Journal of Web Engineering, Vol. 12, No. 5 (2013) 403-421

SEMGSEARCH: AN APPROACH TO SEMANTICALLY RETRIEVE GEOSPATIAL OBJECTS FROM DIFFERENT GEOGRAPHIC SERVERS

JULIO VIZCARRA

Instituto Politécnico Nacional, CIC, Mexico
rvizcarrab08@cic.ipn.mx

MIGUEL TORRES, ROLANDO QUINTERO, MARCO MORENO-IBARRA

Instituto Politécnico Nacional, IPN, Mexico
mtorres@cic.ipn.mx, quintero@cic.ipn.mx, marcomoreno@cic.ipn.mx

Received November 8, 2012
Revised April 29, 2013

In this paper, we propose an approach to semantically retrieve geospatial objects within an Intranet by means of their metadata. It consists of structuring a semantic repository in order to provide the inclusion mechanisms of distributed data to be retrieved, as well as the extraction of those geospatial objects with respect to their conceptual similarity. The similarity measure is based on a conceptual distance (DIS-C algorithm), which consists of determining the levels of similarity among the objects for aiding in the construction of an engine of inclusion and extraction of them. This approach provides a mechanism to handle the knowledge of the geospatial objects distributed on different servers in order to unify the process by means of their semantics. As case study, a web-mapping application called *SemGsearch* has been designed. It provides to the user an engine of semantic retrieval and integration. The result is focused on obtaining a weighting list (ranking) of geospatial objects semantically retrieved by a custom query.

Key words: Semantic retrieval, metadata, application ontology, document ranking

1 Introduction

Nowadays, there are a large volume of geographic data collected, which are obtained by different technologies such as GPS (Global Positioning System), satellite images, geographic databases, maps in analogue format, among other sources [21], and not only by new spatial information systems, but also by data collection technologies that are becoming more sophisticated [14]. In addition, geospatial data are an important issue of any decision support system (DSS), they can be considered as crucial elements for planning and decision-making in a variety of applications. These facts have led to the development of technologies for integration and tools for management and analysis in recent years [11].

On the other hand, there are not agreements in the representation of geospatial semantics in maps and generally in geospatial representations. For example, there are different organizations with an

accuracy to draw certain lines, points or polygons on a plane, in order to represent cities, water supply wells, or altimetry points in the network, transmission lines, road infrastructure, and so on [18]. However there is no consensus or a common approach between organizations or groups of experts about the meaning, semantics or ontology of these representations, which causes that each organization, has a particular representation. In real cases, many organizations handle the concept "artificial lakes" and others the concept "dam". The problem now is not how to accurately represent a geographic feature, the problem is when two spatial representations or geographic databases representing the same, or have a common semantic unit, or the user is familiar with one naming or specific pattern (semantic representation) and he understands other cartographic information made by another organization or another user [13]. This semantic union is the basis for obtaining true interoperability and exchange of geospatial data among different users. Therefore, geographic information systems (GIS) are not exempted from this problem in order to manage and process these data [22, 25].

Most GIS are not originally designed to work with semantic processing, because there are some problems of interoperability, while the integration of heterogeneous geospatial data sources cannot be achieved in these systems [5]. This is mainly because each GIS provides specific requirements for the representation of its data, such as its format and specific query language [3].

The problem in information retrieval is that users can be easily overwhelmed by the amount of information available. The processing time of irrelevant information in documents such as geospatial databases, geographic images, text and tables retrieved by an information retrieval system is very chaotic and this condition is the result of inaccuracies in the representation of the documents in the repository, as well as confusion and imprecision in user queries, since users are frequently unable to express their needs efficiently. These facts contribute to the loss of information and the provision of irrelevant information [6].

Regarding the problems mentioned above, this paper proposes an approach for semantic search geospatial data by means of their metadata. The metadata are structured into a semantic repository, adding *a priori* knowledge of the thematic, spatial and temporal domains, in order to provide mechanisms including the retrieval of distributed data and the extraction of these geographic objects with respect to their conceptual similarity from several distributed servers.

The paper is organized as follows: Section 2 presents the state of the art related to the work in this field. Section 3 describes the proposed methodology to semantically retrieved geospatial information. In Section 4, the experimental results are depicted. The conclusion and future works are outlined in Section 5.

2 Related work

In this section several approaches have been analyzed, used for information retrieval in different data sources to solve the problem of semantic heterogeneity. Some of these works have exploited the metadata, designing descriptors of the information contained in the repositories. Other projects have proposed the use of ontologies, but the common characteristic is the intelligent search, so that models have been developed to retrieve an object or concept from various sources.

In [15] a theory of similarities is presented related to the context on language modelling web services. The implementation of the theory is performed using the framework for WMSL, based on the intersection of logical descriptors. In order to establish the process to give the measure of similarity of a concept, it extracts the attributes or properties of its specification in the WMSL [20]. Later, a matrix is formed in order to compare to other concepts and their respective matrix. For the identical pairs, it is necessary to assign a weight, similar or different to each pair. Additionally, in [19] a framework of semantic annotations is proposed.

Thus, in [16] an approach is described to integrate and retrieve information on distributed heterogeneous resources on the Web. The semantic annotations, defined by the knowledge base within the knowledge discovery processes and control over it, are called entities, which can be generally described interconnected to define a process of discovery of pieces of knowledge in data with little or no human assistant [24].

Other case is presented in [4], who studied the design, syntax and implementation of the semantics and integration within e-business [2]. The integration is performed reconstructing the notion of object-relation with shared components. There are high-level languages to describe knowledge and integrate it into the semantic web by means of techniques of positional and slotted and Artificial Intelligence. A language is created to define the concepts and semantic relations between them, in order to integrate accurate data between different sources.

On the other hand, the work presented in [26] is focused on the discovery and retrieval of spatial data in distributed environments in Spatial Data Infrastructures (SDI). The discovery and recovery tasks are based on three steps: in the first step, the user's search terms are mapped to concepts in domain ontology based on the hybrid ontology approach [17]. In the second step, the concepts are extended on the basis of the hierarchy of concepts in the domain ontology, and the third step consists of the expansion of the query, with which adequate descriptions of geographic information are sought and returned to the users. If the results are adequate, the search is over.

3 The *SemGsearch* Approach

We propose an approach to search geospatial datasets semantically by means of their metadata. This consists of structuring a semantic repository in order to provide the inclusion mechanisms of distributed data to be retrieved, as well as the extraction of those geographic objects with respect to their conceptual similarity.

In order to perform a semantic search, it is necessary to process geospatial datasets at a conceptual level; nevertheless, a geospatial object can be described in many ways, regarding the degree of knowledge, desired abstraction and interpretation. The conceptualization of a domain is used to integrate information. According to the geographic domain, we apply the GEONTO-MET approach [23], in order to build application ontologies for integrating the description of the spatial, temporal and thematic domains proposed as features for the semantic retrieval and integration approach.

Additionally, we defined the DIS-C algorithm to compute the conceptual distance between concepts established in the retrieval process, when the user performs a query. This algorithm avoids obtaining empty answers, whether the exact concept is not found in the sources. Therefore, users will

obtain a concept as a response that can be conceptually closer. The conceptual framework and stages of *SemGsearch* approach is shown in figure 1.

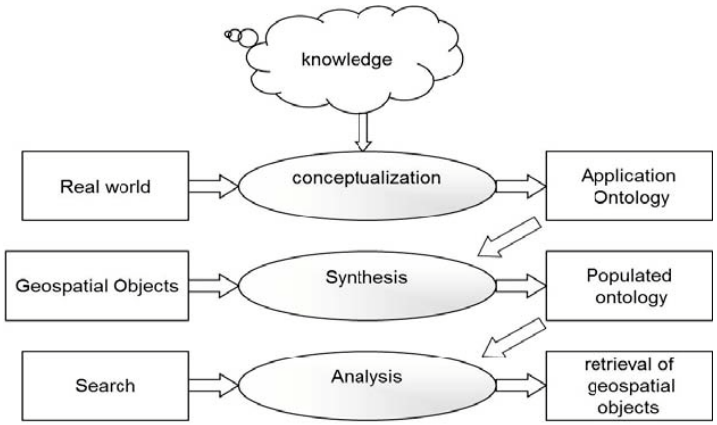

Figure 1 Conceptual framework of the *SemGsearch.*

This approach is composed of three stages: *Conceptualization*, *Synthesis* and *Analysis*. The Conceptualization stage makes the process of building the knowledge base, represented by application ontology. This conceptual structure stores information related to the geographic domain. It also computes the conceptual distance as a basis for the custom query. The Synthesis stage includes the creation of instances that represents geographic concepts of the domain by using the metadata that describe each data source. In other words, the stage has the function of populating the ontology with geospatial objects located in several data sources, composed of an array of repositories. The Analysis stage is used for searching on the spatial, thematic, and temporal domains geospatial objects related to concepts in order to semantically retrieve the most similar concept according to the requested query. The tasks of the three stages of the conceptual framework are depicted in figure 2.

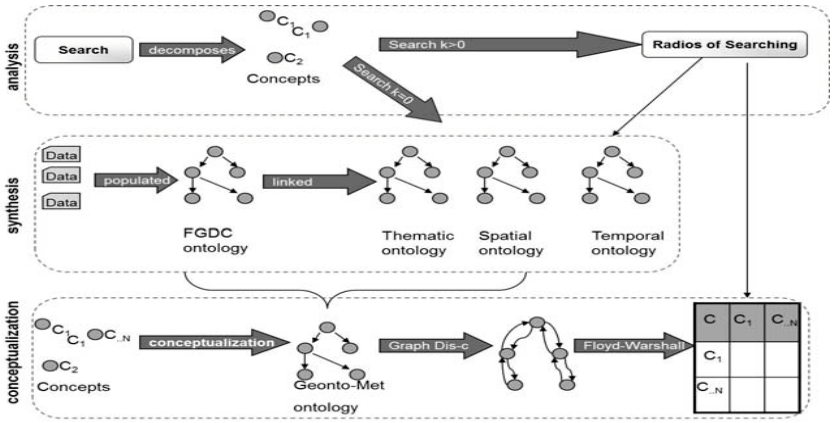

Figure 2 Tasks defined in each stage of the *SemGsearch* approach.

According to figure 2, the Conceptualization stage takes the application ontology that contains the knowledge base of the domains. By applying a semantic similarity algorithm (DIS-C), we build a graph that determines a distance value to each conceptual relationship from the ontology built by GEONTO-MET. Finally, the Floyd-Marshall algorithm, described in [10] is used to determine the smallest value or cost of each concept or node to another, indicating a similarity value. This algorithm reduces the performance of a large number of nodes that works with matrices.

The Synthesis stage picks the geospatial data through their FGDC metadata description [9] and the ontology population is made in the FGDC domain. Later, an instance for each geospatial data is created after that process, with which each object is linked to the other three domains: thematic, spatial and temporal by means of their metadata keywords in order to semantically retrieve geospatial data in the next stage.

In the analysis stage the searching is performed by the decomposition from the query to concepts and after this process, the created vector by the concepts will be used for searching concepts within the ontology and the geospatial object linked in the synthesis stage. Moreover, the retrieval process is based on the expansion of the semantic distance among concepts, increasing the radius. The first approximation is to find equal or exact concepts for the conceptual distance $k = 0$. Later, each domain $k > 0$ from the resulting table of Floyd Warshall in the conceptualization is extended.

3.1. The Conceptualization stage

In this stage, the domains that define geospatial objects are conceptualized. It is possible to use a variable number of domains; however, in this work we only used three domains: thematic, spatial and temporal, which were conceptualized using the GEONTO-MET methodology. It is important to mention while more domains are used, the semantic granularity of the ontology is increased as well as the retrieval approach is more accurate. This fact allows refining a query and making it even more specialized.

3.1.1. DIS-C: Conceptual similarity algorithm

The conceptual distance is defined by the space that separates two concepts within a specific conceptualization [1]. The main idea of the algorithm is to establish the distance value of each relationship within the ontology, and translate it into a weighted directed graph, where each concept becomes a node and each relationship becomes a pair of edges.

The classical step for computing the minimum distances is to find the distance among concepts that are not directly related. Thus, the steps to compute the conceptual distance are the follows:

1. For each type of relationship, it is necessary to assign a conceptual distance for each relationship within the conceptualization. For example, if it has the relation "is" and "road is a communication route", we can set the distance of "road" to "communication route" and “communication route" to "road". As an illustration we establish that $distance(road, \text{communication route}) = 1$ and $distance(\text{communication route}, road) = 0$. Formally, $K(C, \Re, R)$ is a conceptualization where C is the set of concepts, $\Re$ is the set of types of relationships and R is the set of relations in the conceptualization. For each type of relationship $\rho \in \Re$, it has to set the values of δ^{ρ} to the relationship ρ in the

normal sense, and $\overline{\delta}^{\rho}$ to the relationship ρ in reverse sense. In the example above $\delta^{is} = 0$ y $\overline{\delta}^{is} = 1$

2. The directed graph is created $G_K(V, A)$ for the conceptualization *K*. First, it is added to each item $c \in C$ as a vertex in the graph G_K, *i.e.,* $V = C$. Now, for each relationship $a\rho b \in R$, where $a, b \in C$, edges (a, b, δ^{ρ}) and $(b, a, \overline{\delta}^{\rho})$ are added.

3. Once the graph has been built, we apply the Floyd Warshall algorithm to process the minimal distance between two nodes. The table of minimum distances between each pair of vertices, which are directly mapped to the concepts in t he ontology is created. Thus, the result is the conceptual distance disseminated to all concepts in the conceptualization *K*.

3.1.2. Application of DIS-C in GEONTO-MET

There was a very strong feeling (average rating 4.4 on a 0-5 scale) that clients did not well understand the capabilities of the technologies. Similarly it was felt (average rating 4.2) that clients did not understand their own needs as they related to the technology. Perhaps surprisingly, anecdotal evidence indicated that respondents felt that clients had a low understanding of their own organisations and existing processes (most of the time undocumented) that need to be changed to allow for the effective integration of the new system. There was a majority consensus (i.e. 83% responding as *Strongly Agree* or *Agree*) that there needed to be a process at the beginning of the projects focussed on educating their clients.

GEONTO-MET [23] is composed of three axiomatic relations: “is”, “has” and "does". The relation “is”, as we mentioned in the example “road is a communication route”, it establishes the distance of “road” to “communication route” and “communication route” to “road”. It is defined by $distance(\text{communication route}, road) = 1$ and $distance(road, \text{communication route}) = 0$. Thus, we propose that if $a(is)b \in R_R$ then $\delta^{is}(a, b) = 0$ and $\overline{\delta}^{is}(a, b) = 1$.

For the relation “has”, it defines properties, in which the distance is inversely proportional to the number of occurrences of the concept that “has” presents. For example, if an “urban area” “has” a “street of the first order”, then the conceptual distance between the concept “urban area” and the term “street” will be inversely proportional to the number of streets that the urban area presents. That is, if $a(has)b \in R_R$, then $\delta^{has}(a, b) = \dfrac{1}{R(p)}$, where $R(p)$ is the occurrences number of the property $p = a(has)b$ in R_R (this value is normally of 1).

On the other hand, the conceptual distance of “street” to “urban area” will also be inversely proportional to the number of “streets”, Otherwise, “urban area” is directly proportional to the number of properties that “urban area” contains (streets, buildings, parks, etc.) If $a(has)b \in R_R$, then $\overline{\delta}^{has}(a, b) = \dfrac{card(P(a))}{R(p)}$, where $P(a) = \{x \mid a(has)x \in R_R\}$ for any concept $x \in C$ and $R(p)$ is the occurrences number of the property $p = a(has)b$ in R_R.

The relation "does" defines abilities and the conceptual distance is computed in both senses of the relationship, inversely proportional to the number of times that one abilities is referred by one concept, likewise, in the inverse relationship, it is directly proportional to the total number of abilities that a concept contains within ontology.

For instance, we consider the ontology depicted in figure 3, in which the blue arrows indicate the relation "is" and the orange arrows indicate the relationship "has". The conceptual distance is defined between the pairs of concepts by the DIS-C algorithm and the result generated is a weighted directed graph (see figure 4). Once the graph is generated, it is necessary to compute the shortest distance $\Delta_C(a,b)$.

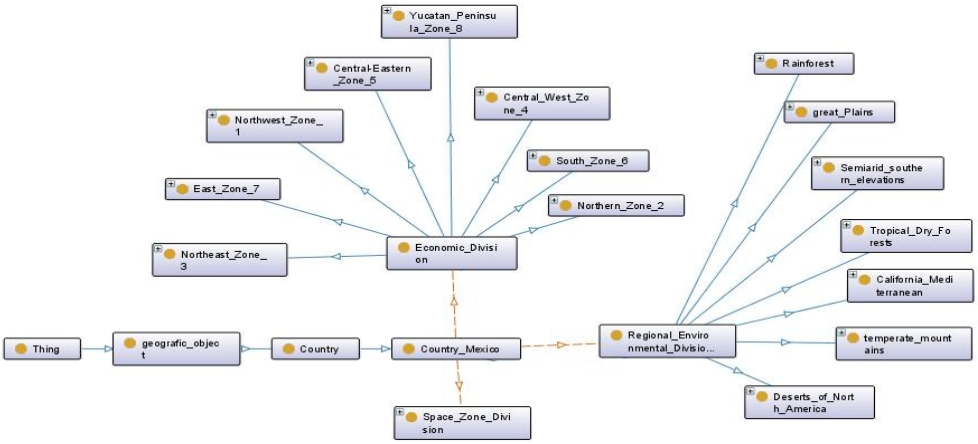

Figure 3 Geospatial ontology for the division of a country (e.g. Mexico).

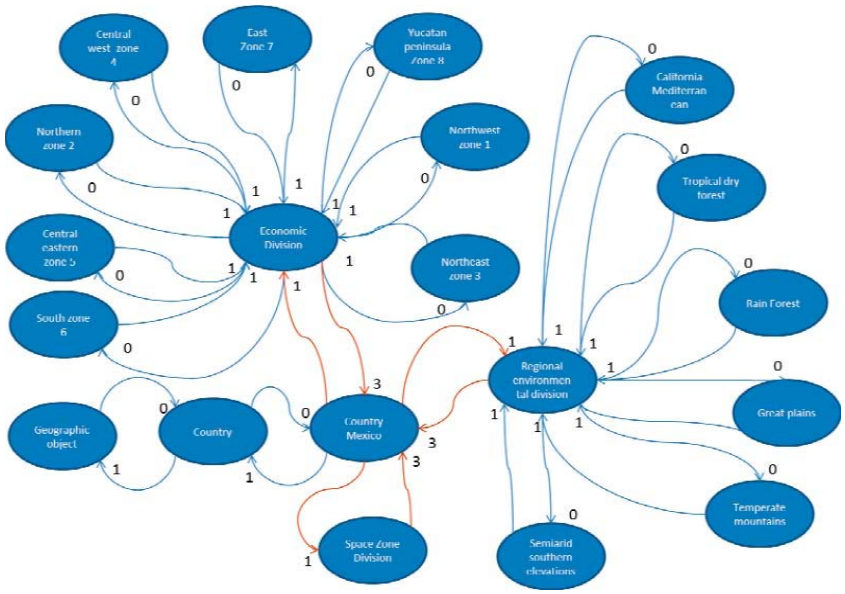

Figure 4 Weighted directed graph obtained from applying the DIS-C algorithm.

To illustrate the DIS-C algorithm, a simple fragment of the ontology with axiomatic relations described in GEONTO-MET, among a set of concepts is shown in figure 5. Likewise, in figure 6 the conceptual distance graph generated from figure 5 is depicted.

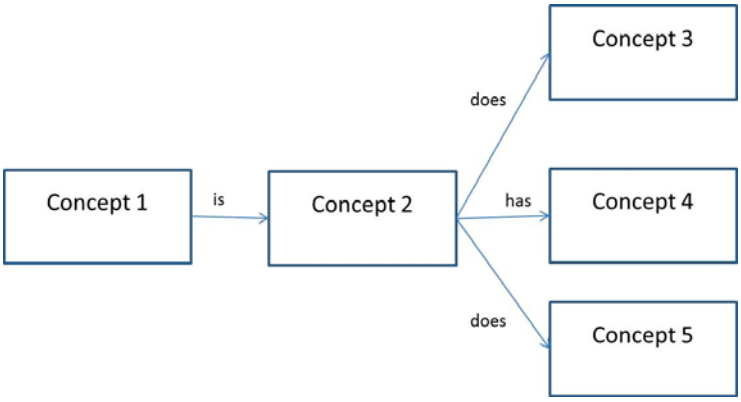

Figure 5 Fragment of ontology with axiomatic relations among concepts.

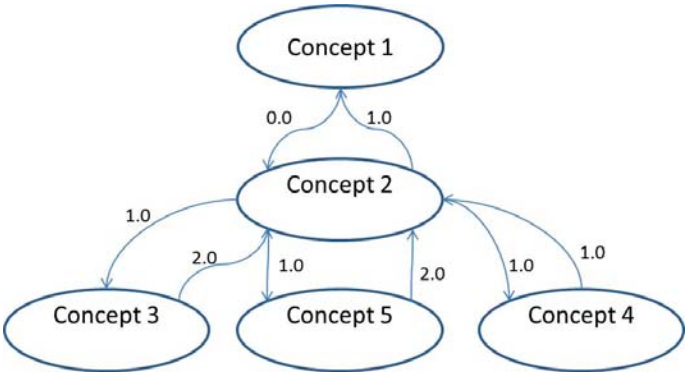

Figure 6 Conceptual graph with semantic distances obtained from the ontology (figure 5).

Once the graph with the weights corresponding to the conceptual distance is obtained, the next step is to transform it in a matrix. Later, it is necessary to measure the similarity of the axiomatic relations.

To illustrate the procedure, we have in the example above the (*concept 1*, *concept 2*) related under the relation "is", according to DIS-C, from the "concept 1" to "concept 2", the conceptual distance or weight $w = 1$ and 0 conversely.

In the case of (*concept 2*, *concept 4*) related under the relation "has" from the "concept 1" to "concept 2", the conceptual distance w is the weight $\frac{1}{R(a\rho b)}$, that is, it has 1/1 and conversely $\frac{card(P(a))}{R(a\rho b)}$, is 1/1.

Finally, the concepts (*concept 2*, *concept 3*) under the relation "does" the weight *w* from the "concept 1" to "concept 2" is $\frac{1}{R(a\rho b)}$, it has 1/1 and conversely $\frac{card(P(a))}{R(a\rho b)}$ is 2/1.

The next procedure is to determine the lowest weight *w* between two concepts *a* and *b* belonging to a given domain ontology. The process is carried out using the Floyd Warshall algorithm, which considers as input the resulting graph as a DIS-C matrix between the weights of the concepts *w*, *Cx*, *Cy*, and another one that belongs to the same domain, which indicates the shortest path or minimum weight between any pair of nodes *Cx*, *Cy* that belong to the graph DIS-C. The matrix that describes the minimum weight of a concept *Cx* to a concept *Cy* is presented in Table 1. In addition, the DIS-C algorithm is described in Table 2.

Table 1 Matrix generated by the Floyd Warshall algorithm using the DIS-C graph, indicating the shortest distance.

Concepts	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{C_4}$	$\mathbf{C_5}$
$\mathbf{C_1}$	0	0	1	1	1
$\mathbf{C_2}$	1	0	1	1	1
$\mathbf{C_3}$	3	2	0	3	3
$\mathbf{C_4}$	3	2	3	0	3
$\mathbf{C_5}$	2	1	2	2	0

Table 2 The DIS-C algorithm.

1. $G_O \leftarrow$ weighted directed graph
2. $R_R \leftarrow$ relations in O
3. For all $a\rho b \pi c \in R_R$ do
4. If $\rho = is$ then
5. Add $(a,b,0)$ to G_O
6. Add $(b,a,1)$ to G_O
7. Else if $\rho = has$ then
8. Add $(a,b,\frac{1}{R(a\rho b)})$ to G_O
9. Add $(b,a,\frac{card(P(a))}{R(a\rho b)})$ to G_O
10. Else if $\rho = does$ then
11. Add $(a,b,\frac{1}{R(a\rho b)})$ to G_O
12. Add $(b,a,\frac{card(H(a))}{R(a\rho b)})$ to G_O
13. End if
14. End for
15. Return Shortest path (Go)

3.2. The Synthesis stage

This stage generates instances from the concepts described in the ontology defined in the previous stage. The task is performed for each site or server where the data sources are located. It is described as follows:

1. For each data source located in the repository, the description of geographic objects is obtained using the metadata file that accompanies each geographic object and it is mapped on FGDC domain (see figure 7).
2. For each domain of metadata, the set of keywords is identified. These keywords are mapped with concepts in the others domains. Thus, it searches in the thematic, spatial and temporal domains (see figure 8).

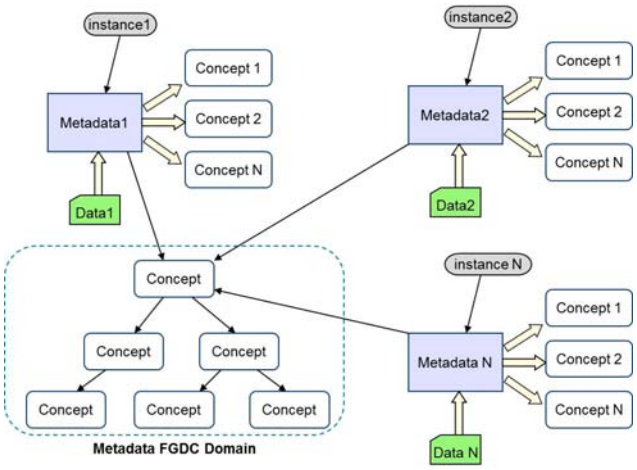

Figure 7 Schema to generate instances of geospatial data.

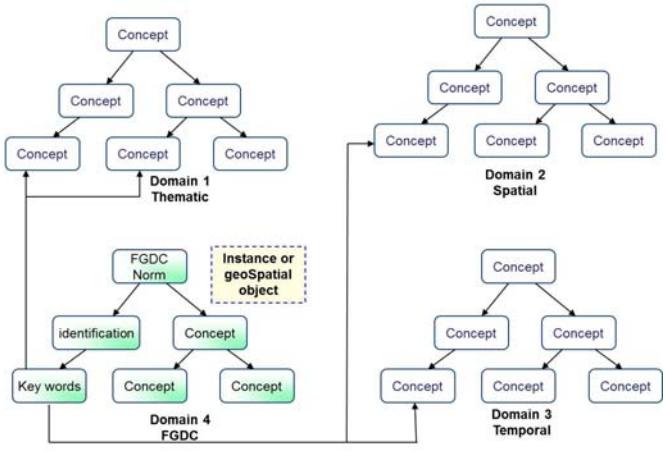

Figure 8 The Synthesis stage applied to the application ontology.

3.3. The Analysis stage

In the analysis stage, the tasks of semantic retrieval and the displaying of the results are carried out. The stage is composed of five basic tasks:

1. From the searching terms, the key concepts to be retrieved in the domains defined in the ontology are obtained. This process identifies the concepts that compose the query. After that, a vector that defines the geographic objects is created.
2. With this vector, we proceed to find each set of concepts with their respective domain.
3. Whether the instance concept in the query has an exact match with an instance concept in the domain, it has the closest conceptual distance ($k = 0$). The process continues by extending for $k >= 0$ (see figure 9).
4. When the radius are increased to values $k >= 0$, more concepts are obtained, thus several instances of geospatial objects in the data sources are semantically retrieved. This procedure is applied to each domain, thus a list of concepts that are semantically related is created by means of the intersection among domains (see figure 10).
5. The last task of the searching is to return the entire set of instances of concepts from the previous procedures for $k = 0$ that represents an exact match and for values $k > 0$ that are semantically close.

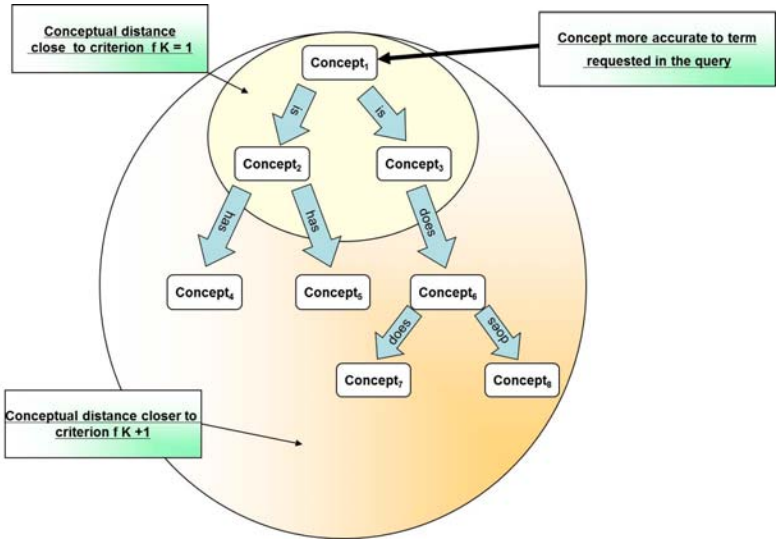

Figure 9 Conceptual distance extending the searching range.

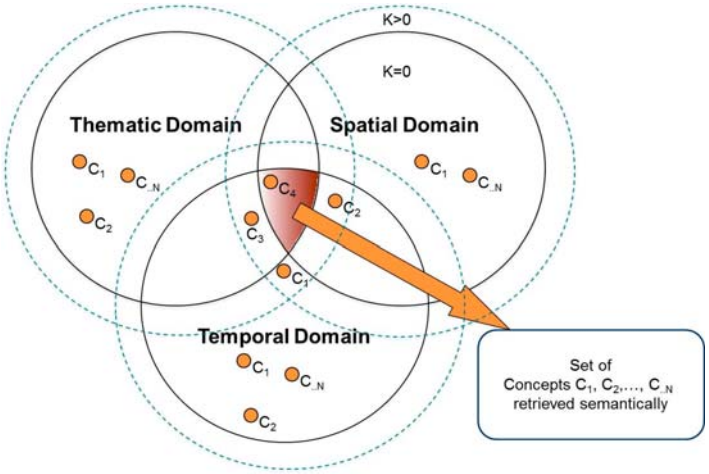

Figure 10 Semantic intersection among the sets of geographic concepts.

We propose to expand the radius in the search as follows: First, we define $A = \{x \mid x$ is a concept $\in$ Thematic Domain$\}$, $B = \{y \mid y$ is a concept $\in$ Spatial Domain$\}$, $C = \{z \mid z$ is a concept $\in$ Temporal Domain$\}$, and $K \in \mathbb{N}$. For instance, to increase the radius for a value $i > 0$ and $i \to \infty$, for each i+1, it is calculated $A' = \{x_i \mid \Delta_C(x,y) < K_i \; \forall(x,y) \in$ Thematic Domain$\}$ and $A' \subseteq A$, $B' = \{x_i \mid \Delta_C(x,y) < K_i \; \forall(x,y) \in$ Spatial Domain$\}$ and $B' \subseteq B$, $C' = \{x_i \mid \Delta_C(x,y) < K_i \; \forall(x,y) \in$ Temporal Domain$\}$ and $C' \subseteq C$. The result set of concepts, according to the expansion of the radius S for a K_i is defined by $S = \{x \mid x \in A' \cap B' \cap C'\}$.

4 Experimental results

As case study, the *thematic domain* is composed of urban areas, communication infrastructure, and forest resources using the specification from INEGI (Mexican National Institute of Statistics, Geography and Informatics) [12]. The *spatial domain* is referred to Mexico with different classifications: regional division by CONABIO (Mexican National Commission for Knowledge and Use of Biodiversity) [8], spatial distribution area by CFE (Mexican Federal Electricity Commission) [7] and economical area as in the classification from INEGI. Finally, *the temporal domain* is conceptualized by the rules of the calendar.

In the *Conceptualization stage*, we built the application ontology that contains 450 concepts (see figure 11), distributed in the domains: thematic (purple line), spatial (blue line), temporal (red line) and metadata (green line). The concepts are related with 150 relationships "*is*", 50 with the relationship "*has*" and 40 with "*does*". We appreciate that whether the number of concepts and relationships are increased, the granularity and precision will be raised in the search engine as well.

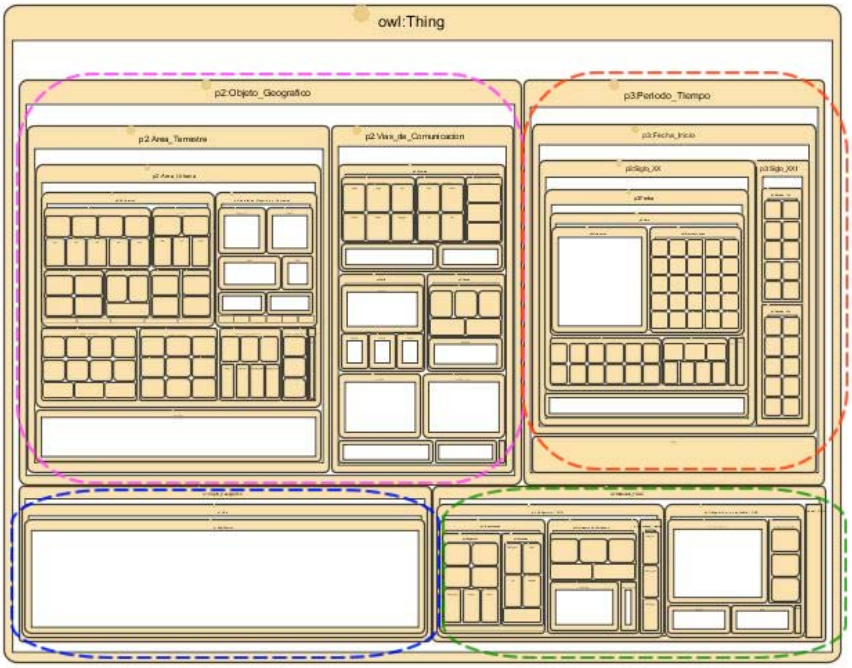

Figure 11 The application ontology designed for the *SemGsearch.*

As an example of the conceptualization on the thematic domain, we present a partition that shows the conceptualization for the "communication routes" and how is divided into a street, according to INEGI (see figure 12).

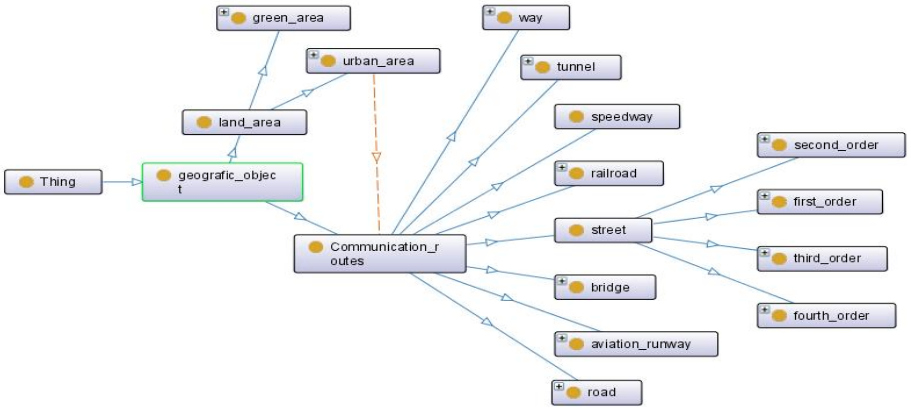

Figure 12 Partition of the "Communication routes" on the thematic domain.

Later, the DIS-C algorithm is applied in order to generate a weighted graph strongly connected, which is directly based on the conceptual distance obtained among the concepts. We only present the

DIS-C graph for the spatial domain in figure 13. It contains 62 concepts; the reduction with 30 concepts depicted in figure 14 is performed in order to appreciate the distance values.

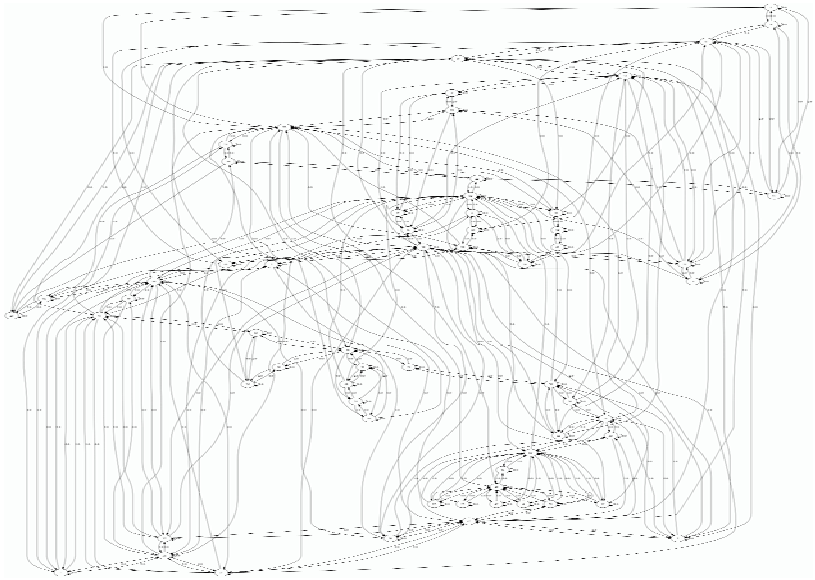

Figure 13 The DIS-C graph for the spatial domain.

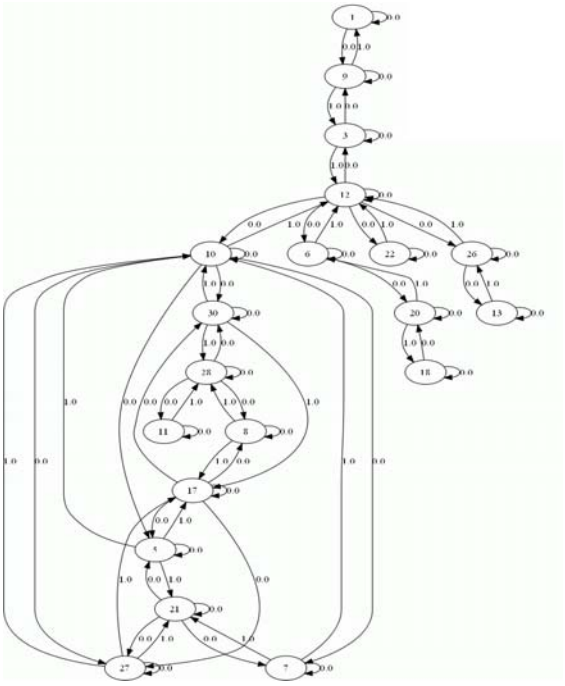

Figure 14 The DIS-C graph with 30 concepts of the spatial domain.

This graph is stored in a database as a table, each object *Cx* is related to another *Cy*, where *Cx* and *Cy* are concepts of a domain in the ontology. The table shows the numerical values that correspond to the weights *w*, according to the conceptual distance. The next process is to apply the Floyd-Warshall algorithm in order to obtain the shortest distances on the concepts that describe the spatial domain (see Table 3).

Table 3 The shortest conceptual distances among the concepts (Floyd Warshall algorithm).

Conceptual Distance	**Concept**	**C_1**	**C 2**	**C 3**	**C 4**	**C 5**	**C 6**	**C 7**	**C 8**	**C_9**	**C 10**	**C 11**	**C_{12}**	**C 13**	**C 14**	**C 15**	**C 16**	**C 17**	**C_{18}**	**C 19**	**C n**
Δ_{C1}	C_1	0	1	1	1	1	2	1	1	0	2	1	2	1	5	1	2	2	1	1	...
Δ_{C2}	C_2	2	0	2	2	2	2	2	1	2	2	1	2	2	6	2	2	2	2	1	...
Δ_{C3}	C_3	1	1	0	1	1	1	1	1	0	1	1	1	1	5	1	2	2	2	1	...
Δ_{C4}	C_4	2	2	2	0	2	2	2	2	2	2	2	2	2	6	1	3	2	2	2	...
Δ_{C5}	C_5	2	2	2	2	0	2	1	1	2	1	1	2	1	6	2	2	1	2	2	...
Δ_{Cn}	C_n	…	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

In the *Analysis stage* the tasks of semantic retrieval and visualization of geospatial data are performed in the *SemGsearch* application. The request is carried out according to the characteristics of geospatial objects that will be searched. The correspondence of concepts to a question or domain, is made by separators ";" and it generates three segments: the first of them is the domain "what" or thematic, the following "where" or spatial, and the last is temporary or "when". The query to the engine is submitted as follows: < thematic domain (geographic objects) ';' spatial domain (location) ';' temporal domain (time period) >.

As an example the next query retrieves the geospatial object "boulevard" in the location of "Veracruz state, Mexico" (see figure 15). Due to "boulevard" is located at "Veracruz", it is not stored in any server; this query will not retrieve any instance with an exact matching ($k = 0$). In this case, to avoid empty answers, the conceptual distance is applied to retrieve similar concepts according to the conceptualization.

Figure 15 Semantic search: thematic "boulevard", spatial "Veracruz, Mexico", year "2006"; *k*=0.

Now, it is necessary to expand the radius for a value $k = 1$, the semantic search is increased in the partition of the ontology, according to the conceptualization, within spatial domain "Eco-regional Division". There is a similarity under the classification of "Warm Humid Jungles", the concepts "Chiapas state" and "Veracruz state" are the closest states semantically with respect to the others (see figure 16).

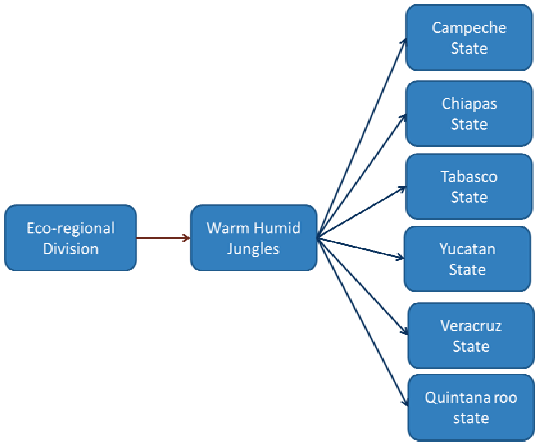

Figure 16 Ontology partition of the class "Eco-regional Division" and "Warm Humid Jungles".

According to the increment of the conceptual distance for $k > 0$, the next search is made as "Boulevard " that is located in "Veracruz", this geospatial object is already located in one server in the repository. The results that accomplish the request query by means of semantic similarity are depicted in figure 17 and figure 18. In this case, for "boulevard" in "Veracruz", with $k = 1$ and $k = 2$ are shown respectively.

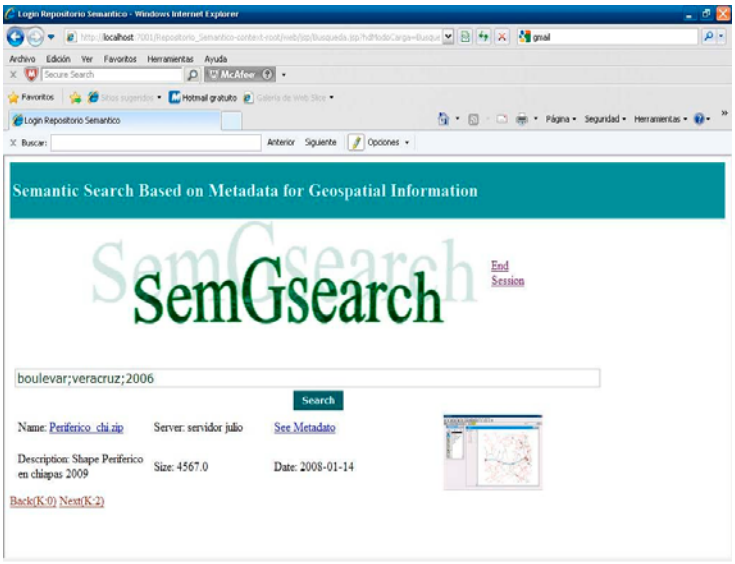

Figure 17 Semantic search: "boulevard", "Veracruz", "2006" with $k = 1$.

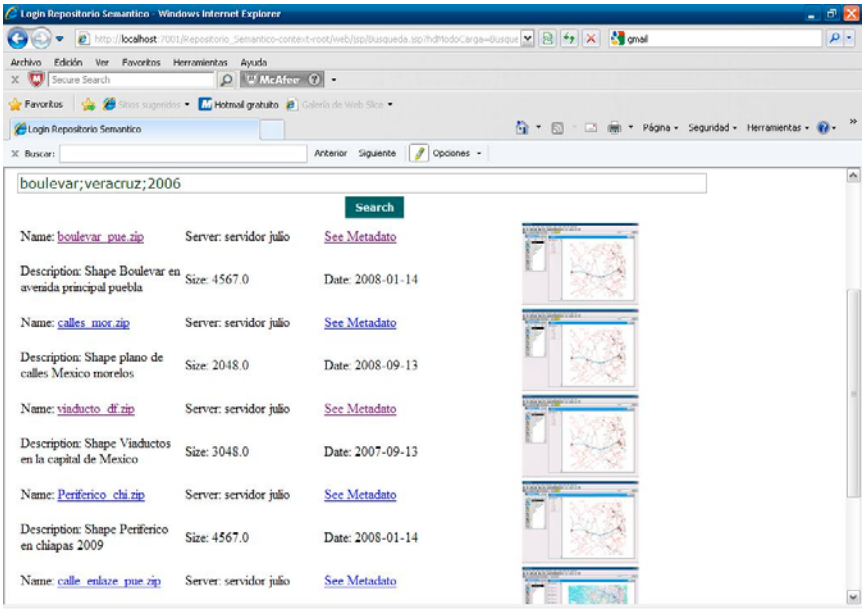

Figure 18 Semantic search: “boulevard”, “Veracruz”, “2006” with $k = 2$.

Finally, in the visualization task the geospatial objects can be downloaded, displayed or reviewed by their metadata file (see figure 19).

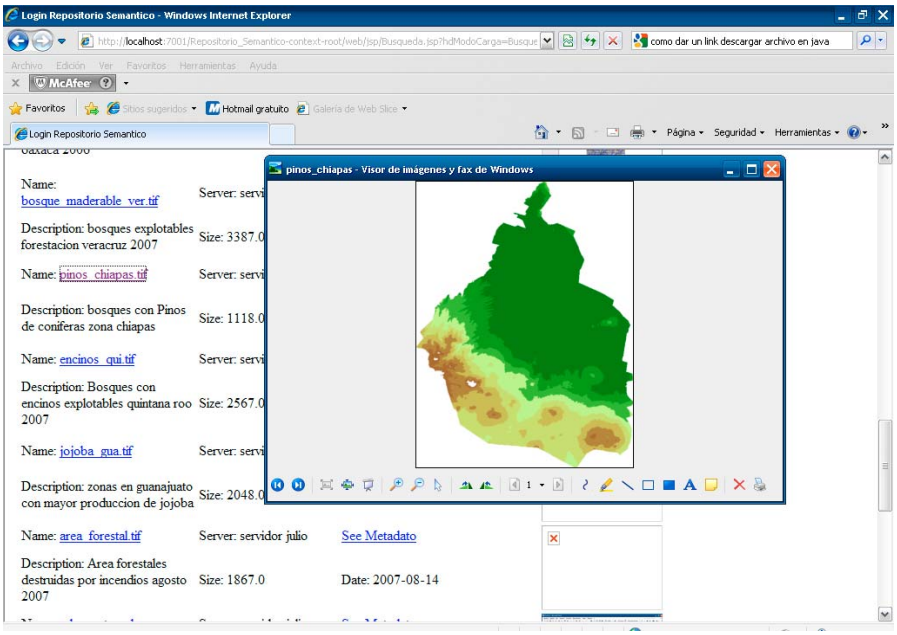

Figure 19 Visualization task for geospatial data.

5 Conclusions and future work

In this paper, a methodology is described focused on the tasks of semantic integration and retrieval of diverse heterogeneous data sources, which can be located on different servers.

Similarly, the main goal of this paper is to be able to share geospatial information by means of semantic retrieval mechanism, using a conceptual representation. Under this approach, a mechanism is proposed to measure the conceptual distance called DIS-C algorithm. This algorithm has the function of measuring how close two concepts conceptually are. It allows the system to resolve queries according to *k* values that represent a conceptual distance. This method avoids returning empty requests to the users.

We implemented the *SemGsearch* system as case study, to test the proposed approach. This application provides ranking list of the geographic objects semantically retrieved.

Future work is oriented for applying purposes of this methodology in particular cases of the semantic web. The conceptual distance is an important issue that covers semantic information integration and semantic interoperability among applications and information in the web.

Acknowledgements

This work was partially sponsored by the Instituto Politécnico Nacional (IPN), the Consejo Nacional de Ciencia y Tecnología (CONACyT) under grant 106692, and Secretaría de Investigación y Posgrado (SIP) under grants 20120563 and 20121661. Additionally, we are thankful to the reviewers for their invaluable and constructive feedback that helped improve the quality of the paper.

References

1. Abu-Hanna, A. and Jansweijer, W.N.H., Modeling domain knowledge using explicit conceptualization. IEEE Expert, 9 (2). 1994, 53-64.
2. Amor, D., The e-business (r) evolution: living and working in an interconnected world. Prentice Hall PTR Upper Saddle River, NJ, 2000.
3. Bishr, Y., Overcoming the semantic and other barriers to GIS interoperability. International Journal of Geographical Information Science, 12 (4). 1998, 299-314.
4. Boley, H., Integrating positional and slotted knowledge on the semantic web. Journal of Emerging Technologies in Web Intelligence, 2 (4). 2010, 343-353.
5. Buccella, A., Cechich A., Gendarmi, D., Lanubile, F., Semeraro, G. and Colagrossi, A., Building a global normalized ontology for integrating geographic data sources, Computers & Geosciences, 37(7). 2011, 893-916.
6. Deborah, M., OWL Web Ontology Language Overview. Knowledge Systems Laboratory, Stanford University. http://www.w3.org/TR/owl-features/. Accessed 20 July 2012.
7. Mexican Federal Electricity Commission. http://www.cfe.gob.mx/. Accessed 20 June 2012.
8. Mexican National Commission for Knowledge and Use of Biodiversity, CONABIO-SEMARNAT. http://www.conabio.gob.mx/. Accessed 20 June 2012.
9. Federal Geographic Data Committee. http://www.fgdc.gov/. Accessed 20 June 2012.
10. Floyd, R.W., Algorithm 97: The Shortest path. Communications of the ACM, 5(6). 1962, 345-367.
11. Heywood, I., Cornelius, S. and Carver, S., An Introduction to Geographical Information Systems. Pearson Education Limited, 2006.

12. Mexican National Institute of Statistics, Geography and Informatics. http://www.inegi.org.mx/. Accessed 20 July 2012.
13. Isla, J.L., Gutiérrez, F.L., Gea, M. and Garrido, J.L., Descripción de Patrones de Organización y su Modelado con AMENITIES. In Proceedings of the IV Jornadas Iberoamericanas de Ingeniería del Software e Ingeniería del Conocimiento (JIISIC04), Madrid, 2004, 3-14.
14. Janissek-Muniz, R. and Moscarola, J., Dinámica del proceso de recolección y analisis de datos vía web. in Proceedings of the Consejo Latinoamericano de Escuelas de Administración. Santiago de Chile. 2005, 1-17.
15. Janowicz, K., Similarity-based retrieval for geospatial semantic web services specified using the web service modeling language (wsml-core). The Geospatial Web-How Geo-Browsers, Social Software and the Web, No. 2. 2007, 235-246.
16. Kiryakov, A., Popov, B., Terziev, I., Manov, D. and Ognyanoff, D., Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web, 2(1). 2004, 49-79.
17. Liang, Y., Bao, H. and Liu, H., Hybrid Ontology Integration for Distributed System. In Proceedings of the Eighth ACIS International on Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, IEEE Press, Vol. 1, 2007, 309-314.
18. MacEachren, A.M., How maps work: representation, visualization, and design. The Guilford Press, 2004.
19. Malik, S.K., Prakash, N. and Rizvi, S., Semantic Annotation Framework for Intelligent Information Retrieval using KIM Architecture. International Journal of Web & Semantic Technology, 1(4). 2010, 2-26.
20. Sabbouh, M., Higginson, J., Semy, S. and Gagne, D., Web mashup scripting language. In Proceedings of the 16th International Conference on the World Wide Web, ACM Press, 2007, 1305-1306.
21. Santos, J.M., The urban transformation of Fuenlabrada: A consequense of the metropolitanization of Madrid. Iberian Studie, University of Keele, United Kingdom. 1986, 39-49.
22. Smith, B. and Bittner, T., Granular partitions and vagueness. in Proceedings of the International Conference on Formal Ontology, IOS Press, 2001, 309-320.
23. Torres, M., Quintero, R., Moreno-Ibarra, M., Menchaca-Mendez, R. and Guzman, G., GEONTO-MET: An approach to conceptualizing the geographic domain. International Journal of Geographical Information Science, 25(10). 2011, 1633-1657.
24. Trochim, W.M.K. and Donnelly, J.P., Research methods knowledge base, Vol. 32, Atomic Dog Pub, 2001.
25. Vilches, L.M., Ramos, J.A., López, F.J., Corcho, O. and Nogueras, J., An Approach to Comparing Different Ontologies in the Context of Hydrographical Information. in Proceedings of the Fourth International Workshop on Information Fusion and Geographic Information Systems, Lecture Notes in Geoinformation and Cartography, Springer-Verlag, 2009, 193-207.
26. Zhan, Q., Zhang, X. and Lic, D., Ontology-Based Semantic Description Model for Discovery and Retrieval of Geospatial Information. The International. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2008.