An Architecture for Ontology-Based Discovery of Geographic Information[1]

Michael Lutz[1], Udo Einspanier[1], Eva Klien[1], and Sebastian Hübner[2]

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

[2] Center for Computing Technologies (TZI), University of Bremen
Universitätsallee 21-23, D-28359 Bremen
huebner@informatik.uni-bremen.de

Abstract: Finding and accessing suitable information in the open and distributed environments of current Spatial Data Infrastructures (SDIs) is a crucial task. Catalogues provide searchable repositories of information descriptions, but the mechanisms to support the tasks of discovery and retrieval are still insufficient. Problems of semantic heterogeneity caused by synonyms and homonyms can arise during free-text search in catalogues. Moreover, once a suitable Web Feature Service (WFS) is found and accessed, the property names of a feature are often difficult to interpret. This paper introduces an architecture for ontology-based discovery and retrieval of geographic information that solves semantic heterogeneity problems of current query capabilities. Based on a (real-world) scenario from the area of flood management, the application of our approach shows that the information requestor can be efficiently supported.

1 Introduction

Geographic information (GI) is a key factor in planning and decision-making in a variety of domains. To facilitate the discovery and access to this often highly distributed information, Spatial Data Infrastructures (SDIs) are currently being set up within regions, countries and even across borders [1, 2]. In these SDIs, catalogue services are used for discovering appropriate GI for a specific task. While a certain level of syntactic interoperability can be achieved through standardising interfaces [3] and metadata schemas [4], a number of problems caused by semantic heterogeneity present challenges during GI discovery. One possible

[1] This is a abridged and substantially revised version of an article [12] published in the Proceedings of the 7th Conference on Geographic Information Science (AGILE 2004).

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). It can be referenced as publication no. GEOTECH-78.

approach to overcome these problems is the explication of knowledge by means of ontologies. In this paper we introduce an approach for ontology-based discovery of GI.

2 Semantic Heterogeneity Problems During GI Discovery

In current standards-based catalogues, users can formulate queries using keywords and/or spatial filters. Semantic heterogeneity can lead to problems during the discovery process if a requester performs a simple keyword-based search. To illustrate these problems, we introduce a simple GI discovery scenario. A requester (who we will subsequently call 'John') is interested in the measurement of the water level at a specific control point of the river Elbe. Two services that potentially provide suitable information for John are listed in table 1. The keywords used in a catalogue to annotate the GI provided are also given.

Table 1: Two examples for services providing information on water level measurements and the keywords used in the metadata

	Service 1	Service 2
provides	water level measurements (including date and time) in centimetres	water level measurements (including date and time) in metres
keywords	depth, control point	water level, control point

If John uses "water level" as a search term, he will discover service 2 but not service 1, i.e. recall is low (there are false negatives). Conversely, if he uses "control point" as a search term, he will discover both services. However, one of these services might be unsuitable if John is looking for water level measurements *given in a certain unit of measure*, e.g. centimetres. In this case, precision is low (there are false positives). These examples show that keywords used in free-text entries have to be considered a poor way to capture the semantics of a query or item [5].

3 Ontology-Based Discovery

In this chapter we introduce our approach for ontology-based discovery of GI, which is based on the classification of DL concepts [6] representing geographic feature types (i.e. classes of geographic objects with common characteristics) on the one hand and the user's query on the other hand.

For the classification, RACER [7] is employed as a terminological reasoner, which works with concepts described in the Description Logic SHIQ [8]. We have chosen RACER because of its ability to reason about concrete domain slots, which we already use, and inverse roles, which we plan to use in the future.

For the development of the ontologies used in the approach we adopt a modified version of the *hybrid ontology approach* [9], which is based on the idea of having a source-independent shared vocabulary for each domain based on which several application ontologies can be defined. It is assumed that the shared vocabulary consists of the basic terms of a domain, which are understood by all members without further explanation. We consider the shared vocabulary to consist of several ontologies (Figure 1). We believe that using full-grown on-

```
MEASUREMENTS DOMAIN ONTOLOGY
(define-concept  Quantity (and
  (some  value              Double)
  (some  unitOfMeasure      Unit)))
(define-concept  Measurement (and
  (some  observable         Phenomenon)
  (some  quantityResult     Quantity)
  (some  location           Locator)
  (some  timeStamp          TimePrimitive)))

HYDROLOGY DOMAIN ONTOLOGY
(define-concept  Depth (and
  Phenomenon
  (some referenceSurface  Surface)))
(define-concept  WaterLevel (and
  Depth
  (some referenceSurface WaterTable)))
```

Figure 1: Extract of the shared vocabulary consisting of the measurements and hydrology domain

```
(define-concept FeatureType1 (and
  Measurement
  (some quantityResult
    (all unitOfMeasure Centimetre))
  (some observable WaterLevel)))
(define-concept FeatureType2 (and
  Measurement
  (some quantityResult
    (all unitOfMeasure Metre))
  (some observable WaterLevel)))
```

Figure 2: Examples for defining application concepts

```
(define-concept QueryConcept1 (and
  (some observable WaterLevel)
  (some quantityResult *top*)
  (some location *top*)
  (some timeStamp *top*)))
(define-concept QueryConcept2 (and
  (some observable WaterLevel)
  (some quantityResult
    (all unitOfMeasure Centimetre))
  (some location *top*)
  (some timeStamp *top*)))
```

Figure 3: Examples for defining query concepts

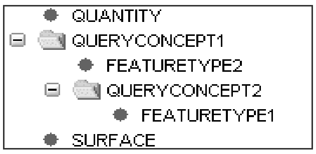

Figure 4: Subsumption hierarchy including two query concepts and application concepts for the two feature types introduced in section 2

tologies as a shared vocabulary (rather than just simple lists of terms or simple hierarchies) facilitates the understanding and use of the shared vocabulary terms for specifying semantic descriptions of geographic data and user requirements.

Feature types are annotated using specific application concepts that are built using terms from the shared vocabulary (Figure 2). In our example, two concepts are defined for the feature types provided by services 1 and 2. Both are defined as a Measurement that observes a *WaterLevel* with only the difference that the measurement is either given in *Centimetre*s (FeatureType1) or *Metre*s (FeatureType2). By referring to the domain concept Measurement, both definitions also imply that the feature types provide a *quantityResult*, a *location* and a *timeStamp*.

The user's query concept can either be a concept from an existing ontology, or it can be defined by the user based on roles and concepts from the shared vocabulary. In our example, John can also specify two concepts representing feature types that have *WaterLevel* as an *observable* and provide a *location* and a *timeStamp* (Figure 3). While the first concept does not specify a specific unit of measure, the second concept requires the unit of measure to be *Centimetre*.

A classification of the domain, application and query concepts in RACER (Figure 4) shows that both FeatureType1 and FeatureType2 are subsumed by the first query concept. Thus, in contrast to the keyword-based search, both services would be correctly discovered, i.e. recall would be increased. The second query concept only subsumes FeatureType1. Again, this is the desired result as only this feature type provides water level measurements in centimetres. This illustrates that compared to keyword queries, the ontology-based approach can also increase precision.

4 Architecture

In order to support the advanced query capabilities described above, some new service interfaces and information items are needed in addition to the well-known components as catalogues of current SDIs:

- First, we have to provide the ontologies. For each application schema there is one application ontology that is described with the shared vocabulary of the corresponding domain. These ontologies provide the formal description of the application schema of a data source. They are referenced in the *keyword* metadata section of the corresponding ISO 19115 documents for that data source. The keyword section has been chosen because it is a compound metadata entity that does not only provide the keyword term and category, but also a reference to its formal definition.
- To provide access to the ontologies, two new interfaces are defined: The *Concept Definition Service* interface allows access to the concepts of the shared vocabulary and application ontologies. The *Concept Query Service* (CQS) interface allows to reason about possible matches with simple and defined concept search. A concept is considered a match if it is equal to or subsumed by the query concept. To increase recall, it might also be sensible to consider other types of matches, e.g. those proposed in [10].
- The second component is a cascading catalogue service. It provides access through the standard *OGC Stateless Catalogue Service* interface, thus implementing the decorator design pattern [11]. It extends the functionality of the conventional catalogue service by analysing and manipulating the filters of metadata queries. If a catalogue query contains a DL query concept, the matchmaking capabilities of the CQS are used. The returned list of matching (sub)concepts from existing application ontologies is added to the query, which can then be sent to any conventional standard catalogue service because the expanded query requires only the usual exact word match.
- Finally, a *client* supports the user to formulate catalogue queries containing a DL query concept for the required feature type. For this task, it uses the *Concept Definition Service* to ask for *Query Templates*. Query Templates contain expedient combinations of slots and fillers and prevent the inexperienced user from defining queries that do not make sense. After a template has been chosen, is can be filled unto a selectable level. This two-step procedure reduces the amount of terms – slots and fillers – that are presented to the user significantly. For experienced users, unlimited access to the whole vocabulary has been considered, although this is not implemented yet.

The information flow between these components is described in more detail in [12]. A prototypical implementation can be accessed from http://www.meanings.de/.

5 Conclusion and Future Work

We have presented an ontology-based approach and architecture that can contribute to solving existing problems during keyword-based discovery of GI. The tested scenario comprises information items with simple structures. Future tests of the architecture will include more complex application schemas and examples from other domains. Also, we plan to extend the terminological reasoning capabilities by implementing “emergent, context-dependent concept equivalence”. This will allow negotiation about service and data properties, instead of quitting cooperation immediately when a request does not match the features fully.

The presented architecture is component-based, i.e. it is extendable in various directions. So far, the Cascading Catalogue Service and the reasoning component are tightly coupled in the architecture. However, the standardized interfaces allow to extend the architecture with multiple and exchangeable components. It is also planned to extend the architecture with modules for spatial and temporal reasoning [13] as well as gazetteer services.

Also, the approach and architecture will be expanded to enable ontology-based *retrieval* of (previously unknown) GI. The user will then be able to formulate his actual question on the data straight away (i.e. without having to perform a query for relevant sources on the metadata first) using terms from the familiar shared vocabulary. The discovery and the filter formulation for retrieval will then be automated within the system. This “intelligent” query capability will enhance the usability of existing GI even further.

References

[1] L. Bernard, "Experiences from an implementation Testbed to set up a national SDI", presented at 5th AGILE Conference on Geographic Information Science 2002, Palma de Mallorca, 2002.

[2] J. Riecken, L. Bernard, C. Portele, and A. Remke, "North-Rhine Westphalia: Building a Regional SDI in a Cross-Border Environment / Ad-Hoc Integration of SDIs: Lessons learnt", presented at 9th EC-GI & GIS Workshop - ESDI: Serving the User, A Coruña, Spain, 2003.

[3] GDI-NRW, "Catalog Services für GeoDaten und GeoServices, Version 1.0", International Organization for Standardization & OpenGIS Consortium 2002 2002.

[4] ISO/TC-211, "ISO 19115:2003. Geogaphic information - Metadata", International Organization for Standardization 2003.

[5] A. Bernstein and M. Klein, "Towards High-Precision Service Retrieval", presented at The Semantic Web - First International Semantic Web Conference (ISWC 2002), Sardinia, Italy, 2002.

[6] H. Beck and H. S. Pinto, "Overview of Approach, Methodologies, Standards and Tools for Ontologies", 2002, pp. 58.

[7] V. Haarslev and R. Möller, "Description of the RACER System and its Applications", presented at International Workshop on Description Logics (DL-2001), Stanford, USA, 2001.

[8] I. Horrocks, U. Sattler, and S. Tobies, "Reasoning with Individuals for the Description Logic SHIQ", presented at 17th International Conference on Automated Deduction (CADE-17), Pittsburgh, Pennsylvania, USA, 2000.

[9] U. Visser and H. Stuckenschmidt, "Interoperability in GIS - Enabling Technologies", presented at 5th AGILE Conference on Geographic Information Science, Palma de Mallorca, Spain, 2002.

[10] J. González-Castillo, D. Trastour, and C. Bartolini, "Description Logics for Matchmaking of Services", HP Laboratories, Bristol, UK HPL-2001-265, 2001.

[11] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: elements of reusable object-oriented software*. Boston, MA, USA: Addison-Wesley, 1995.

[12] E. Klien, U. Einspanier, M. Lutz, and S. Hübner, "An Architecture for Ontology-Based Discovery and Retrieval of Geographic Information", presented at 7th Conference on Geographic Information Science (AGILE 2004), Heraklion, Greece, 2004.

[13] T. Vögele, S. Hübner, and G. Schuster, "BUSTER - An Information Broker for the Semantic Web", *KI - Künstliche Intelligenz*, vol. 03, pp. 31-34, 2003.