in section 4.3: How many meters away from the water body still counts as adjacent? What degree of flatness still counts as flat? How is the difference of altitude between floodplain and adjacent river determined? In our future work, we will have to consider vagueness of spatial relations when specifying the associated analysis methods.

The performance of the spatial query techniques have to be evaluated and, if necessary, optimized. Methods for spatial query optimisation have been discussed in [12, 38].

We are aware that a fully automated process is out of scope. Therefore, we plan to develop a user interface that guides the data provider through the annotation process.

Acknowledgements

We would like to thank Werner Kuhn and Florian Probst for their valuable input at various stages of this work. Our thanks also go to the anonymous referees for providing valuable comments that helped to improve the content of the paper. The work

presented in this paper has been supported by the German Federal Ministry for Education and Research as part of the GEOTECHNOLOGIEN program (grant number 03F0369A) and can be referenced as publication no. GEOTECH-142.

References

1. OGC: OpenGIS Reference Model. Open GIS Consortium (2003)
2. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering. 25(1-2) (1998): 161-197
3. Klien, E., Lutz, M., Einspanier, U., Hübner, S.: An Architecture for Ontology-Based Discovery and Retrieval of Geographic Information. Presented at 7th Conference on Geographic Information Science (AGILE 2004). Heraklion, Greece. (2004)
4. Lutz, M., Klien, E.: Ontology-Based Retrieval of Geographic Information. International Journal of Geographical Information Science (IJGIS), forthcoming
5. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-Based Integration of Information — A Survey of Existing Approaches. Presented at IJCAI-01 Workshop: Ontologies and Information Sharing. Seattle, WA. (2001)
6. Papadias, D., Kavouras, M.: Acquiring, Representing and Processing Spatial Relations. Presented at Sixth International Symposium on Spatial Data Handling. Edinburgh, Scotland. (1994)
7. Schmudde, T.H.: Floodplains. In: Fairbridge, R.W. (ed.): The Encyclopedia of Geomorphology. New York. (1968) 359-362
9. Richardson, R., Smeaton, A.F.: Using WordNet in a Knowledge-based Approach to Information Retrieval (Technical Report CA-0395). Dublin City University: Dublin, Ireland (1995)
10. Bernstein, A., Klein, M.: Towards High-Precision Service Retrieval. Presented at The Semantic Web - First International Semantic Web Conference (ISWC 2002). Sardinia, Italy. (2002)
11. OGC: Web Feature Service Implementation Specification. Open GIS Consortium (2002)
12. Shariff, A., Egenhofer, M., Mark, D.: Natural-Language Spatial Relations between Linear and Areal Objects: the Topology and Metric of English-Language Terms. International Journal of Geographical Information Science. 12(3) (1998): 215-246
13. Clementini, E., Sharma, J., Egenhofer, M.: Modeling Topological Spatial Relations: Strategies for Query Processing. Computers and Graphics. 18(6) (1994): 815 - 822
14. Herskovitz, A.: Language and Spatial Cognition. In: Joshi, A. (ed.): Studies in Natural Language Processing. Cambridge University Press: Cambridge. (1986)
15. Mark, D., Svorou, S., Zubin, D.: Spatial terms and spatial concepts: geographic, cognitive, and linguistic perspectives. Presented at International Geographic Information Systems (IGIS) Symposium. Arlington, VA, USA. (1987)
16. Talmy, L.: How Language Structures Space. In: Pick, H. ,Acredols, L. (eds.): Spatial Orientation. Theory, Research and Application. Plenum: New York. (1983): 225-282
17. Egenhofer, M., Herring, J.R.: A Mathematical Framework for the Definition of Topological Relationships. Presented at the 4th International Symposium on Spatial Data Handling. Zurich, Switzerland. (1990)
18. Frank, A.U.: Qualitative Spatial Reasoning about Distances and Directions in Geographic Space. Journal of Visual Languages and Computing. 3 (1992): 343-371
19. Cohn, A.G.: A Hierarchical Representation of Qualitative Shape Based on Connection and Convexity. In: Frank, A.U., Kuhn, W. (eds.): Spatial Information Theory-A Theoretical Basis for GIS. Springer, Berlin-Heidelberg-New York. (1995): 311-326

20. Papadias, D., Sellis, T.: The Semantics of Relations in 2D Space Using Representative Points: Spatial Indexes. In: Frank, A.U. ,Campari, I. (eds.): Spatial Information Theory - Theoretical Basis for GIS. Springer Verlag, Heidelberg-Berlin. (1993): 234-247
21. Egenhofer, M.: Reasoning about Binary Topological Relations. Presented at Advances in Spatial Databases, 2nd International Symposium. Zurich. (1991)
22. ISO: ISO 19107 - Spatial Schema. ISO TC 211 (2002)
23. ISO/TC-211: Text for DIS 19109 Geographic information - Rules for application schema Vs. 2.0. Draft Version. International Organization for Standardization. (2001)
24. Shekhar, S., Zhang, P., Huang, Y., Vatsavai, R.: Trends in Spatial Data Mining In: Kargupta, H., et al. (eds.): Data Mining: Next Generation Challenges and Future Directions. AAAI Press. (2004): 357 - 380
25. Koperski, K., Han, J., Adhikary, J.: Mining Knowledge in Geographical Data. Communications of the ACM. 26(1) (1998): 65 - 74
26. Roddick, J., Lees, B.G.: Paradigms for spatial and spatio-temporal data mining. In: Miller, H. ,Han, J. (eds.): Geographic Data Mining and Knowledge Discovery (2001)
27. Handschuh, S., Staab, S. (eds): Annotation for the Semantic Web. Frontiers in Artificial Intelligence and Applications. Vol. 96. IOS Press: Amsterdam, The Netherlands. (2003)
28. Handschuh, S., Staab, S., Ciravegna, F.: S-CREAM -- Semi Automatic Creation of Metadata. Presented at the Semantic Authoring, Annotation and Markup Workshop, 15th European Conference on Artificial Intelligence (ECAI02). Lyon, France. (2002)
29. Dingli, A., Ciravegna, F., Wilks, Y.: Automatic Semantic Annotation Using Unsupervised Information Extraction and Integration. Presented at the Knowledge Markup and Semantic Annotation Workshop at the Second International Conference on Knowledge Capture (K-CAP 2003). Sanibel, Florida, USA. (2003)
30. Manov, D., Kiryakov, A., Popov, B., Bontcheva, K., Maynard, D., Cunningham, H.: Experiments with geographic knowledge for information extraction. Presented at the NAACL-HLT 2003, Workshop on the Analysis of Geographic References. Edmonton, Alberta, Canada. (2003)
31. Hiramatsu, K., Reitsma, F.: GeoReferencing the Semantic Web: ontology based markup of geographically referenced information. Presented at the Joint EuroSDR/EuroGeographics workshop on Ontologies and Schema Translation Services. Paris, France. (2004)
32. Heinzle, F., Sester, M.: Derivation of Implicit Information from Spatial Data Sets with Data Mining. Presented at the XXth Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS). Istanbul, Turkey. (2004)
33. Manso, M.A., Nogueras-Iso, J., Bernabe, M.A., Zarazaga-Soria, F.J.: Automatic Metadata Extraction from Geographic Information. Presented at the 7th Conference on Geographic Information Science (AGILE 2004). Heraklion, Greece. (2004)
34. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (eds.): The Description Logic Handbook. (2003)
35. Grosof, B., Horrocks, I., Volz, R., Decker, S.: Description Logic Programs: Combining Logic Programs with Description Logic. Presented at 12th Intl. Conf. on the World Wide Web (WWW-2003). Budapest, Hungary. (2003)
36. W3C: OWL Web Ontology Language Overview. (2004)
37. Russell, S., Norvig, P.: Artificial Intelligence - A Modern Approach. (2002)
38. Borchert, R.: How can a knowledge base run executables on the frame level? Presented at the International Protege Workshop. Manchester, England. (2003)
39. ISO/TC-211: ISO 19115:2003. Geographic information - Metadata. International Organization for Standardization. (2003)
40. Papadias, D., Theodorodis, Y.: Spatial relations, minimum bounding rectangles, and spatial data structures. International Journal of Geographical Information Systems. **11**(2) (1997): 111-138