Automatic Representation of Semantic Abstraction of Geographical Data by Means of Classification

Rainer Larin Fonseca and Eduardo Garea Llano

Advanced Technologies Application Centre, 7ma # 21812, Siboney,
Playa, Havana - 12200, Cuba
{rlarin,egarea}cenatav.co.cu

Abstract. Providing Geographical Information Systems (GIS) with the mechanisms for processing geographical data based on their semantic abstraction is a task that at present is carried out in a number of research given their scope of applications. Tackling this issue may help to solve many problems of geographical data like its heterogeneity, since the SIG could process geographical data focusing on their meaning and not on their syntax and/or structure, thus reducing the Man-Machine semantic gap. An important aspect for achieving these objectives is the establishment of an automatic way of correspondence between geographical data and their conceptualization in a Domain Ontology. In this work, we propose a new type of Ontology, a Data-Representation Ontology. We also propose a new method for the automatic generation of the Data-Representation Ontology from geographical data and his interrelationships with the Domain Ontology. For this we use pattern classification techniques and a dissimilarity measure. The experiments showed that once the Data-Representation Ontology was generated, the classifier using dissimilarities could correctly classify all the data.

Keywords: Ontology, Classification, Semantic, Geographical data.

1 Introduction

For some years, scientists have been working with the aim of having a uniform access to geographical data. One of the principal problems of geographical data [1] is the heterogeneity in them. This implies that it becomes very difficult to work with this data in a uniform way, mainly by compatibility problems between them.

The problem of getting a uniform access to heterogeneous data is known as integration of geographical data. Efforts in this direction are focused mainly on these two types of integration:

- *Syntactic-Structural Integration:* It proposes the existence of a technical interconnection between data that may be in different reference systems or in different formats.
- *Semantic Integration:* It proposes the integration of heterogeneous data based on their meaning and not based on what they are, ensuring a mutual understanding over a context defined between different systems including human beings who may interact with them.

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 557–568, 2010.

In the literature there are recent papers [2-7] that deal with the issue of semantic integration of geographical data. In them, the use of Ontologies as the knowledge representation mechanism for the integration process is proposed, precisely because Ontologies are based on both Object-Oriented (OO) and Relationship-Entity (RE) paradigms. These paradigms are essential to phenomena modeling in geographical scope.

One of the major problems for geodata processing from the semantic point of view is precisely the way in which they will be conceptualized. Firstly, one must have knowledge about the nature of data based on conceptual domain that it is wished to model. On the other hand we have the geographical data complexity and finally the conceptualization way of these data; this means the way to represent the structure semantic abstraction of data and their interrelationships with the Domain Ontology. In this paper we focus on the task of representing geographical data based on their semantic abstraction supported by a Domain Ontology that models their nature with a higher level of abstraction.

This paper continues with a brief section where some key concepts for the understanding are defined. After that, the types of Ontologies existing in the literature are presented followed by the proposal of a new type of Ontology and its structural description. Then, the principal steps for establishing the correspondence between the semantic abstraction of geographical data and the Domain Ontology is described using a classification technique based on a distance. Paper continues with experiments and its results and fallow by the conclusions.

2 Definitions

- *Geographical Datum:* The geographical dictionary ESDIG[8] defines Geographical Datum as "*Object or Entity resulting from an abstraction of the real geographical space. (...), its definitive characteristic is a spatial reference in two or three dimensions*". It also states that in some cases the following terms are considered synonymous with geographical datum: geospatial datum, geographical object and others that correspond with these definitions. These synonyms will be used throughout this paper indistinctly.
- *Class:* Class could be defined as *"A set of similar objects"* [9] e.g.: Objects that share common features, taking this into our context we could say that a Class is defined as a concept that could contain other sub-concepts and represents the semantic nature of geographical objects that have common features on a set.
- *Semantic Abstraction:* We understand by semantic abstraction the process that implies a reduction of the main components of the information from a phenomenon so as to preserve its most important features aiming at extrapolating this phenomenon to a semantic space in which it is defined according to their meaning.

3 Ontologies

There are forms of semantic representation and in general they are limited or focused to a specific semantic domain such as shown in [10-11]. Ontologies are the most widely used since they provide formal specifications of the logical models in which the data is based. Ontologies have appeared to provide a common vocabulary in a knowledge domain and to specify, at different formalism levels, the meaning of terms and their relationships. Therefore Ontologies provide a shared and accepted understanding of the knowledge of a domain, which can be communicated among human beings, between heterogeneous systems and between human beings and systems. One of the most popular and quoted definitions of ontology is the one proposed by Gruber and later extended by Studer "*An ontology is an explicit specification of a shared conceptualization*"[12-13], which shows that they have been developed for interchange and use of knowledge efficiently.

Guarino in [14] defines several levels of generality that give rise to different types of ontologies, see Fig. 1:

- *Top-Level Ontologies:* Contains reusable generic terms in different domains.
- *Domain Ontologies and Task Ontologies:* Contain terms that are specific in a particular domain (e.g.: Soils or Geology) or specific task (e.g.: Selling). These terms are usually defined as specializations of existing concepts in Top-Levels Ontologies.
- *Application Ontologies:* Contain all necessary terms to model a particular application. They are often specializations of Domain Ontologies or Task Ontologies.

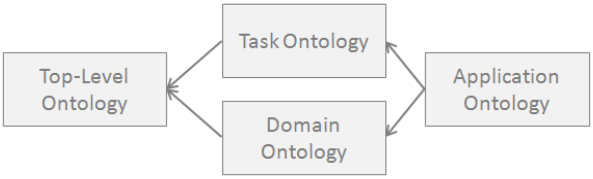

Fig. 1. Graphic representation of kinds of Ontologies proposed by Guarino

4 Data-Representation Ontology (DRO)

To face the semantic integration problems of heterogeneous geographical data it is then necessary to extrapolate this data into a common space independent of type and/or format in which they have been stored based on their semantic abstraction. The Ontologies mentioned above are designed to to capture the semantics from the different geographical domains but in a broad manner. These Ontologies only express the different concepts and their relationships up to a specific abstraction level, since it does not take into account the semantic embed in the geographical data integrated in GIS. As result of this, the characteristic of these data and the relationships between them (see Fig.3-A) are not used, therefore valuable information may be lost. Furthermore, these Ontologies are neither capable of discovering new and more specialized concepts that could be embedded in the data, see Fig.3 –C. These new concepts may be obtained from the data processing, e.g. a process of data clustering. It contributes to a major granularity in the Ontology and therefore the new abstraction levels. Both the information embedded in the data and the new abstraction levels contributes to better accuracies in the results after its use in many tasks like information retrieval and/or data analysis for the decision making.

To tackle these issues we consider it necessary to define a new type of Ontology that covers this semantic emptiness above the geographical data integrated in GIS. To this end we propose the definition of a new type of Ontology, "*Data-Representation Ontology (DRO)*" it will represent the features that describe the nature of data and the existing relationships between them.

It would be formally defined as:

- *Data Representation Ontology:* Contains the necessary definitions for the representation of features and relationships that model and give meaning to objects belonging to a domain from a semantic point of view.

Based on the scheme proposed by Guarino[14] , see Fig.1, we have included the Data Representation Ontology (DRO) in the lowest level of generality since the DRO represents the major degree of specialization with respect to the other Ontologies, see Fig. 2:

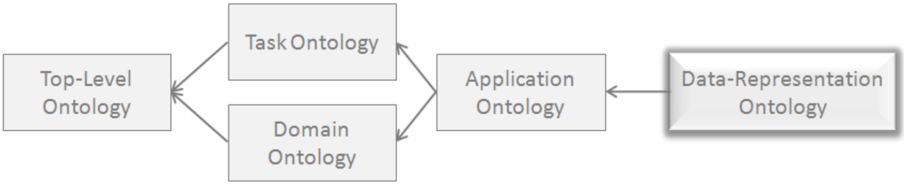

Fig. 2. Graphic representation of kinds of Ontologies, including DRO

The DRO is essentially a dynamic ontology since its structure; terms and relationships are always going to depend on the data the user is working with. The use of DRO allows the integration of heterogeneous data, see Fig.3 and it also provides a greater semantic enrichment from the complement generated between DRO and the Ontology employed by the user (e.g.: Domain Ontology (DO)). This complement is addressed in both directions, from DO to DRO and from DRO to DO, see Fig. 3. This is explained by the fact that on the one hand the DO will have exact information of the data being worked with allowing a major level of specialization and therefore a major granularity; and on the other hand the DRO will contain more levels of abstraction from the semantic point of view, which is provided by DO. Here a process of synergy is shown in which the results obtained through the use of the two ontologies (DRO and DO) are better than the sum of the results obtained by each of them separately.

The user could change (add, delete or modify) his working data, this implies the restructuring of Ontology (generation of new terms and relationships), here we would like to distinguish that the Domain Ontology does not change, retaining its original structure; only the DRO undergoes the changes. These changes occur below the interconnection layer, see Fig. 3.

4.1 Structure of the DRO

As mentioned above, the DRO is the projection in the semantic space of geographical data, see Fig. 3, in which all the represented geographical data is described in the same structure and consequently a uniform access to it is possible, taking a step forward in the process of integration of heterogeneous data.

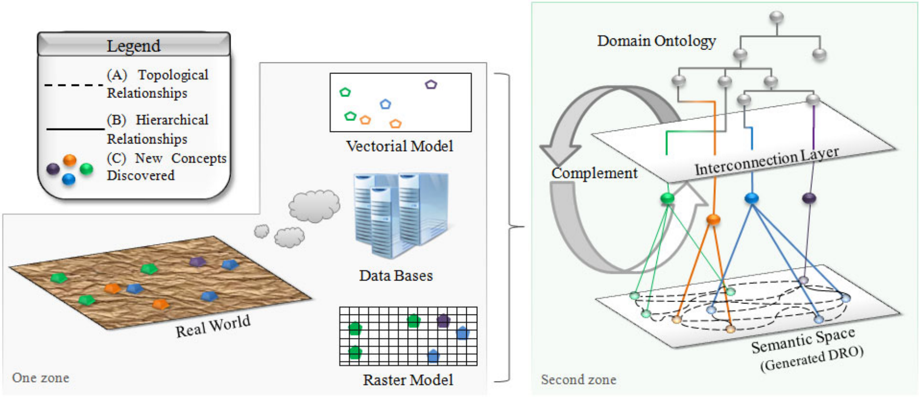

Fig. 3. Graphic representation of the different models in which the geographical data can be stored (one zone) and the projection in the semantic space of geographical data and the complement between the DRO and the DO (second zone). The Legend shows in (A) the representation of the relationships between geographic objects (e.g. topological relationships), in (B) the hierarchical relations are shown (e.g. sub_class or super_class) and (C) shows the new concepts that may be obtained from the data processing e.g. a process of data clustering.

In this paper, an architecture for the DRO construction is proposed. The basic unit of this architecture is constituted by Data-Representation Nodes (DRN) and edges that interconnect these DRN. Each DRN represents the semantic abstraction of a single datum and the edges representing the existing relationships between data, e.g.: topological relationships see Fig. 4. The structure of these DRN is based on the representation of thematic features, spatial features and temporal features of geographical data, by means of three substructures (*Thematic Component, Spatial Component and Temporal Component*) in semantic space.

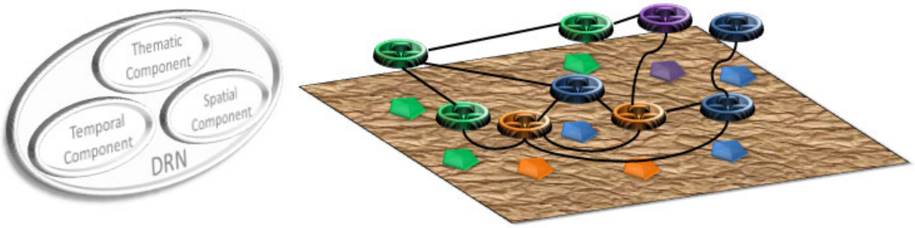

Fig. 4. Graphic representation of Data-Representation Node

4.2 Principals Steps for Automatic Generation of DRO

The DRO is generated from the data information the user is working with. In this section the main steps of the algorithm for automatic generation of the DRO are shown, these are:

1. Definition of the structure in which the data is stored:
   1.1. To specify which the structures of a geographical datum that refer to its three main components (*Thematic Component, Spatial Component, and Temporal Component*) are and how to access their values.
2. Extraction of values:
   2.1. *Thematic Component:* All thematic attributes which characterize the geographical datum are extracted and normalized. These are attributes that answer the question: What is it?
   2.2. *Spatial Component:* All spatial attributes that define location in space and cartographic projection are extracted and normalized. These are attributes that respond to the question: Where is it?
   2.3. *Temporal Component:* All temporal attributes that define the moment in which the datum is manifested are extracted and normalized. These are attributes that respond to the question: When? Also from these components the changes occurred in time with respect to both the thematic component and spatial component are extracted, e.g.: the changing values of its properties or its position in space.
3. Extraction of Relations:
   3.1. To extract all existing relations in the geographical datum with respect to other data, e.g.: Topological relationships.

The way to carry out these steps depends mainly on the formats and standards of the data. Each kind of formats and standards has its own characteristics that can modify the way in which the steps for automatic generation of DRO are carried out.

5 Semantic Abstraction of Geographical Data

As mentioned above the semantic abstraction of data refers to the operation by which certain properties of a geographical phenomenon are isolated for their processing from a semantic point of view. In this sense, geographical data can be represented in different semantic spaces, since it is possible to do several observations of the same phenomenon from different viewpoints, e.g.: a biologist can see a lake as a fish habitat while a hydrologist can see it as a body of water. Fonseca in [7] defines this phenomenon as roles. Therefore, in our context, the role played by the data will be determined by the ontology that defines its application domain, e.g. the lake object seen previously could play the role of fish habitat or body of water depending on the domain ontology used.

To increase levels of semantics abstraction of data it is necessary to link it with the domain ontology, i.e. to establish a correspondence with the concept it belongs to in the Domain Ontology. In such a way the data acquires more expressiveness and this turns out to be vital for the processing of data from an abstract point of view. To achieve this we propose the following steps:

1. For the Domain Ontology (DO):
   1.1. To extract the features present in each concept of type leaf[1] with the aim of identifying all the features that are to be processed. The concepts in the

[1] *Concept of type Leaf:* A concept does not contain sub concepts and allows being instantiated.

Domain Ontology represent the samples and the classes that will be used in the classification process.

1.2. To build a vector of occurrences and absences of the features present in classes taking into account the features extracted in step (1.1).

2. For the Data-Representation Nodes (DRN):

2.1. To extract the features present in each Data Representation Nodes.

2.2. To build vector occurrences and absences from the features present in the DRN, taking into account the features extracted in step (1.1).

3. To classify the DRN with respect to the classes present in the Domain Ontology using the classification process that is presented in the next section.

Since these vectors represent the occurrence or absence of features it is very convenient to represent these vectors with binary values. These values are usually encoded with one or zero denoting whether the property exists or not in the datum or the sample of the class. With these steps we automatically provide higher levels of semantic abstraction and definition of the role the data is playing according to the Domain Ontology being used.

5.1 Classification of Data Representation Nodes (DRN)

There are several techniques for data classification; between these techniques we can find those that use distances. The k-NN [15] is an example of a classifier based on distances that can even work with dissimilarities that do not meet metric properties such as symmetry or the triangle inequality. In essence, when it comes to classifying new data, the method calculates the distances to all the classes, then sorts the distances and assigns to the new data the label of the class that had the smallest distance to it; this means that the smaller distance between the data and the class, the bigger will be the correspondence.

As we only have one sample per class, where the sample "i" and the class "i" are precisely represented by the concept "i" in the Domain Ontology, then the K-NN classifier with K=1 (1-NN) is used. It is precisely for this reason that the 1-NN classifier is proposed. It is suitable furthermore to establish a threshold for classification, in order to avoid the risk of assigning a class with a low probability of being the correct one. In this way it is ensured that classification is made offering a certain guarantee, therefore a datum will not be classified in a class if it is not likely to belong to this class. To this end the 1-NN with reject is used, this variant of 1-NN classification excludes those data for which the threshold was not reached.

On the other hand, for computing the distances between objects there are a several measures that differ essentially in the type of data for which they have been designed. These measures are grouped into two main groups: Similarity Measures and Dissimilarity Measures.

- *Similarity Measures:* Measures that make more emphasis on the nearness between objects, where smaller values indicate that the elements are more different.
- *Dissimilarity Measures:* Measures that make more emphasis on the remoteness between objects, where smaller values indicate that the elements are more similar.

In our case, to classify a new DRN "n" is equivalent to finding the class whose distance is minimal with respect to "n", therefore the use of a dissimilarity measure to calculate the resemblance between DRN and classes is proposed. This dissimilarity measure is explained in details in the next section.

5.2 Dissimilarity Measure

In general, there is not a dissimilarity measure for all kinds of data, this must be chosen or adapted depending on the kind of data of the problem at hand. In the case of DRN classification we can take into account the present features in the *Thematic Component* since it represents the most adequate point of contact, at semantic level, between DRN and its conceptualization in the Domain Ontology.

As explained above, for the classification, the DRN and classes will be represented by vectors that will contain the occurrence or not of common characteristics between them, thus simplifying the problem in the sense of working with binary data. Among the measures to calculate dissimilarities between binary data, it was chosen the Simple Matching Distance [16] due to the nature of binary vectors with which we are working. As its name suggests this dissimilarity is a distance. It uses the coefficient shown in equation (1) and the distance is defined as shown in expression (2):

$$S_{(T,K)} = \frac{a+d}{a+b+c+d} \qquad (1) \qquad\qquad D_{(T,K)} = 1 - S_{(T,K)} \qquad (2)$$

Where:

- T: Classi.
- K: DRNp.
- S(T,K): Simple Matching Coefficient between Classi and DRNp.
- D(T,K): Simple Matching Distance between Classi and DRNp.
- a: Number of properties where T and K have an occurrence.
- b and c: Number of properties where T and K have different values.
- d: Number of properties where T and K have an absence.

This measure should be used in data in which the existence of occurrences and absences of the same features have a significant contribution in the classification. From the other distance measures for binary data it could also be used the Jaccard´s distance, but this measure only considers the common features for both individuals, which means that important information could be discarded in the classification process. Therefore the type of measure to be used must be selected depending on the nature of data.

The classification of geographical data based furthermore on its semantic abstraction tends to be more robust than the classification based only on spatial and geometrics features, since e.g., different data as Soil and Geology can be represented in the same form from the spatial and geometrical point of view and visually they are similar objects but from the semantic point of view they are different.

6 Experimental Results

In order to be able to represent automatically the semantic abstraction of geographical data a case study to illustrate the above methods is shown in this section. For this case study, Soil and Geological data layers were taken from the stored database in Spatial Data Infrastructure of the Republic of Cuba (IDERC)[17] (Geology.shp and Soil.shp). These data were stored in ESRI Shapefile format. This is a geospatial vector data format developed by ESRI company[18] and lately it has become de facto standard format for the exchange of geographical data. For each geographic object stored in the shapefile layers the information stored in both files * .dbf and * .shp were extracted for the creation of the Thematic and Spatial component in the NRD. The Temporal component has not been taken into account since these data lacks temporal characteristics. The ESRI Shapefile lacks the capability for the topologic information storage, therefore the relations between these objects has been defined based on 9-Intersections model [19] and the topological relationships proposed in the paper of JDARE'10 event[20].

To determine the semantic abstraction of data in the classification process we used a Land Cover Ontology, where apart from other concepts we can cite Soil and Geology concepts. Soil data and Geology data have been chosen to illustrate this study case, precisely because these data are similar from a geospatial representation point of view but not from a semantic point of view. Samples of classes, as noted above, are represented by each concept in the Domain Ontology, e.g. the Soil concept in the Land Cover Ontology itself represents the class of Soil and the sample of the Soil class. These samples for Soil and Geology classes contain the following features:

- *Soil Class:* ID, NAME, GROUP, TYPE, TEXTURE, EROSION, ACIDITY and SALINITY.
- *Geology Class:* ID, CODE, NAME, DESCRIPTION, AGE, TEXTURE and HARDNESS.

These classes contain some common characteristics, since they are likely to happen. The vector extracted per class contains the occurrences or absences of the common features that have been taken into account and it has constructed the following array of features per class, see Table.1:

Table 1. Vectors of occurrences and absences of common features per classes

Features / Classes	*id*	code	*name*	description	age	Texture	…	salinity
Geology	1	1	1	1	1	1	…	0
Soil	1	0	1	0	0	1	…	1

All the used data in the classification contains a subset of all these features. Each row of Table 3 represents a data layer that belongs to Geological or Soil classes; see Table.3 and Fig. 5, in which these features were collected. Therefore, a good classification of the layer represents a good classification of the objects therein. Table 2 shows the number of geographical objects both Geological and Soil layers.

The classification accuracy using 1-NN classifier of Distools toolbox[21] and using the structures "dataset" of PRTools [22] was 100% for all layers analyzed. Fig.7 shows the vectors of dissimilarity of each data layer regarding Geology and Soil classes, which shows that the layers belonging to the same class are grouped.

Table 2. Number of geographic objects both Geological and Soil layers

Geology Layer		Soil Layer	
Layer Name	Number of objects	Layer Name	Number of objects
geoLayer_1	928	sueLayer_1	307
geoLayer_2	1145	sueLayer_2	385
geoLayer_3	712	sueLayer_3	524
geoLayer_4	1098	sueLayer_4	233
geoLayer_5	977	sueLayer_5	189

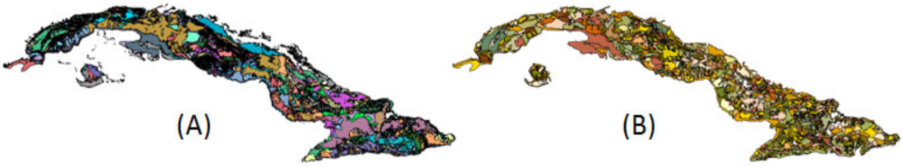

Fig. 5. Graphic representation of data layers of Geology (A) and soil (B)

Table 3. Vectors of occurrences and absences of common features per Layers

Features / Classes	id	code	name	description	age	Texture	…	salinity
geoDat_1	1	0	1	0	1	1	…	0
…	…	…	…	…	…	…	…	…
geoDat_5	0	1	1	1	1	1	…	0
sueDat_1	1	0	1	0	0	0	…	1
…	…	…	…	…	…	…	…	…
sueDat_5	1	0	0	0	0	1	…	1

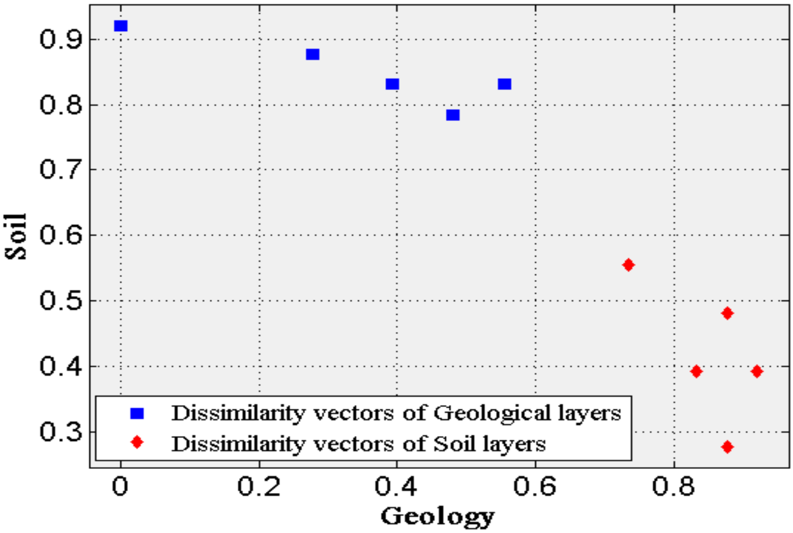

Fig. 6. Graphic representation of dissimilarity vectors of each layer respecting to Geology and Soil classes

This case study demonstrates the feasibility of automatically representing the semantic abstraction of geographical data, which means a step forward in the integration and processing of geographic data from a semantic point of view.

7 Conclusions

The work with geographical data from a semantic point of view and with the use of Ontologies as a way to represent knowledge, allows new and better ways of analyzing and exploiting these data. These ways undoubtedly improve existing tasks in the conventional Geographical Information Systems such as information retrieval and decision making. Nowadays it is necessary to create new mechanisms to represent automatically the semantic abstraction of geographical data given the great volume and heterogeneity that they present. The Ontologies that have been proposed in the literature cannot represent the characteristics and relationships that may exist in the geographical data integrated on GIS. We consider that the use of the information extracted from these data can improve the conventional task like the analysis task and/or information retrieval. For that reason in this paper a new type of Ontology has been proposed (*Data-Representation Ontology (DRO)*), which is automatically generated from the user data using the algorithms also proposed in this paper. The DRO represents the semantic abstraction of user data providing a bigger degree of specialization in the results and therefore a bigger granularity from interrelationship between the DRO and DO. Furthermore we have proposed a method to provide with more abstraction levels in the geographical data across the use of Domain Ontology based on classification techniques based on distances using a dissimilarity measure, something which is currently being done by hand.

References

1. Leung, Y.: Knowledge Discovery in Spatial Data. Springer, Heidelberg (2010)
2. Visser, U.: Intelligent Information Integration for the Semantic Web. LNCS. Springer, Heidelberg (2004)
3. Kavouras, M., Kokla, M., Tomai, E.: Comparing categories among geographic ontologies. In: Computers & Geosciences, Special Issue, Geospatial Research in Europe: AGILE 2003 (2003)
4. Schwering, A., Raubal, M.: Spatial relations for semantic similarity measurement. In: Akoka, J., Liddle, S.W., Song, I.-Y., Bertolotto, M., Comyn-Wattiau, I., van den Heuvel, W.-J., Kolp, M., Trujillo, J., Kop, C., Mayr, H.C. (eds.) ER Workshops 2005. LNCS, vol. 3770, pp. 259–269. Springer, Heidelberg (2005)
5. Hakimpour, F.: Using Ontologies to Resolve Semantic Heterogeneity for Integrating Spatial Database Schemata. PhD thesis Zurich University (2003)
6. Hess, G.N., Iochpe, C.: Ontology-driven resolution of semantic heterogeneities in gdb conceptual schemas. In: Proceedings of the GEOINFO 2004: VI Brazilian Symposium on GeoInformatics (2004)
7. Fonseca, F.T.: Ontology-Driven Geographic Information Systems, The University of Maine (2001)

8. ESDIG. Diccionario del Espacio Digital Geografico ESDIG (2010), http://infoteca.semarnat.gob.mx/website/diccionario/diccionario_d.html (cited 2010 Enero)
9. Pekalska, E., Duin, R.P.W.: The dissimilarity representation for pattern recognition. Foundations and Applications 64 (2005)
10. Lehmann, F.: Semantic networks. Computers Math. Applic. 23, 1–50 (1992)
11. Minsky, M.: A framework for representing knowledge. In: Winston, P.H. (ed.) The Psychology of Computer Vision. McGraw-Hill, New York (1975)
12. Gruber, T.: Ontolingua: A mechanism to support portable ontologies. Stanford University, Stanford (1992)
13. Studer, S., Benjamins, R., Fensel, D.: Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering (1998)
14. Guarino, N.: Formal Ontology and Information Systems. In: Proceedings of FOIS 1998. National Research Council, LADSEB–CNR (1998)
15. Fix, E., Hodges, J.L.: Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas (1951)
16. Backhaus, K., et al.: Multivariate analysis methods. In: An application-oriented introduction. Springer, Berlin (2000)
17. IDERC: Infraestructura de Datos Espaciales de la República de Cuba (2010), http://www.iderc.co.cu/ (cited 2010 Marzo)
18. ESRI: ESRI Home Page (2010), http://www.esri.com/ (cited 2010 Enero)
19. Egenhhofer, M.J.: A model for detailed binary topological relationships. Geomatica 47(3&4) (1993)
20. Larin-Fonseca, R., Garea-Llano, E.: Topological Relations as Rule for Automatic Generation of Geospatial Application Ontology. In: Proceedings of VII Jornadas para el Desarrollo de Grandes Aplicaciones de Red (2010) (in press)
21. Duin, R.P.W., et al.: DisTools A Matlab Toolbox for Pattern Recognition Delft Pattern Recognition Research, Faculty EWI - ICT, Delft University of Technology, The Netherlands (2009), http://prtools.org
22. Duin, R.P.W., et al.: PRTools4 A Matlab Toolbox for Pattern Recognition Version 4.1.5. Delft Pattern Recognition Research, Faculty EWI - ICT, Delft University of Technology, The Netherlands (2009), http://prtools.org