1. Introduction

Because of the advances in information, communication and sensing technologies, we are witnessing a constant growth in the amount of globally available geospatial information. This growth is driven by a necessity of different stakeholders (citizens, public institutions, scientists) to share information. Since different stakeholders generate geospatial information simultaneously and make it available over the Internet, the diversity of the geospatial data and geospatial data sources becomes inevitable. The diversity introduced in this way poses great challenges for individuals who are trying to discover and assemble geospatial data from heterogeneous and distributed geospatial data sources.

Since the emergence of (national) spatial data infrastructures ((N)SDI) (The European Parliament and The Council of the European Union, 2007), the problem of discovering and accessing geospatial data assembled from heterogeneous and distributed geospatial data sources gains importance. As access is foreseen as a central component in an (N)SDI, significant effort has been put into the development of geoportals - a single points of discovery and access to information within (N)SDI. The primary objective of geoportal is rather simple: among all referenced data (services, geospatial layers, documents, etc), geoportal users need a mechanism to easily find (discover) what they are searching for – using their own words, their own language (Tellez-Arenas, 2009). Discovery process should include both data description (metadata) and content. However, the discovery process details hidden behind these sentences are much more complex. Hence, geoportal usability issues emerge (Aditya and Kraak, 2005; Resch and Zimmer, 2013).

Recent geoportal research and development report usability issues from the standpoint of individuals who do not belong to Geographic Information System (GIS) world, also referred to as “non GIS professionals” (Aditya and Kraak, 2005). As stated in (Tellez-Arenas, 2009), “non professionals require rapid performance, good background information, and simple but elegant GUI”. In case of geospatial data discovery, this request has proven to be a difficult one to achieve using current geoportal architecture. Contemporary geoportals relay on the usage of metadata catalogues (usually the OGC standard CSW (Catalogue Service)) (Bernhard et al., 2013; Sakkopoulos et al., 2012; Salas et al., 2012; Bernard et al., 2005) which results in their inability to: obtain relevance percentage of the returned results, search with approximation of the spelling, classify the result using keywords or analyse relationships between words used for geospatial data discovery (Tellez-Arenas, 2009). The described state inspired the research in this paper. Our research addresses a part of these problems by presenting a novel geospatial data discovery methodology which can be effectively used to solve some of the geoportal usability issues. In particular, the implementation of our methodology enables users to discover heterogeneous geospatial data source by simply providing a natural language of geo-information they are interested in. Our research started with an analysis of the geo-information dissemination infrastructures which provide the ability to resolve geospatial data source heterogeneity. Geoportal is foreseen as a front end of these infrastructures, thus the methodology we propose enhances geoportal
usability by simplifying the discovery of geospatial data sources within these infrastructures.

Over the years, scientists and engineers have struggled to develop an interoperable geoinformation dissemination environment by taking two seemingly different approaches: syntactic standardization and semantic annotation of Web-accessible geo-information sources (Martin et al., 2004; Lutz et al., 2009; Meditskos and Bassiliades, 2010), and the development of ontology-driven geoinformation integration architectures (Hakimpour, 2003; Buccella et al., 2009).

As concerns syntactic standardization, probably the most prominent contribution was given by Open Geospatial Consortium (OGC) in the form of standards the geo-information Web services and data structures (Open Geospatial Consortium, 2006; Open Geospatial Consortium, 2011; Open Geospatial Consortium, 2005a; Open Geospatial Consortium, 2007). Research community devoted lots of attention to enhancing geospatial data source discovery in architectures relaying on these standards through means of semantic annotation of geospatial data and Web services (Open Geospatial Consortium, 2009; Kuhn, 2003; Buccella et al., 2009). Even in cases where ontologies are used for semantic annotation, recent researches like (Tian and Huang, 2012) have demonstrated that these approaches are still highly dependent of transformation between UDDI and languages used for ontology development and can suffer from restricted semantic reasoning capabilities. If geportals would rely solely on this type of architecture, these characteristics indicate that geportals would suffer from reduced geospatial data source discovery capabilities. It is our opinion that a solution for these shortcomings can be found among ontology-driven geoinformation integration architectures.

Ontology-driven geoinformation integration architectures provide powerful semantic reasoning capabilities and utilize ontologies for the discovery and retrieval of geo-information (Wache et al., 2001; Buccella et al., 2009), whereas the process of geoinformation retrieval utilizes connections (mappings) between ontologies and geo-information sources (Bizer, 2003; Cullot et al., 2007; Baglioni et al., 2007; Stanimirović et al., 2012). This form of geoinformation retrieval can be also observed as a part of ontology-based data access (OBDA) paradigm that tackles the problem of accessing data sources with a complex structure. Geportals relying on these architectures, e.g. on architectures relying on OBDA, could benefit in terms of their usability by taking advantage of powerful discovery mechanisms these architectures proclaim. However, there still seems to be a gap between the way geportal users (e.g. “non GIS professionals”) may perceive the underlying geospatial data, and the way this data is conceptualized from the standpoint of GIS professional developing the semantic description (e.g. ontology) of the same data. This can lead to a situation in which users may not know what keywords they should use to discover appropriate geo-information. The main contribution of the methodology we propose in this paper is to fill this gap by allowing users to discover geospatial data sources using their own words, which will in turn lead to geportal usability improvement. Our proposal is based on a number of best of the breed existing mechanisms combined to form a novel methodology used for geportal usability improvement. The methodology described in this paper utilizes terms extracted from a natural language description of geo-information, disambiguates the terms through means of a combination of unsupervised word sense disambiguation methods and matches their sense with the sense of the ontology concept names. The proposal we present in this paper can be equally applied on both domain and local ontologies. We consider local ontologies as a combination of domain ontology and task ontology in order to fulfill the specific purpose of an application (Fonseca et al., 2000). Our methodology simplifies
geospatial data discovery and can be easily implemented because it uses ontological components of the underlying architectures in their original form. It does not require the existence of semantic annotation of geospatial data source interface and can be implemented as a separate engine/Web service.

The rest of this paper is organized as follows. Section 2 discusses related work involved in semantic annotation and integration of geospatial information. In section 3, related work considering word sense disambiguation (WSD) is discussed. In particular, section 3 discusses unsupervised WSD methods based on the use of computational lexicons. Section 4 describes the methodology for enhancing discovery of geospatial data sources in ontology-driven geo-information integration architectures. The details of the prototype implementation and several evaluation results are given in Section 5. Section 6 concludes with an outlook to future work.

2. Related work

Contemporary Geographic Information Systems (GISs) rely on data from distributed geo-information sources to provide users with a single uniform access point over the refined data, whereas the single access point is commonly implemented in the form of a geoportal (Tellez-Arenas, 2009). While searching for the data they are interested in, GIS users expect to be provided with search results in the form of homogeneous data set(s) shorn of details regarding the data origin. This introduces a necessity to perform integration of data and computation resources belonging to several autonomous systems (Calvanese and Giacomo, 2005), which leads to the development of a federated geo-information system (Sayar A., 2009). To be able to integrate data originating from distributed and heterogeneous geospatial data sources, federated GISs need to overcome the problem of semantic heterogeneity between different geospatial data sources (Hakimpour, 2003; Buccella et al., 2009). Within different (geo-)information integration architectures and federated (geo-)information systems, semantic heterogeneity problems are most commonly solved through means of ontologies (Gruber, 1993; Buccella et al., 2009).

Regardless of the approach used for the development, if a (geo-)information integration architectures is to provide a single access point used for data discovery, it has to be capable of solving two additional problems: develop mappings between (global/local) ontologies and information sources (W3C RDB2RDF Incubator Group, 2009; W3C RDB2RDF Working Group, 2012), and define discoverable Web interfaces which represent information source access points (Meditskos and Bassiliades, 2011; Tian and Huang, 2012). If distributed geo-information sources can be discovered by utilizing their Web interfaces and integrated through means of their semantic description (mostly ontologies), then mappings between ontologies and information sources can be used to retrieve integrated data. If a geoportal would be implemented as a front-end application of an architecture that fulfils these prerequisites, in that case a geoportal would be capable of providing users with integrated data originating from heterogeneous sources. It is our opinion that this ability would significantly improve the usability of a geoportal.

Although problems encountered in the process of developing mappings between (global/local) ontologies and information sources and defining discoverable geo-information source Web interfaces belong to the same scope, they have mostly been investigated independently: the former within the development of ontology-to-relational database (RDB) mapping methodologies, ontology matching and integration (W3C RDB2RDF Incubator Group, 2009; Malhotra, 2009; Auer et al., 2009; W3C RDB2RDF Working Group, 2012; Euzenat and Shvaiko, 2007) and the latter within the development of a semantic description of geospatial Web services and Web service discovery systems (Wache et al., 2001; Euzenat and Shvaiko, 2007; Meditskos and Bassiliades, 2010; Tian and Huang, 2012). However, a majority of proposals from both
groups is missing explicit support for geospatial data, or at least an empirical study which confirms that a proposal can be equally applied within geospatial domain. Another common characteristic of these proposals is lack of an explicit definition of a discoverable Web geospatial service interface used to access the mapped data source. Both advantages and disadvantages introduced by any of solution proposals for each of the aforementioned problems are directly reflected onto the mechanism that can be implemented for information discovery. Thus, both advantages and disadvantages directly influence the usability of a geoportal. Since geoportal usability improvement is an overall goal of the methodology described in this paper, we will describe the reasons why we have decided to use geo-information integration architecture over similar proposals by briefly describing the shortcomings of similar proposals.

Previously reported approaches have proven that semantic annotation of geospatial information and services can be utilized as mean for overcoming semantic heterogeneity problems and enabling geospatial data source discovery (Wache et al., 2001; Euzenat and Shvaiko, 2007; Meditskos and Bassiliades, 2010). A number of proposals on how to enrich geospatial Web services with semantics using ontologies have been published (Martin et al., 2004; Lutz et al., 2009; Cabral et al., 2004). Mostly, these proposals perform semantic enrichment of geospatial Web services by generating explicit relationship between the data schema and domain ontologies. A majority of domain ontologies is modelled using formalized languages like Web Service Modeling Ontology (WSMO) or OWL-S (Meditskos and Bassiliades, 2010; Klien et al., 2007; de Bruijn et al., 2005). In the GIS domain, significant effort has been put into semantically annotating OGC Web services by means of linking OGC capabilities documents to ontologies (Open Geospatial Consortium, 2009). In addition, OGC released a standard used to support representing and querying geospatial data on the Semantic Web - OGC GeoSPARQL (Open Geospatial Consortium, 2012). This standard defines a vocabulary for representing geospatial data in RDF. It defines an extension to the SPARQL query language for processing geospatial data. Although these proposals represent a significant research contribution, there seems to be a lack of discovery systems developed on the bases of these research results so the usability of this work is yet to be discovered.

Other than utilizing Web service ontologies, the development of discovery systems in GIS domain has taken a different path recently. In a majority of implementations, reported systems discover OGC services through UDDI interface (Lieberman et al., 2003) using service catalogs (Stock et al., 2010). Since first implementations had to support for semantic querying, service catalogs have grown into semantically-enhanced service registries. Such registries use mappings of OWL, OWL-S and WSMO constructs to UDDI data structures (Goodwin et al., 2007; Meditskos and Bassiliades, 2011; Herzog et al., 2004). Although these services enhance the capability to discover geospatial Web services, they also expose some weaknesses. Reported systems use different standards (OGC capabilities XML document, OWL, OWL-S, WSMO and UDDI) within different architectural tiers to express Web service characteristics, resulting with them being highly dependent on transformation between these standards. This makes semantic reasoning over the entire architecture very hard to implement.

Also, there are few systems, like the one proposed by (Tian and Huang, 2012), which perform service matchmaking through semantic reasoning. Even if it exists, service matchmaking through semantic reasoning is restricted to reasoning over domain ontology. For example, (Tian and Huang, 2012) is restricted to a custom lightweight domain ontology. Though (Tian and Huang, 2012) make a step forward in service catalog development, their system does not have the ability to use and reason over existing or third-party developed ontology, or to become usable
in architectures based on multiple or hybrid ontology approach. Furthermore, the usage of word sense disambiguation (WSD) methodologies to aid services discovery, or at least the usage of computational lexicons for the same purpose, was not found by the authors of this paper in the reported systems. It is our opinion that WSD algorithms have the potential to enhance service discovery by being integrated into this process as an intermediate task. For this reason, our methodology envisions the use of a combination of unsupervised word sense disambiguation methods for geospatial data source discovery.

3. Word Sense Disambiguation – unsupervised methods based on the use of computational lexicons

Word Sense Disambiguation (WSD) is considered to be one of the core tasks in Natural Language Processing (NLP) (Navigli, 2009). The aim of WSD is to assign for each word of a text the appropriate sense(s). Algorithms used to perform WSD are classified into supervised and unsupervised methods. A supervised algorithm depends on labelled training data to compare information, whereas the unsupervised method does not. Aside from this classification, (Zesch & Gurevich, 2010) proposed another classification of WSD methods: path-based, information content based, gloss based, and vector based.

A majority of WSD methods use external knowledge sources as fundamental components to perform WSD. There is a variety of resources used as external knowledge sources: corpora of texts, computational lexicons, thesauri, glossaries, ontologies etc. An external knowledge source that has been widely used in the context of WSD is WordNet (Fellbaum, 1998). WordNet is a computational lexicon organized over ‘synonym sets’ (synsets). The latest version of WordNet (3.1) is available online and it contains over 155000 terms for 117000 synsets. Each synset represents a structure, which contains a term (word), its class (verb, noun, adjective etc.) and connections to all semantically related words along with a brief definition (‘gloss’) illustrating the use of the synset members. Semantic relations defined for a single synset apply to all its members. Among the semantic relations, the most frequently used ones are the following: hypernymy (also called kind-of or is-a), hyponymy (the inverse relations of hypernymy), meronymy (also called part-of) and holonymy (the inverse of meronymy).

WordNet can be effectively used within a majority of unsupervised WSD methods, which introduce semantic similarity measures used to perform word disambiguation. Usually, these methods use semantic relations defined within WordNet to determine the similarity between terms (synsets). For example, path-based methods measure the length of the path between two words in a graph-like structure and WordNet can be used as a resource which supplies the paths. A number of path-based methods, such as (Rada et al., 1989; Leacock & Chodorow, 1998; Wu and Palmer, 1994), have successfully utilized WordNet as a graph-like structure to perform word similarity measurement. Furthermore, information-based methods use tree structure (WordNet) to measure how much information the two words share, by measuring the number of nodes in the tree structure that the two words have in common. Information-based methods reported in (Resnik, 1995), (Jiang & Conrath, 1997) and (Lin, 1998) modify this idea to define similarity measures which indicate the probability of a word appearing in a corpus. WSD methods, which further exploit semantic relationships contained in WordNet, can be found among gloss-based and vector-based methods. As a starting point, both gloss-based and vector-based methods often use an approach defined by (Lesk, 1986) that relies on word definition usage. This approach determines the similarity of all the words in the two word definitions and measures the similarity by overlapping the two word definitions. This approach was successfully adapted by (Patwardhan et al., 2002), (Patwardhan, 2006) and (Zesch and
Gurevich, 2010) to create approaches which make a better use of semantic relations (synonym, hypernym, hyponym, holonym, meronym, troponym, and attribute) defined within WordNet.

4. Geospatial data source discovery approach

The approach we have developed is focused on geoportals that rely on federated geo-information systems as their spatial data infrastructure. In particular, federated geo-information systems utilized in our methodology resolve semantic heterogeneity through means of meta-data, used to describe geospatial data sources. Although it can be adapted for different meta-data, our approach was intended to be used within federated GISs which utilize ontological components, e.g. domain/local ontologies, for geospatial data integration purposes (Buccella et al., 2009). In such systems, domain/local ontology elements are usually matched against geospatial data sources through means of different mapping files or schema transformations (Stanimirović et al, 2012; Tian and Huang, 2012). Thus, geospatial data sources become discoverable at the level of ontology elements. In particular, our approach is used to determine an appropriate sense of a user-defined geospatial data search description, and match this description against global/local ontology concepts. Figure 1 illustrates a general architecture of a system in which our approach can be implemented.

Figure 1: The environment used for geospatial data source discovery

The main facilities of the system are described as follows:

- Natural language terms – a set of terms extracted from a natural language description of geo-information, given by a user.
- Federated Geographic Information Systems (Federated GISs) – GISs which relay on data from distributed and heterogeneous geo-information sources; these systems utilize meta-data to overcome semantic heterogeneity problems and enable geospatial data source discovery.
- Meta-data – data used to describe the content of different geospatial data sources. This data can be stored in different forms: semantic annotations of geospatial data and services, ontological components (local/global ontologies), UDDI documents, OGC capabilities documents etc. Each federated GIS maintains its meta-data within a central meta-data repository.
- Meta-data repository – a repository capable of storing different forms of meta-data for each of the federated GISs; it is an optional component, e.g. it can be implemented if federated GISs use hybrid ontology approach to overcome semantic heterogeneity problems, or omitted if each part of federated GIS stores meta-data locally.
- Computational lexicon – structured machine-readable lexicon of terms; it is used as a knowledge source to associate the most appropriate senses with terms (words) given by the user.
- Geospatial Data Source Discovery Engine – a stand-alone geospatial data source discovery module. It implements the discovery process by matching federated GIS meta-data elements with user-defined geospatial data description; for this purpose, discovery engine utilizes computational lexicons as knowledge sources.

According to the aforementioned focus of our approach, the core of the geospatial data source discovery is a matching process, based on a similarity measurement. The problem of similarity measurement among geospatial concepts has been previously studied in the geospatial domain (Rodríguez and Egenhofer, 2004; Janowicz et al., 2011). In the methodology we propose, similarity measurement process is performed between
terms extracted from the user-defined geospatial data description and global/local ontology concepts. The matching process is performed by Geospatial Data Source Discovery Engine and starts with extracting terms from user description. After stripping the extracted terms from suffixes, Geospatial Data Source Discovery Engine will load a global/local ontology of a federated GIS, extract ontology concepts and determine similarity between the extracted terms and the ontology concepts. The similarity measurement is based on the use of a combination of unsupervised word sense disambiguation methods, which utilize WordNet computational lexicon. The global/local ontology concepts, whose similarity with user-defined terms exceeds a predefined threshold value, will be added to a result set.

In the rest of this section, we will go into the details of the approach we have defined for the purpose of geospatial data source discovery. Our approach will be presented in the form of an algorithm which uses user-defined geospatial data description as its input. As the output, this algorithm will determine a set of (domain/local) ontology concepts mapped to geospatial data sources.

Step 1: Disambiguate user-defined description of geospatial data
The algorithm starts with utilizing natural language description of geo-information. This description, defined by the end users, is tokenized into a list of words through means of a regular expression (Figure 2a). Afterwards, the correct part of speech (noun, verb, pronoun or adverb) is identified for each of the words by utilizing WordNet computational lexicon (Figure 2b). The identification process is based on the usage of WordNet library functions, starting with “findtheinfo” method as a primary search function for WordNet database. Also, the identification process includes the stripping of suffixes from words which we have implemented according to algorithm proposed in (Porter, 1980). At the end of step 1, words identified as nouns are extracted to a separate term set, which will be referred to as ‘input term set’.

Step 2: Expand the input term set through means of WordNet computational lexicon
Among the semantic relations defined in WordNet, this algorithm utilizes synonymy, hypernymy and hyponymy semantic relations. For each term from the input term set, a new term set is created and it consists of the term’s synonyms, hypernyms and hyponyms (Figure 3). Each of the newly created term sets will be referred to as ‘expanded term set’.

Step 3: Repeat steps 4 to 7 for each of the expanded term sets

Step 4: Create a term set which contains the names of (domain/local) ontology concepts
This step can be equally applied to both domain and local ontologies which are mapped to geospatial data sources. Regardless of its position in the geo-information integration architecture, the ontology is loaded, and traversed for the purpose of extracting concept names. Each of the concept names is added to a new term set which will be referred to as ‘concept term set’ (Figure 4). The result of this step is cached e.g. the concept term set is created during the first iteration and reused in the following ones.

Step 5: Initialize a result container
As previously stated, the output of the algorithm is a set of (domain/local) ontology concepts mapped to geospatial data sources. Therefore, each item within the result container represents a structure which consists of two elements:
1. an ontology concept whose similarity to expanded term set terms is being measured,
2. a dictionary of semantic similarity measurement results – each item in the dictionary is a key-value pair, whereas key represents a term from the expanded term set while the value represent a similarity measured between the term and the ontology concept (Figure 5)

Figure 5: An example of a dictionary created for an ontology concept named ‘Road’

Step 6: Semantic similarity measurement
For each pair of terms $T_{EX}$ and $T_{C}$, $T_{EX}$ being a term from the expanded term set and $T_{C}$ being a term from the concept term set for the current iteration, perform the geospatial data source discovery process by repeating the following steps:

1. Measure a similarity between the terms $T_{EX}$ and $T_{C}$
   a. Compute edit distance similarity for terms $T_{EX}$ and $T_{C}$
      Edit distance similarity is measured according to Levenshtein distance (Levenshtein, 1966). The Levenshtein distance calculates the least number of edit (insert, delete) operations that are necessary to modify one string to obtain another string. Edit distance similarity is given by
      \[
      dist_{T_{EX},T_{C}}(i_{T_{EX}}, j_{T_{C}}) = \begin{cases} 
      \max(i_{T_{EX}}, j_{T_{C}}), & \text{when } \min(i_{T_{EX}}, j_{T_{C}}) = 0 \\
      \min\left( \begin{array}{c}
      \text{dist}_{T_{EX},T_{C}}(i_{T_{EX}} - 1, j_{T_{C}}) + 1 \\
      \text{dist}_{T_{EX},T_{C}}(i_{T_{EX}}, j_{T_{C}} - 1) + 1 \\
      \text{dist}_{T_{EX},T_{C}}(i_{T_{EX}} - 1, j_{T_{C}} - 1) + 1 \end{array} \right), & \text{otherwise.}
      \end{cases}
      \]
   b. Compute semantic similarity $sim(T_{EX},T_{C})$ between the terms $T_{EX}$ and $T_{C}$ according to the algorithm described in (Wu and Palmer, 1994). (Wu and Palmer, 1994) approach measures the path length to the root node from the least common subsumer (LCS) of the two concepts compared. The least common subsume can be interpreted as a concept in a lexical taxonomy (in this case WordNet), which has the shortest path length from the two concepts compared. The measured distance between the root node and the least common subsumer of the two concepts is scaled by the sum of the path lengths from the individual concepts to the root. According to this algorithm, $sim(T_{EX},T_{C})$ is determined by considering the depths of the $T_{EX}$ and $T_{C}$ synsets in the WordNet computational lexicon, along with the depth of their least common subsumer (LCS). The LCS of synsets $T_{EX}$ and $T_{C}$ is the most specific synset that is an ancestor of both synset $T_{EX}$ and $T_{C}$.
      \[
      sim(T_{EX},T_{C}) = \frac{2*\text{depth(LCS}_{T_{EX},T_{C}})}{\text{depth}(T_{EX}) + \text{depth}(T_{C})}
      \]
   c. Determine final semantic similarity according to the following equation:
\[
\text{semsim}(T_{EX}, T_C) = \max(\text{dist}(\text{length}(T_{EX}), \text{length}(T_C)), \text{sim}(T_{EX}, T_C))
\]

2. For this method a fixed constant has to be defined as a similarity threshold, usually 0.8 or greater. If the measured semantic similarity exceeds the threshold value, populate the result container in the following way:
   a. If the ontology concept \( C \), which corresponds to the term \( T_C \), doesn’t exist in the result container, create an item in the result container as defined in step 5.
   b. Create a dictionary item as a key-value pair, whereas the key represents the term \( T_{EX} \) while the value represents \( \text{sim}(T_{EX}, T_C) \), and add it to the dictionary which corresponds to the ontology concept \( C \) (Figure 5).

**Step 7: Ordering the results**

Results will be ordered according to the following criteria:

a. Precedence will be given to ontology concept whose name was determined to be similar to the largest number of terms from the input set.

b. If several ontology concepts were determined to be similar to the same number of terms from the expanded term sets, precedence will be given to ontology concept which has the highest average semantic similarity value, where this value is computed for each concept as an average of all outputs from step 6.c.

**Evaluation**

The solution presented in this paper is implemented as a part of our GeoNis framework. GeoNis is a framework for interoperability of GIS applications that have to provide infrastructure for data interchange in the local community environment (Stoimenov, 2002). The GeoNis framework was developed to perform the intelligent, ‘on-demand’ integration of information from multiple heterogeneous GIS (spatial and geographic) and nonspatial (thematic) data sources, which consist of local community services and offices that own geodata in some format. Semantic interoperability in GeoNis, resolved by Semantic Mediator (Stoimenov, 2004), is the ability of sharing geospatial information at the application level, without knowing or, understanding terminology of other systems.

In GeoNis Semantic Mediator we propose a semantic based integration approach that uses multiple ontologies, instead of an integrated view (Stoimenov, 2005a). Our Semantic Mediator uses hybrid ontology approach. Our solution is to formally specify the meaning of the terminology of each Geoinformation Community (GIC) (i.e. local service or office) using local ontologies and to define a translation between each GIC terminology (local ontology) and shared domain terminology (in top-level ontology). In this context, ontologies are virtually linked by inter-ontology relationships, which are then used to indirectly support query processing. Semantic Mediator provides a methodology and software support for semantic mismatches (conflicts) resolving between terminologies. This methodology uses the defined ontology mappings between each community terminologies and a top-level ontology or the common data model (reference ontology).

GeoNis ontology (local or domain) is an abstraction of domain of interest \( D \), represented by triple \( O = (C, R, is-a) \), where \( C = \{c_i | i = 1, n\} \) is a set of concepts, \( R = \{r_i | i = 1, n\} \), is a set of binary typed roles (or relations) between concepts, and is-a is a set of inheritance relationships defined between concepts. In an (local or top-level) ontology, inheritance relationships define a partial order over concepts and carry subset semantics. Set of semantic relations between concepts defines semantics of concepts and their relevance. We
have defined the following set of relations between concepts in ontology:

\[ R = \{ \text{synonym}, \text{hypernym}, \text{hyponym}, \text{meronym}, T \} \]

where \( T \) is a set of ‘topological’ relations. More details about formal definition of GeoNis ontologies are given in (Stoimenov, 2003) and (Stoimenov, 2005b).

**Figure 6: The position of Geospatial Data Source Discovery Engine in GeoNis framework**

For the evaluation purposes of the proposed algorithm, GeoNis framework was extended with a software component named Geospatial Data Source Discovery Engine (Figure 6). As stated in its name, this component facilitates the discovery of existing geospatial data source to GIS application users within GeoNis environment. For these purposes, Geospatial Data Source Discovery Engine utilizes local ontologies and domain ontology implemented for GeoNis environment, as well as a natural language description of geo-information defined by users (this process is described in section 4). During the implementation of Geospatial Data Source Discovery Engine prototype, the usage of GeoNis ontologies, implemented according to formal definitions given in (Stoimenov, 2003) and (Stoimenov, 2005b), has proven to be highly complex. Since we were performing feasibility testing, we have decided to use a different approach, which is in accordance with ontology standardizations efforts in geospatial domain (Lieberman, 2007).

To evaluate our research efforts we tried to implement ‘lightweight’ version of GeoNis ontology. This ‘lightweight’ version of GeoNis ontology is implemented in accordance with formal definition of GeoNis ontology, is interoperable with full geospatial ontologies and takes into account ontology standardizations efforts in geospatial domain. ‘Lightweight’ GeoNis ontology is a rather simple ontology that defines some of the geospatial features and a number of spatial relationships between them. The term ‘geospatial feature’ in this ontology refers to any entity with an inherently or indirectly associated spatial dimension. Expressiveness of formalization of implemented ontology is limited using ontology OWL Lite representation language (W3C, 2009). After evaluation of different geospatial ontologies, for implementation of ‘lightweight’ GeoNis ontology, we decided to reuse GeoOWL standard (Lieberman, 2007) and Geonames feature type hierarchy (GeoNames, 2012).

**Figure 7: Geometry definition in GeoNis ontology**

W3C Geospatial Incubator Group developed GeoOWL (Lieberman, 2007) as a minimum geo-vocabulary which follows GeoRSS guidelines (GeoRSS, 2013). ‘Lightweight’ GeoNis ontology imports GeoOWL as the core ontology, thus supposing a Simple GML Geometry definition and an easy alignment with existing GeoRSS feeds (Figure 7). The central part of the GeoOWL ontology (Lieberman 2007) is a object property geo:where which takes as its domain any OWL/RDF class that it makes sense (after ISO 19109) to cast as a geographic feature. This property takes as its range the abstract class _Geometry. Subclasses of _Geometry include gml:Point, gml:Linestring, gml:Polygon, and gml:Envelope after the corresponding GML objects. The properties of these classes are a subset of the corresponding properties defined in the GML model and schema. This represents GeoRSS GML. Other subproperties of geo:where represent GeoRSS Simple and include geo:Point, geo:Line, geo:Polygon, geo:Circle, and geo:Box. These properties each take a literal list of doubles as their range, but are equivalent in definition to (are a shorthand for) geo:where plus the corresponding GeoRSS GML classes and their properties. For backwards compatibility, geo:lat and geo:long are retained as subproperties of geo:where.

GeoOWL just offers two free text tags for representing feature types and feature relations, and it expects the deployment of folksonomies from this tagging. In our scenario we need some complex classification schemas.
For this reason, we decided to reuse Geonames ontology (GeoNames, 2012). Geonames ontology provides a rich taxonomy of feature types based in SKOS thesauri (W3C, 2004). We have defined all concept schemes from Geonames in ‘lightweight’ GeoNis ontology as subclasses of GeoNis_Feature class (Figure 8). Figure 8 shows a fragment of GeoNis taxonomy presenting only the particular subclasses that are used in examples outlined in this paper. This design allows the alignment with Geonames database, with more than 6.5 million features.

Feasibility testing
To evaluate the feasibility of the proposed algorithm within the GeoNis framework, a prototype Geospatial Data Source Discovery Engine was developed. The prototype, shown on Figure 9, represents a stand-alone desktop application which implements the algorithm described in section 4. The implementation was performed by utilizing Microsoft Windows Forms as a graphical application programming interface (API), included as a part of Microsoft .NET Framework 4.0 (Microsoft, 2010). The functionalities of the implemented prototype will be presented through a simple walkthrough which consists of five steps. Each walkthrough step will be explained through details considering three components: input, discovery algorithm step(s) implemented within the current walkthrough step, and output. On Figure 9, user interface (UI) parts conforming to each walkthrough step are placed in a separate frame. Frame numbers (1 to 5) conform to walkthrough steps.

For the purpose of simplifying feasibility evaluation, discovery example will be conducted and presented under the following assumptions:

- User describes the data he/she is looking for as ‘streets and rivers in Serbia’.
- First step of our algorithm (Step 1 in Section 4) has partially been performed e.g. the description given by the user has been tokenized into a list of words and each word has been stripped of suffixes. For this purpose, the prototype application uses an implementation of the Porter stemming algorithm (Porter, 1980) developed by Leif Azzopardi (Azzopardi, 2002). The output is the following list of words: {street, river, Serbia}.
- The prototype application will be given only the word ‘street’. Thus, the discovery process will be performed on the basis of a single term.

Walkthrough step 1 (frame 1 on Figure 9):
- Input: word ‘street’, which was previously stripped of suffixes; user will activate the ‘search’ button within prototype application UI.
- Discovery algorithm step: a part of step 1; word ‘street’ is identified as a noun with five different descriptions in WordNet lexicon.
- Output: word ‘street’ is identified to be a term ‘street’ in WordNet lexicon.

Walkthrough step 2 (frame 2 on Figure 9):
- Input: WordNet term (noun) ‘street’; user will activate ‘create term set’ button within prototype application UI.
- Discovery algorithm step: step 2; an expanded term set is created for term ‘street’; it consists of synonyms, hypernyms and hyponyms discovered in WordNet lexicon for the term ‘street’.
- Output: expanded term set contains 66 terms: {artifact, whole, thoroughfare, abstraction, entity, gathering, abstract_entity, attribute, physical_entity, object, chance, road, street, opportunity, physical_object, possibility, route, way, neighborhood,
Walkthrough step 3 (frame 3 on Figure 9):

- Input: GeoNis.owl ontology; user will activate ‘load ontology’ button within prototype application UI and will be provided with a dialog used to select the appropriate ontology from the file system.
- Discovery algorithm step: step 4; concept term set is created by traversing GeoNis ontology.
- Output: concept term set:

  - {Accommodation, Cable, Road_Feature, Substation, Taxi_Stop, PopulatedPlace_Feature, Shopping, Hotel, Main_Street, Vegetation_Feature, Food, Consumer_Connection, Area_Feature, Electric_Utility, Bus_Stop, Hydrographic_Feature, Population_Feature, Utility_Feature, Road, Street, Building_Feature, Semantic_Topological_Relation, PointOfInterest_Feature, Administrative_Feature, Hostel, Local_Street, GeoNis_Feature, Sight, Service}; once the term set is created, prototype application will report that the chosen ontology has been successfully loaded by displaying the name of the loaded ontology, as shown in frame 3 on Figure 9.

Walkthrough step 4 (frame 4 on Figure 9):

- Input: expanded term set created for the term ‘street’ and concept term set created for GeoNis ontology; user will activate ‘perform discovery’ button within prototype application UI.
- Discovery algorithm step: step 5 and step 6; a result container is initialized and semantic similarity measurement is performed for each pair of terms from the expanded term set and concept term set; these term sets were previously created in walkthrough steps 2 and 3, respectively.
- Output: the result container is populated for each pair of terms whose semantic similarity exceeds a similarity threshold, which was in this evaluation example set to 0.7; results are divided into the following three groups according to semantic similarity value (column SCORE in the grid placed within frame 4 on Figure 9):
  - Group 1: semantic similarity value between 0.7 and 0.8 – these pairs of terms are marked using yellow colour in the prototype application UI, as shown in frame 4 on Figure 9.
  - Group 2: semantic similarity value between 0.8 and 0.9 – these pairs of terms are marked using blue colour in the prototype application UI, as shown in frame 4 on Figure 9.
  - Group 3: semantic similarity value is 0.9 or higher – these pairs of terms are marked using red colour in the prototype application UI, as shown in frame 4 on Figure 9.

Walkthrough step 5 (frame 5 on Figure 9):

- Input: the result container populated in step 4; user will activate ‘display results
in semantic tree’ button within prototype application UI.

- Discovery algorithm step: step 7; discovered ontology concepts will be ordered according to criteria defined in the step 7 of the discovery algorithm.

- Output: discovered ontology concepts are the following: \{Street, Road, Main_Street, Local_Street, Service, Shopping, Cable, Accommodation, Food, Substation, Hostel, Hotel, Administrative_Feature, Point_Of_Interest_Feature, Sight, Hydrographic_Feature, Area_Feature\}; each of the discovered ontology concepts is displayed as a tree node; also, for each of the concepts, users can traverse the semantic tree of its parent ontology classes, as shown in frame 5 on Figure 9.

**Figure 10: Summarized discovery results**

Summarized discovery results are shown in Figure 10. This figure displays average semantic similarity computed for each ontology concept against all terms from the expanded term set. The number of terms whose similarity to an ontology concept exceeds predefined threshold value of 0.7 is also shown for each of the ontology concepts. Some of the results are rather straightforward. As expected, ontology concept ‘Street’ has the largest number of matches and the highest average semantic similarity. Similarly, ontology concepts ‘Main_Street’ and ‘Local_Street’ have been highly ranked mostly because a high level of semantic similarity was detected of the basis of edit distance similarity measurement. The discovery of ontology concept ‘Road’ reveals the good sides of the proposed algorithm. Although edit distance similarity for this concept was quite low, the usage of path-based method for semantic similarity measurement resulted in this concept becoming the second highest ranked concept within the analyzed ontology. The rest of the ontology concepts have been matched against a significantly lesser number of terms which indicates that this search should result in acquiring the data from (geo-)information sources mapped to ontology concepts \{Street, Road, Main_Street, Local_Street\}.

**An analysis of performances and discovery results**

Besides evaluating the feasibility of the proposed algorithm within the GeoNis framework, our proposal was also tested for the following purposes:

- determining the quality of the discovery results
- determining the most adequate similarity threshold value
- estimating an overhead the algorithm might cause in the GeoNis system

In order to achieve these goals, the proposed algorithm was tested against a testing term set. We have defined this term set as a set of terms that can be used for naming spatial data feature types. The testing term set used for the analysis was extracted from schemas belonging to the Spatial Data Standards for Facilities, Infrastructure, and Environment (SDSFIE) (Defense Installation Spatial Data Infrastructure Group, 2014). The same term set was used for the evaluation of performances of the developed prototype engine.

Each term belonging to testing term set was extracted as a key part of feature type name proposed by SDSFIE standard. In this way, a term set consisting of 37 terms has been extracted and used as a testing term set, as shown in Table 1. Although the term set has been extracted from an official standard, the adequacy of the selected terms was confirmed by examining the description of feature codes in Geonames ontology (GeoNames, 2012). This examination has shown that 81% (30 terms) appear as keywords within feature descriptions. Since we have defined all concept schemes from Geonames in ‘lightweight’ GeoNis ontology as subclasses of GeoNis_Feature class, we considered this term set to be adequate for the analysis of results and performances of the algorithm we propose.
The testing was conducted using GeoNis ontology through 37 runs – one per each term from the testing term set. During each run, an expanded term set was created for each term from the testing term set. As described in algorithm steps 4 to 7, the similarity was measured between each term from the expanded term set and each GeoNis ontology concept. Similarity measurement results, varying from 0.0 to 1.0, were divided into 10 groups whereas each group occupies a range of values with step 0.1 between groups (0.0-0.1, 0.1-0.2..., 0.9-1.0).

The output of all 37 runs was analysed for the purpose of determining the quality of the discovery results and the most adequate similarity threshold value. The first step of the analysis was a determination of the correct output in the form of a set of ontology concepts expected to be found similar to a particular term from the testing term set. The correct output was determined for each term from the testing term set as a union of the sets of ontology concepts defined as correct output by 3 domain ontology experts (the authors of the paper). During the analysis, proposed algorithm considers a term from the testing term set similar to an ontology concept if the expanded term set, created for the observed term, contains at least one term whose similarity value belongs to at least one of the groups whose lower range bound value is greater or equal to a predefined threshold value.

To determine the most adequate similarity threshold value, the following indicators were used:

- discovery rate (DR) – number of ontology concepts determined to be similar to a particular term
- false discovery rate (FDR) – number of false discoveries among ontology concepts determined to be similar to a particular term
- discovery rate matching domain expert expectations (DR-MDEE) – number of ontology concepts determined to be similar to a particular term and matching domain expert’s expectation to be similar to a particular term

These values were analyzed in the context of different similarity threshold values. The analysis results are given in Table 1. The result analysis has shown that in cases where the similarity threshold value is set under 0.7, the average percentage of false discoveries in some cases exceeds 70%. For similarity threshold value set to 0.7 and larger, the average percentage of false discoveries is significantly lower. For example, the average percentage of false discoveries will be decreased by 25.06% if the similarity threshold value is set to 0.8 instead of being set to 0.7. At the same time, the discovery rate matching domain expert expectations will be decreased by only 6.85% (from 62.39% to 55.61%). If the similarity threshold value is set to 0.9 instead of being set to 0.8, the average percentage of false discoveries will be decreased by 59.26%. However, the discovery rate matching domain expert expectations will be decreased by 25.06% which can lead towards significant loss of relevant discovery results. Therefore, we consider that similarity threshold value can be set to 0.8 without jeopardizing the quality of the discovery process while lowering the false discovery rate at the same time.

Aside from the previously described result analysis, the testing process was used to observe the time needed for the discovery process to be performed, as shown in Table 1. The observation was needed in order to estimate the overhead proposed algorithm might cause in the GeoNis system. The observation was conducted on the machine using Intel(R) Core(TM) i5-4440 CPU (3.10GHz) with 8GB of RAM on Windows 8.1 64-bit operating system. The application was considered to be a black box using 10 background working threads for each of 37 discovery runs. The time needed for a single run to be performed was analyzed in the context of two indicators: the number of WordNet synsets (SSN in Table 1) and the number of terms in the expanded terms set (ETSN in Table 1) used within the particular run. Terms with several distinct meanings are represented in as
many distinct synsets and all synsets are used during the discovery process. Since the proposed algorithm determines the depth and the least common subsumer for each synset of the term, whereas each synset has a different hierarchical structure in the lexicon, we expected the number of used synsets and their lexicon structure to highly influence the time needed to perform discovery. Also, we expected the increased number of terms in the expanded term set to negatively influence the discovery process performances.

The expectations we envisioned can be easily identified from the results shown in Table 1. The shortest time was needed to perform a test run for the term “marina” – 550ms. Within the WordNet, this term has only one synset. Also, the expanded term set created for this term consists of only 16 terms. On the other hand, the longest time was needed to perform a test run for the term “transportation” – 12297ms. This term has 6 synsets and the expanded term set created for this term consists of 113 terms. However, this is not the term with the largest number of synsets and expanded term set terms. Term “land” has 11 synsets and the expanded term set created for this term consists of 346 terms. This example demonstrates that the hierarchical structure in the lexicon of the used synset highly influences the algorithm performances. The average depth of synsets within WordNet for the term “transportation” is 8.3 while the average depth of synsets within WordNet for the term “land” is 5.9, which resulted in discovery process for the term “transportation” taking longer time to perform.

| DR-MDEE | DR (TH = 0.5) | DR-MDEE IN DR (TH = 0.5) | DR (TH = 0.6) | DR-MDEE IN DR (TH = 0.6) | DR (TH = 0.7) | DR-MDEE IN DR (TH = 0.7) | DR (TH = 0.8) | DR-MDEE IN DR (TH = 0.8) | DR (TH = 0.9) | DR-MDEE DECREASE (TH=0.5 vs. TH=0.8) | DR (TH=0.8 vs. TH=0.9) | DR-MDEE DECREASE (TH=0.5 vs. TH=0.9) | FDR DECREASE (TH=0.5 vs. TH=0.8) | FDR DECREASE (TH=0.5 vs. TH=0.9) | TIME (ms) | SSN | ETSN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMINISTRATION | 5 | 14 | 100.00% | 9 | 100.00% | 6 | 60.00% | 0.00% | 40.00% | 55.56% | 66.67% | 6206 | 6 | 81 |
| AGRICULTURE | 2 | 13 | 50.00% | 7 | 0.00% | 4 | 0.00% | 50.00% | 50.00% | 41.67% | 66.67% | 4433 | 4 | 136 |
| BRIDGE | 7 | 19 | 85.71% | 17 | 71.43% | 9 | 42.86% | 14.28% | 42.48% | 7.69% | 53.85% | 4094 | 9 | 226 |
| BUILDING | 5 | 17 | 80.00% | 16 | 80.00% | 8 | 60.00% | 0.00% | 20.00% | 7.69% | 61.54% | 4927 | 4 | 344 |
| CABLE | 3 | 17 | 33.33% | 12 | 33.33% | 6 | 33.33% | 0.00% | 0.00% | 31.25% | 68.75% | 2187 | 6 | 190 |
| CHANNEL | 3 | 20 | 15.00% | 17 | 17.65% | 13 | 23.08% | 0.00% | 0.00% | 17.65% | 41.18% | 5503 | 8 | 319 |
| CONSTRUCTION | 7 | 18 | 57.14% | 17 | 57.14% | 9 | 42.86% | 14.28% | 7.14% | 57.14% | 8587 | 7 | 499 |
| CULTURE | 3 | 15 | 100.00% | 10 | 100.00% | 7 | 100.00% | 0.00% | 0.00% | 41.67% | 66.67% | 3259 | 7 | 562 |
| DISPOSAL | 2 | 17 | 50.00% | 16 | 50.00% | 9 | 50.00% | 0.00% | 0.00% | 6.25% | 50.00% | 3295 | 4 | 587 |
| ENVIRONMENT | 8 | 17 | 50.00% | 15 | 50.00% | 8 | 37.50% | 0.00% | 12.25% | 15.38% | 61.54% | 1978 | 2 | 620 |
| FLOOD | 2 | 18 | 50.00% | 14 | 50.00% | 9 | 50.00% | 0.00% | 0.00% | 23.53% | 52.94% | 8159 | 6 | 97 |
| FOREST | 2 | 10 | 50.00% | 5 | 50.00% | 0 | 0.00% | 0.00% | 50.00% | 55.56% | 100.00% | 1701 | 2 | 47 |
| GAS | 4 | 24 | 100.00% | 22% | 75.00% | 17 | 25.00% | 25.00% | 75.00% | 5.00% | 20.00% | 10284 | 6 | 135 |
| HISTORY | 3 | 11 | 66.67% | 9 | 66.67% | 1 | 0.00% | 0.00% | 66.67% | 22.22% | 88.89% | 1938 | 5 | 53 |
| IMAGE | 4 | 18 | 100.00% | 14 | 50.00% | 8 | 25.00% | 50.00% | 75.00% | 14.29% | 50.00% | 11433 | 9 | 125 |
| LAND | 4 | 21 | 100.00% | 19 | 100.00% | 15 | 100.00% | 0.00% | 0.00% | 11.76% | 35.29% | 11458 | 11 | 346 |
| LEVEE | 3 | 18 | 66.67% | 17 | 66.67% | 11 | 66.67% | 0.00% | 0.00% | 6.25% | 43.75% | 1256 | 3 | 33 |
| MARINA | 3 | 11 | 66.67% | 8 | 66.67% | 0 | 0.00% | 0.00% | 66.67% | 33.33% | 100.00% | 550 | 1 | 16 |
| NATURE | 4 | 11 | 75.00% | 8 | 75.00% | 4 | 25.00% | 0.00% | 50.00% | 37.50% | 62.50% | 3755 | 3 | 57 |
| PLAYGROUND | 2 | 14 | 50.00% | 11 | 50.00% | 2 | 0.00% | 0.00% | 50.00% | 23.08% | 84.62% | 1156 | 2 | 32 |
| POLLUTION | 4 | 13 | 75.00% | 7 | 75.00% | 5 | 25.00% | 0.00% | 50.00% | 60.00% | 60.00% | 1669 | 3 | 37 |
| PUMP | 5 | 17 | 60.00% | 16 | 60.00% | 10 | 20.00% | 0.00% | 66.67% | 7.14% | 35.71% | 3227 | 3 | 55 |
| RAIL | 3 | 16 | 100.00% | 13 | 100.00% | 6 | 66.67% | 0.00% | 33.33% | 23.08% | 69.23% | 6756 | 5 | 83 |
Conclusion

For the purpose of enhancing geospatial data source discovery, this paper proposes a methodology used in ontology-driven geo-information integration architectures, which implement the retrieval of geo-information through means of connections (mappings) between ontologies and geo-information sources. The methodology core is a process of matching terms extracted from a natural language description of geo-information, defined by the end users, with the sense of the (domain/local) ontology concepts. The matching process utilizes a combination of unsupervised word sense disambiguation methods, based on the use of WordNet computational lexicon. The output of the matching process is a set of (domain/local) ontology concepts mapped to geospatial data sources. To evaluate our methodology, we developed a prototype desktop application which implements the matching process. Our methodology provides a number of benefits over methods that have been previously used for geospatial data discovery. These benefits can be summarized as follows:

- Simplify geospatial data discovery using natural language terms rather than requiring the users to fill any kind of interactive form
- The method can be implemented as a separate engine/Web service. Therefore, it is independent of the system’s architectural tier that implements geo-information data source interfaces.
- Doesn’t require the existence of semantic annotation of geospatial data source interface, such as UDDI records, OWL-S and WSMO elements. Thus, the problem of translation between different semantic annotation schemes is completely avoided.
- Ontological components of federated GISs are used in their original form. Therefore, this method utilizes existing ontologies while retaining their full expressiveness and reasoning capabilities.
- Previously presented evaluation confirms the feasibility of our approach within (geo-)information integration architectures based on single ontology approach. By utilizing interontology mappings (relationships between the ontologies), this method can be used within (geo-)information integration architectures based on multiple and hybrid ontology approaches.

Despite this methodology represents a feasible approach for discovering heterogeneous and distributed geospatial data.
sources, further efforts are required to make geospatial data sources fully discoverable. In future work we will address following issues:

- Enhance the discovery process by introducing natural language processing algorithm capable of parsing complex text expressions involving conjunctions, disjunctions, negations, context-specific indexicals and metaphors. This enhancement will be addressed by authors in the nearest future since we consider it to be very important. The level of importance can be observed in the following example. Currently, if a geoportal used would enter a description of information such as “not streets nor rivers”, our discovery methodology would produce the exact opposite of the expected – the result would be data sources containing streets and rivers since those words would the identified as nouns.

- Enhance the discovery process by introducing gloss-based word sense disambiguation methods. Currently, natural language terms are tokenized from a user-defined description of geospatial data, and for each of them an appropriate sense is determined. However, user-defined description of geospatial data can be utilized as a sentence which can be compared to the description of terms from WordNet computational lexicon, e.g. user-defined description can be compared to WordNet synset glosses. For this purpose, separate heuristic techniques will have to be developed and coupled with existing gloss-based methods.

- Enhance the discovery process by taking advantage of toponyms extracted from user’s natural language description. The toponyms will be extracted through usage of Web-accessible gazetteer services and used for improving the process of filtering discovery results.

- Utilize natural language spatial relations for semantic similarity measurement. Spatial relations are very important parts of semantic description of geospatial data. Previously reported research, such as (Schwering, 2008), have proven spatial relations to be very significant for semantic similarity measurement. Therefore, this methodology will be enhanced with mechanisms to detect, extract and utilize spatial relations from the user-defined natural language description.

- Organize discovery results into an OGC-compliant document. The methodology should take advantage of discovered geospatial data source interfaces by organizing them in the form of an OGC-compliant document. For this purpose, we envision the usage of OGC Web Map Context Documents Implementation Specification (Open Geospatial Consortium, 2005b). In this way, each user could be provided with a standardized document describing access points towards geospatial data sources containing the data he/she needs.

- Usability testing with the prototype. Preliminary prototype testing only approves its feasibility. To evaluate performances and effectiveness, appropriate usability study is necessary. For these purposes, current prototype will be replaced with a Web service which will implement the same algorithms.

References

Azzopardi L., 2002., Porter stemmer in CSharp, [http://tartarus.org/martin/PorterStemmer/csharp 2.txt](http://tartarus.org/martin/PorterStemmer/csharp 2.txt) [05 June 2013]


GeoRSS, 2013. [http://georss.org/Main_Page](http://georss.org/Main_Page) [05 June 2013]

Goodwin J. C., Russomanno D. J., Qualls J., 2007. Survey of Semantic Extensions to UDDI:


Meditskos G., Bassiliades N., 2010. Structural and role-oriented web service discovery with taxonomies in OWL-S, Ieee Transactions on Knowledge and Data Engineering, 22, pp. 278–290


Rodríguez A., Egenhofer M., 2004, Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure,
International Journal of Geographic Information Science 18(3): 229-256


Figure 1: The environment used for geospatial data source discovery.

- Natural language terms are input into the Geospatial Data Source Discovery Engine.
- The engine generates a metadata repository and a computational lexicon.
- The metadata repository contains semantic annotations, ontologies, OGC Capabilities documents, ontology-to-information source mapping files, UDDI documents, etc.
- The computational lexicon includes WordNet, BabelNet, FrameNet, etc.
- The metadata generated from the repository and the lexicon are output.
- This output is used to populate the metadata elements and the taxonomy of terms.

Federated geo-information systems are connected to the metadata repository, allowing for integrated access to geospatial data.
"streets and rivers in Serbia"

Input term set = {street, river, serbia}
Figure 3: Expanding the input term set through means of WordNet

expanded term set for term "street"
{entity, physical entity, object, unit, artifact, way, road, thoroughfare,...}

expanded term set for term "river"
{entity, physical entity, thing, water, stream, river, Amazon...}

expanded term set for term "Serbia"
{entity, physical entity, object, location, region, unit, geographic area, Serbia}
Figure 4: Creating and caching a concept term set

concept term set

{thing, feature, areafeature, roadfeature, hydrographicfeature, street, road, localstreet...}
Figure 5: An example of a dictionary created for an ontology concept "road".

<table>
<thead>
<tr>
<th>KEY</th>
<th>thoroughfare</th>
<th>street</th>
<th>opportunity</th>
<th>physical_object</th>
<th>possibility</th>
<th>route</th>
<th>way</th>
<th>two-way_street</th>
<th>main_street</th>
<th>avenue</th>
<th>boulevard</th>
</tr>
</thead>
<tbody>
<tr>
<td>VALUE</td>
<td>0.93</td>
<td>0.88</td>
<td>0.27</td>
<td>0.6</td>
<td>0.29</td>
<td>1</td>
<td>0.93</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Figure 6: The position of Geospatial Data Source Discovery Engine
Click here to download high resolution image
Figure 7: Geometry definition in GeoNis ontology
Click here to download high resolution image
Figure 9: A prototype of Geospatial Data Source Discovery Engine
Click here to download high resolution image
<table>
<thead>
<tr>
<th>Concept name</th>
<th>Street</th>
<th>Road</th>
<th>Main_Street</th>
<th>Local_Street</th>
<th>Service</th>
<th>Shopping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average semantic similarity</td>
<td>0.830151515</td>
<td>0.675909091</td>
<td>0.673484848</td>
<td>0.643030303</td>
<td>0.5719697</td>
<td>0.536364</td>
</tr>
<tr>
<td>Number of matched terms</td>
<td>54</td>
<td>40</td>
<td>39</td>
<td>36</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Cable</td>
<td>0.537727273</td>
<td>0.518484848</td>
<td>0.384848485</td>
<td>0.49969697</td>
<td>0.49151515</td>
<td>0.48697</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Administrative Feature</td>
<td>0.203636364</td>
<td>0.18</td>
<td>0.331363636</td>
<td>0.167424242</td>
<td>0.15924242</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>