ADDING SEMANTICS TO SPATIAL CONTENT - A LAND COVER SCENARIO

Mariana Belgiu1*, Josef Strobl1, 2, Manfred Mittlboeck3
Institute for Geographic Information Science, Austrian Academy of Sciences, Salzburg, Austria, mariana.belgiu@oeaw.ac.at
2josef.strobl@sbg.ac.at
3Research Studio iSpace, Salzburg, Austria, Manfred.mittlboeck@researchstudio.at

Abstract

For successful implementation of geospatial information sharing platforms, we need solutions for overcoming syntactic and semantic ‘noise’ that may occur while sharing spatial data across communities, enterprises and application domains. For example, querying existing spatial data repositories is commonly based on thematic, spatial and temporal criteria. Semantic heterogeneity problems caused by ambiguity of thematic keywords pose great challenges for discovering appropriate spatial information. Over the last years, ontologies have been identified as solutions to overcome the challenges associated with spatial information discovery by adding semantics to (spatial) content. Therefore, an increasing number of (formalized) knowledge bases are now available. This is an important step towards enhancing information discovery experience, but it is not enough. Unfortunately, existing domain knowledge bases operate as standalone solutions and do not create the envisioned semantic synergies. This paper describes a conceptual framework flexible enough to integrate existing knowledge bases. It focuses mainly on the development of a lightweight ontology, rather than on technical implementation of it into discovery capabilities of existing repositories. Proposed framework utilizes lightweight ontologies for explicit specification of domain conceptualization and maps domain concepts against related concepts defined in other formalized concepts schemas. SKOS (Simple Knowledge Organization System) model is used to specify both hierarchical (general/specific or broader/narrower relations) and associative relations between concepts defined in different Knowledge Organization Systems (consistent schemata). Land cover datasets serve as running scenario across this paper. Challenges associated with domain ontology development are also discussed.

Keywords: ontologies, SDI, SKOS, knowledge base, geospatial portals, land cover

* Corresponding author
1. **INTRODUCTION**

Information intensive domains, such as spatial planning, environmental monitoring, security or disaster management pose new challenges for information evaluation, retrieval, extraction and usage. This requires broad availability of and accessibility to reliable spatial data sets and services, distributed across domains, communities and geographic boundaries. Spatial Data Infrastructures (SDIs) facilitate the exchange of spatial data between stakeholders (Rajabifard, Feeney and Williamson, 2002) creating the premises for developing sustained “spatial information highways” (Strobl and Nazarkulova, 2009) and for “maximizing overall usage of published resources” (Masser, 2007). Existing SDI initiatives facilitate discovery, access and seamless integration of disparate spatial resources (Li et al, 2011). They supply the information infrastructures required for creating a spatially enabled society (Sadeghi-Niaraki et al, 2010) where data providers and users are interacting for their mutual benefits. Thus, data providers (national agencies, cadastral offices, academia) have the opportunity to overcome the information silos status of their spatial content and users can focus on making well informed decisions based on existing spatial information.

For successful implementation of data sharing platforms, syntactic and semantic ‘noises’ occurring while sharing spatial data across communities, enterprises and application domains (Rajabifard et al, 2002) need to be overcome. The standardized encoding formats (Geographic Markup Language - GML), together with standardized metadata models (e.g. ISO 19115, ISO 19119, ISO 19113 & ISO 19157), schemata (ISO 19119, ISO 19139) and standardized web service interfaces (Catalog Services for Web, Web Feature Services – ISO 19142, Web Coverage Services, and Web Map Service – ISO-19128 etc.) put the basis for syntactic interoperability of geographic resources. Nevertheless, syntactic interoperability is only one facet involved in the process of spatial data sharing. Semantics or the meaning of the shared information needs to be considered as well, if successful and sustained sharing platforms are to be developed (Lutz et al, 2009). In this paper we are addressing the problems posed by semantic heterogeneity occurring on metadata level and which hinders the efficiency of spatial information discovery. Figure 1 underlines the main challenges associated with the process of geographic information discovery. Users query existing geospatial catalogue services using keywords, very often in conjunction with additional temporal and spatial criteria. This approach is a useful starting point for spatial information discovery and retrieval, but it has shortcomings generated by differences in semantics of disparate geographic resources (Bishr, 1998). As users’ terminologies may differ from the terminology used by data providers, keyword-based search may have low recall (Athanasis et al. 2009). This means that users cannot discover all relevant sources to answer the question they are aiming at. Query results often show low precision, i.e. some of the discovered geographic information is not relevant (Klein and Bernstein, 2004).

![Figure 1: Semantic heterogeneity issues challenging information discovery](image-url)
Today, extensive work is being carried out to enhance spatial information discovery in GI portals by adding semantics to its repositories either through “semantic annotation of spatial resources (data and services) or through direct mapping of semantics in associated registries” (Stock et al, 2010). All these approaches involve the development of ontologies (formalized knowledge bases), seen as solution for explicating the meaning of spatial concepts (Klien, 2008). Unfortunately, existing ontologies are operating as standalone solutions for overcoming semantic heterogeneity. Appropriate methods for integrating existing Knowledge Organization Systems (Thesauri, ontologies) are to be developed (De Martino, 2011). This paper underlines the need to create synergies between approaches dedicated to semantic enhancement of spatial content. Thus, the present paper aims to (1) underline the main challenges associated with ontology developing, (2) develop a land cover lightweight ontology with SKOS model (3) map domain knowledge to concepts defined in other knowledge bases.

The remainder of the paper is structured as follows: section 2 introduces the present challenges in spatial information retrieval with a special focus on land cover information and offers a brief overview of previous work on semantics–based search. The proposed concept and methodology is outlined in section 3 introducing lightweight ontology development approaches. Section 4 is dedicated to conclusions and future outlook.

2. BACKGROUND AND MOTIVATION

2.1 Semantic heterogeneity of Land Cover data sets

Today, climate change together with urban sprawl, accelerated world population growth, land degradation caused by erosion, deforestation or desertification influence societies. For developing strategies responding to these challenges, reliable information on ecological conditions and indicators (e.g. land cover, land use etc.) is required. This information helps to understand the interaction between human and natural systems, to make informed decisions and to adopt sustainable development policies. Land Cover represents “the physical and biological cover of earth’s surface including artificial surfaces, agricultural areas, forest, semi-natural areas, wetlands, water bodies” (INSPIRE Data Specifications – Land Cover, 2011) linking many aspects of physical and human environments. Land cover surveys are conducted through various land cover monitoring initiatives at various levels, using different categorization systems (classification systems), nomenclatures or legends (Jansen, Groom and Carrai, 2008):

- EEA CORINE Land Cover Program
- LUCAS Survey, carried out by EuroStat
- High Resolution Layers GMES: [http://www.geoland2.eu](http://www.geoland2.eu)
- LiSA (Land Information System Austria): [http://www.landinformationsystem.at](http://www.landinformationsystem.at)

Due to various land cover mapping initiatives, several Classification Systems have been developed: UN FAO Land Cover Classification System LCCS, USGS Anderson System (Anderson, 1976), European CORINE System (CORINE Land Cover, 2007). Besides the mentioned classification systems, a plethora of other national or local classifications exist.

In the last years, sustained efforts have been dedicated to the harmonization of land cover data set (Ahlqvist and Shortridge, 2010). For example, a Land Cover Meta Language (LCML), derived from FAO Land Cover Classification System LCCS has been proposed as ISO standard (ISO/DIS 19144-2:2011). It proposes a solution for
harmonized documentation of land cover classifications and nomenclature, but validation is still required (INSPIRE Data Specifications – Land Cover, 2011). HarmonISA project (http://harmonisa.uni-klu.ac.at/) aimed at harmonization of land cover and land use data across three regions from Italy, Slovenia and Austria. It addresses the need to add semantics to land use classes in order to enable seamless integration of disparate data sources (in relation to trans-boundary issues). Despite existing initiatives for land cover data harmonization, land cover datasets are yet semantically heterogeneous information sources.

Semantic heterogeneity of land cover data may cause synonymy problems, i.e. different terms refer to the same concept, and polysemy problems, i.e. use of the same term when referring to different concepts (Athanasis et al, 2009). For example, the CORINE Land Cover Classification uses the concept arable_Land to define an (agricultural) area used for agricultural production, while GEMET Thesaurus and AgroVoc Thesauri use the concept arable_Farming to describe the same entity. To overcome these problems an integrated and flexible framework for semantic enrichment of spatial information on Land Cover is required. A modularized framework is proposed as solution to adding semantics to spatial content. In this paper we do not focus on assessing semantic similarity between different land cover classification systems, the goal is rather to develop a land cover knowledge base, formalized using a standardized model and envisioned as important contribution towards integrating existing knowledge bases.

2.2 Land Cover Data Discovery

To inform users about available geographic resources, providers need to document them using standardized metadata models. Various metadata elements (title, keywords, author, date, spatial extent etc.) serve as search or browse elements. In Austria, public geportals like INSPIRE Österreich (http://www.inspire.gv.at/Geoportale/National.html) or Geoland (http://www.geoland.at), and research geportals like GIsceience-research.org (http://geoportal.giscience-research.org) (Figure 2) collect references to available geographic resources and enable users to query registered repositories by spatial or temporal criteria and various topics (ISO defined topics), INSPIRE themes (within European environment for spatial sharing), keywords:

- **What**: data category, content type, keywords,
- **Where**: spatial filtering (‘anywhere’, ‘intersecting’, ‘fully within’)
- **When**: temporal filtering (‘start date’, ‘end date’)

For example, to retrieve data on arable land on a global level, users query data repositories using keywords and additional constraints: spatial and temporal filtering of results, scale specification and data type selection (Q to Qn stands for Query):

\[ Q = Q_1 \land Q_2 \land Q_3 \land Q_4 \land Q_n \]

- **Q1 Keyword**: Arable Land
- **Q2 Spatial Filter**: Q2.1 Site name or Geographic Coordinates
- **Q3 Scale (Observational Scale)**
Q4 Data type (Coverage or Vector)

Q5 Temporal Filter

This searching technique works well as long as the provided keyword list fits the user-defined keyword. Otherwise, information requesters cannot find all relevant datasets (Lutz et al, 2009). Therefore, semantics needs to be added to spatial content for constraining the use and interpretation of domain concepts (Kuhn, 2001).

Figure 2: Sample Geoportal (GIScience Research Cluster)

2.3 Previous Work

This section underlines the existing approaches and methods developed for improving searching capabilities by adding semantics to spatial content stored in catalogue services. Available Spatial Knowledge Organization Systems are also presented.

2.3.1 Catalogue Services

Data registries or clearinghouses are a core component of any SDI. A comprehensive assessment of implemented spatial data clearinghouses can be found in Crompvoets et al (Crompvoets et al, 2004). Existing cataloging services provide interfaces to available data and service references, and allow updating, querying, discovering and evaluating spatial data and services. Usually resource
documentation encompasses information on resource content, purpose, quality, validity, spatial coverage and temporal extent. Semantic annotation of these references is important "if more advanced automation approaches are going to be employed within SDIs" (Stock et al, 2010) with the general purpose to improve search recall and search precision (Salton and McGill, 1983, Klein and Bernstein, 2004). Considerable research efforts have been oriented towards supporting users in GI retrieval. To enable automated discovery and ultimately orchestration of services, an ontology-based architecture has been proposed for Geographic Information (GI) discovery and GI retrieval through Web Feature Service (Lutz and Klien, 2006). Important research work has been conducted to assess the role of spatial relations in GI retrieval. Attribute Relation Graph – ARG has been used to query and retrieve geographic information based on spatial relations (Liu et al, 2010). Yue et al (2011) addressed the need to represent geographic resources semantics using Web Ontology Language and to register these semantics in CSW-ebRIM profile (OWL) (Yue et al, 2011). The concept and technologies of Geospatial Semantic Web have been also employed as solution to guarantee semantic interoperability (Zhang, Li and Zhao, 2007).

2.3.2 Knowledge Organization Models

There is a single geographic reality which establishes the foundations for different conceptualizations, formalizations and perceptions. Domain experts divide this geographic reality into entities whose intrinsic characteristics and interdependencies reflect domain specific theories, goals, methodologies and terminologies (Mark, Smith and Tversky, 1999). Important research work has focused on developing ontologies with a general view to explaining the meaning of geospatial concepts. These efforts lead to the development of both upper-level ontologies and domain ontology:

- SUMO (The Suggested Upper Merged Ontology)
- SWEET (Semantic Web for Earth and Environmental Terminology)
- DOLCE (Descriptive Ontology for Linguistics and Cognitive Engineering) (Sieber, Wellen and Jin, 2011)
- DIGEST (Feature and Attribute Coding)
- USGS Spatial Data Transfer Standard (SDTS)
- Geographic Data Description Directory (GDDD)
- Alexandria Digital Library feature Type Thesaurus
- GEMET (General Multilingual Environmental Thesaurus)
- AGROVOC (Agricultural Information Management Standards)
- EuroVoc (Multilingual Thesaurus of the European Union)

Higher level ontologies have been designed with the general purpose to explain the meaning of generic concepts, while domain ontologies attempt to capture domain knowledge into structured and formalized models (De Martino and Albertoni, 2011). Although these knowledge bases are important building blocks in achieving semantic interoperability of disparate data sources, they are operating as standalone approaches and do not create the envisioned semantic synergies. Therefore, conceptual frameworks and methods are still required to integrate and to develop interoperable domain ontologies (Kliën, 2008).

3. FRAMEWORK DESCRIPTION

3.1 Methodological framework
In this section, a general methodological framework for semantic enrichment of spatial content is described. This conceptual framework consists of several components. It starts with spatial data sets (in our case study: land cover data) acquired using different methods and stored in legacy management systems (Figure 3). Existing datasets are described using standardized Metadata model (e.g. ISO 19139/19115). On the other hand, knowledge about land cover is formalized using appropriate knowledge models. In this work, SKOS has been used for explicit specification of concepts and relations between them. The resulting knowledge base (this concept is used interchangeably with Knowledge Organization System and concepts schemata) is interlinked with concepts captured in other concepts schemata (more details are provided in Table 1 and following subsections). This paper focuses only on defining a land cover SKOS model, while the mapping between formalized land cover terms and metadata and its validation will be the subject of a future paper.

![Figure 3: Methodology Overview](image)

<table>
<thead>
<tr>
<th>Methodology Overview</th>
<th>Description</th>
<th>Examples</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>More general knowledge base</td>
<td>Explication of general concepts</td>
<td>EuroVoc Thesaurus</td>
<td>Common denominator supporting creation of interoperable knowledge bases</td>
</tr>
<tr>
<td>Domain Knowledge base</td>
<td>Explicit specification of factual domain knowledge</td>
<td>Land Cover</td>
<td>Define both hierarchical and associative relations between concepts within the same concept schema or to other concepts schemata</td>
</tr>
</tbody>
</table>

3.1.1 Knowledge Representation Models
As already mentioned, data source heterogeneity generates semantic problems preventing seamless integration and sharing of spatial data sets. Ontology has been seen as solution to overcome semantic-related problems, its role in information integration, semantic augmentation and knowledge management being widely accepted (Uschold and Grueninger, 1996, Chandrasekaran, Josephson and Benjamins, 1998). It is defined as “formal, explicit specification of a shared conceptualization” (Gruber, 1993). Thus, domain ontology captures domain concepts, their properties and relations holding them together. Resulting ontological data structures mediate and translate between disparate systems and data sources, and enhance Information Retrieval (IR) tasks (Lopez-Pellicer et al., 2011).

Several models to encode domain knowledge are available. Description Logics for example is a family of knowledge representation languages, subsets of first-order logic. They provide the basis for OWL (Web Ontology Language), proposed as standard for Semantic Web (Lutz and Klien, 2006). By using OWL-DL, available reasoners can be used to check the consistency of developed ontologies, and logical reasoning can be performed to infer implicit relations between defined concepts. RDF (Resource Description Framework) is one W3C standard used to represent information in the WWW (RDF, 2004). It is a triple-based (subject, predicate, object) representation of knowledge and uses URIs (Uniform Resource Identifiers) as a mechanism to uniquely identify information on the Web.

SKOS (Simple Knowledge Organization System) is a “data model for sharing and linking knowledge organization systems via Web” (SKOS, 2009) and is being used to encode thesauri, taxonomies and various classification schemes. It allows us to define both hierarchical and associative relations between defined concepts. While hierarchical relations define the links between a super-class and a sub-class (broader-narrower, parent-child relations), associative relations help to specify relations between concepts belonging to different concept schemata. To encode a land cover thesaurus, the SKOS model has been used. It is a standard model for data interchange on the web (SKOS, 2009) and a building block towards developing a web of linked data (Linked Data).

3.1.2 Knowledge acquisition and integrating of existing ontologies

The first step in domain knowledge representation is domain concepts acquisition by performing “an effective ontological analysis of some field of knowledge” (Chandrasekaran et al., 1998). Ontology engineering is not a straightforward task as it requires deep understanding of domain knowledge. Usually, knowledge engineers lack domain knowledge and domain specialists lack the required expertise for developing a formalized knowledge model. To deal with this problem, we use the solution proposed by Kuhn (Kuhn, 2001) for ontology grounding in natural language descriptions of domain information. [Natural language related vagueness and ambiguity problems are the subject of an ongoing research and it will not be addressed in this paper].

Thus, we analyze natural language descriptions of European CORINE land cover classification. Existing reports (CORINE Reports, 2000) contain a detailed and comprehensive description of both land cover nomenclature and descriptors involved in domain categorizations. For semi-automatic extraction of relevant terms, text-mining technique has been used (QDAMiner - http://www.provalisresearch.com). Extracted concepts and relations between them were listed in a concept mapping tool (CmapTools - http://cmap.ihmc.us/) and then imported into the Protégé ontology editor.
The second step consists of mapping (or linking) land cover concepts against related concepts defined in other knowledge base as we need to take advantage of already existing Knowledge Organization Systems and avoid “reinventing the wheel” (Molina and Bayarri, 2011). Land cover domain concepts have been mapped against environmental concepts defined in EuroVoc European multilingual Thesaurus for facilitating knowledge bases inter-operation and for leveraging already existing knowledge formalization efforts. More details are provided in next section.

### 3.1.3 Defining semantic relations and properties

We define domain concepts (forest, water, meadows etc.), their intrinsic characteristics, and relations with each other (hierarchical and associative relations, Figure 4), using SKOS standard model. The Protégé Ontology Editor (http://protege.stanford.edu/) and associated SKOSed plugin (http://code.google.com/p/skoseditor/) have been used for thesauri creation and maintenance. Below is an example of knowledge representation using the SKOS model and Table 2 introduces the SKOS concepts used for domain specification:

```xml
<skos:Concept rdf:about="http://www.semanticweb.org/ontologies/2011/10/LandCoverThesaurus.owl#Agricultural_Areas">
    <skos:prefLabel>Agricultural Areas</skos:prefLabel>
    <skos:exactMatch rdf:resource="http://www.semanticweb.org/ontologies/2011/10/LandCoverThesaurus.owl#Agricultural_Land"/>
    <skos:narrower rdf:resource="http://www.semanticweb.org/ontologies/2011/10/LandCoverThesaurus.owl#Arable_Land"/>
    <skos:broader rdf:resource="http://www.semanticweb.org/ontologies/2011/10/LandCoverThesaurus.owl#Land_Cover"/>
    <skos:inScheme rdf:resource="http://www.semanticweb.org/ontologies/2011/10/LandCoverThesaurus.owl#Land_Cover_Thesaurus_CORINE"/>
</skos:Concept>
```

![Figure 4: Hierarchical and associative relations](image)

To link land cover concepts to concepts defined in EuroVoc Thesaurus, appropriate SKOS specific mapping relations can be used. Thus, if two concepts are
to some degree similar and can be used interchangeably in various applications, skos: exactMatch mapping property is used. For example, Arable Land concept is similarly defined in both EuroVoc and Land Cover Thesaurus. In this paper, we defined only the exact and related match assertions, but in the future SKOS broad and narrow match assertion will be also used to interlink concepts. The links between concepts have been established manually, but an ‘automatic’ approach for integrating existing knowledge bases has been also proposed by De Martino (De Martino et al, 2011).

<table>
<thead>
<tr>
<th>URI</th>
<th>Example</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>skos:Concept</td>
<td>ArableLand</td>
<td>Captured Domain Concept</td>
</tr>
<tr>
<td>skos:inScheme</td>
<td>Land Cover Thesaurus</td>
<td>Defined schema</td>
</tr>
<tr>
<td>skos:prefLabel</td>
<td>Arable Land</td>
<td>One value of skos:prefLabel per language tag</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Instances: owl:AnnotationProperty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sub-properties: rdfs:label</td>
</tr>
<tr>
<td>skos:allLabel</td>
<td>Arable Area</td>
<td>Instances: owl:AnnotationProperty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sub-properties: rdfs:label</td>
</tr>
<tr>
<td>skos:hiddenLabel</td>
<td>“Arble”</td>
<td>Instances: owl:AnnotationProperty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sub-properties: rdfs:label</td>
</tr>
<tr>
<td>skos:broader</td>
<td>Land Cover</td>
<td>Semantic relation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Used to assert a direct (immediate) hierarchical link between two SKOS concepts.</td>
</tr>
<tr>
<td>skos:narrower</td>
<td>Rice Fields</td>
<td>Hierarchical Semantic relation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Used to assert a direct (immediate) hierarchical link between two SKOS concepts.</td>
</tr>
<tr>
<td>skos:related</td>
<td></td>
<td>Associative Semantic Relation</td>
</tr>
<tr>
<td>skos:broadMatch</td>
<td></td>
<td>Hierarchical semantic relations between concepts defined in different concept schemes</td>
</tr>
<tr>
<td>skos:narrowMatch</td>
<td></td>
<td>Hierarchical semantic relations between concepts defined in different concept schemes</td>
</tr>
<tr>
<td>skos:exactMatch</td>
<td>Arable Land (EuroVoc Thesaurus)</td>
<td>Associative Semantic Relation Between two concepts belonging to different concept schemes - defined in SKOS model as sub property of skos:closeMatch</td>
</tr>
</tbody>
</table>

In this section, a general framework for adding semantics to land cover concepts has been described. After a short description of existing knowledge models, challenges associated with capturing domain concepts in a systematic and structured way have been described. Currently, the developed knowledge base includes only land cover concepts defined by the CORINE Land Classification System and their mapping against environmental concepts defined in the EuroVoc Thesaurus. A comprehensive integration with other thesaurus together is also foreseen in the future.
4. CONCLUSION

In this paper we describe an approach towards adding semantics to land cover data, by developing a formalized knowledge base and addressing once again the need to interlink knowledge bases. An integrated framework design is described as solution to add semantics to spatial content with a general goal to enable:

- flexibility and extendibility (adding new concepts or for incorporating new knowledge bases to the general ontological framework), and
- interoperability by using a standardized knowledge modeling approach (SKOS)

Described model adds semantics to land cover data sets by defining hierarchical and associative relations between concepts (explicit specification of general-specific relation between concepts or the relation between concepts defined within different concept schemas).

Despite enormous efforts dedicated to land cover harmonization, spatial information on land cover still faces semantic heterogeneity challenges. Considering the efficiency of SKOS implementation strategies in other domains like geology (Ma, 2011), we assume that the described approach lays the conceptual knowledge foundation for improving land cover datasets discovery, sharing and usage. Future work includes further extension of the developed knowledge base and concepts interlinking with other knowledge bases such as GEMET and AgroVoc. As a further step this will also be technically integrated into the search capabilities of the GI Research Cluster Geoportal for enhancing the effective information retrieval of disparate spatial resources and improving discovery experience.

REFERENCES


Web-based articles

Alexandria Digital Library Feature Type Thesaurus, http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/, [accessed 18 September 2011]


LiSA (Land Information System Austria), http://www.landinformationsystem.at, [accessed 2 September, 2011]


SWRL (Semantic Web Rule Language), [http://www.w3.org/Submission/SWRL/], [accessed 12 October 2011]

Proceedings


Books

