Revista Cubana de Ciencias Informáticas
Vol. 8, No. 2, 2014
ISSN: 1994-1536 | RNPS: 0547
Pág. 132 - 143
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Inteligencia artificial
Recibido: 20/03/2014 | Aceptado: 11/04/2014

Recuperación de objetos geoespaciales utilizando medidas de similitud semántica

Recovery geospatial objects using semantic similarity measures

Neili Machado-García[1*], Lilibeth González-Ruiz[1], Carlos Balmaseda-Espinosa[2]

[1] Departamento de Informática. Universidad Agraria de la Habana "Fructuoso Rodríguez Pérez", Carretera Tapaste y Autopista Nacional, km 23 ½, San José de Las Lajas, Mayabeque, Cuba, CP 32700
[2] Departamento de Gestión de la Calidad, Universidad Agraria de la Habana "Fructuoso Rodríguez Pérez". Carretera Tapaste y Autopista Nacional Km 23 ½, San José de Las Lajas, Mayabeque, Cuba, CP 32700
[*] Autor para la correspondencia: {neili, lilibeth, cbalma}@unah.edu.cu

Resumen

En este artículo se propone una metodología basada en el procesamiento semántico de objetos geográficos para la clasificación de los suelos según la Nueva Versión de Clasificación Genética de los suelos de Cuba. El método se compone de cinco etapas: conceptualización, síntesis, procesamiento de la solicitud, recuperación y ordenamiento. Como resultado se obtiene un sistema de gestión semántica de información geoespacial que aplica la medida de similitud semántica de Resnik. Como caso de estudio se considera el municipio de San José de las Lajas ubicado en la provincia de Mayabeque.

Palabras clave: Objetos geográficos, ontologías, semántica espacial, similitud semántica.

Abstract

In this paper we propose a methodology based on the semantic processing of geographic objects for the classification of soils according to the New Version of Genetic Classification of soils of Cuba. The method consists of five stages: conceptualization, synthesis, queries processing, retrieval and management. The result is a system of geospatial information management applying semantic similarity measure of Resnik. As a case study considering the region of San Jose de las Lajas located in the province of Mayabeque.

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 2, 2014
ISSN: 1994-1536 | RNPS: 0547
Pág. 132 - 143
http://rcci.uci.cu

Keywords: *Geographic object, ontology, semantic similarity, spatial semantics.*

Introducción

La Evaluación de Tierras es un sistema de clasificación aplicado que evalúa la capacidad del suelo para su utilización óptima, es decir, obtener máximos beneficios con mínima degradación. Puede definirse como "cualquier método que mida, o sea capaz de predecir, el uso potencial de una tierra" (McRae y Burnham, 1981).

Los distintos tipos de suelos presentan propiedades muy diferentes, por lo que su respuesta frente a cada tipo de utilización es muy diferente y depende de estas propiedades, por tanto, conociéndolas se puede predecir el comportamiento del suelo frente a una determinada utilización. Para realizar cualquier toma de decisión acerca de la utilización o conservación de los suelos es necesario conocer el tipo de suelo que se está valorando y sus características.

En el mundo hay gran cantidad de información de levantamientos de suelos. Esos estudios constituyen el legado de las ciencias del suelo, por ello se necesitan esfuerzos para su localización y catalogación (arqueología de datos), preservación (rescate de datos) y conversión a formatos de Sistemas de Información Geográfica (renovación de datos) de manera que no desaparezcan (Rossiter, 2006).

En los últimos años, las técnicas de organización y búsqueda de la información geográfica han cobrado gran importancia para poder extraer de estos datos toda la información útil que sea posible. Los datos geográficos poseen características específicas que dificultan su manipulación, la ubicación espacial, o sea el estar en una localización sobre la superficie de la tierra referida a un sistema de coordenadas, la temporalidad y las relaciones espaciales con otros objetos o datos, además, presentan gran heterogeneidad y volumen de almacenamiento. Existe por tanto una creciente necesidad de encontrar una solución que permita la integración de datos geográficos de una manera mucho más abstracta en la que el conocimiento juegue un rol esencial, en la que se explote con mayor efectividad la información semántica existente que se encuentra embebida en los datos almacenados.

La comparación de los conceptos que describen el significado de los datos en fuentes de información distribuidas es una de las operaciones básicas para solucionar la heterogeneidad semántica. Prueba de ello es la tendencia que se manifiesta en numerosas investigaciones recientes que abordan esta problemática (Fonseca, *et al*., 2002; Aparício, *et*

al., 2006; Giger and Najar, 2003; Gómez-Pérez, 2002). Estas soluciones se basan principalmente en el uso de ontologías como mecanismo de representación del conocimiento.

Las ontologías han sido analizadas en la Geociencia como un procedimiento de estandarización que facilita la traducción entre diferentes fuentes de información (Chandrasekaran, *et al*. 1999, Smith 1999, Fonseca et al. 2002). El estudio de los sistemas de recuperación de información en el campo de las Geociencias, concentra sus esfuerzos en estudiar otras formas de representar el conocimiento, modificar la manera en que se almacenan y organizan los datos, así como la búsqueda de mecanismos que extraigan información controlando su precisión, para que correspondan de manera exacta o similar las respuestas arrojadas con las consultas que realizan los usuarios.
La similitud semántica es fundamental para el procesamiento semántico de datos geoespaciales. Establece el grado de interoperabilidad semántica entre los datos o los diferentes SIG y constituyen las bases para la recuperación y la integración de información semántica (Janowicz, *et al.*, 2007).

En los SIG la similitud es particularmente importante debido a la dificultad para obtener representaciones satisfactorias de los fenómenos geográficos y a la variedad de formalizaciones que existen de las propiedades espaciales tales como su forma, localización y relaciones espaciales (Fonseca, 2001).

En el presente trabajo se presenta un sistema de gestión semántica de información geoespacial, utiliza una ontología de la Nueva Clasificación Genética de los Suelos de Cuba y aplica la medida de similitud semántica de Resnik entre los conceptos representados para identificar y recuperar los elementos que comparten propiedades similares.
Este documento se organiza en cinco secciones principales. En la sección 2 se presenta un resumen de los trabajos más relevantes relacionados con las anotaciones semánticas de datos geográficos y las medidas de similitud. En la sección 3 se presenta la metodología propuesta. Luego en la sección 4 se comentan los principales resultados de una primera implementación y experimentación del sistema. Finalmente se exponen en la sección 5 las conclusiones y líneas futuras de investigación.

Trabajos relacionados

En la actualidad muchas investigaciones están encaminadas a encontrar la manera de codificar formalmente el contexto y las relaciones geográficas. Un tipo de relación de este ámbito son las que existen entre datos geoespaciales, el conocimiento entorno a estos y su interrelación.

En (Sotnykova, 2005) se propone una metodología para la integración de esquemas conceptuales espacio-temporales mediante modelos conceptuales y Lógica Descriptiva (LD). En (Hakimpour, 2005) se introduce una arquitectura y una metodología basada en LD para crear un sistema integrado de información geográfica. En el sistema GioNis (Stoimenov, 2006) se define una propuesta de ontología híbrida basada en una arquitectura semántica que combinada con LD permite descubrir correspondencias entre conceptos de diferentes ontologías. En (Aerts, 2006) describe una metodología para desarrollar un SIG integrado principalmente para resolver los problemas de heterogeneidad semántica en bases de datos topográficas.

En su propuesta (Li, B. y F. T. Fonseca, 2006) proponen una medida de similitud que integra cuatro modelos, el modelo geométrico, el modelo de características, el modelo de transformación y el modelo de alineación estructurada para calcular las igualdades y las diferencias entre escenas espaciales y a nivel de capa. Aplica el orden de prioridad topología, dirección, distancia y se disminuyen los costos de transformación. Ambas características son implementadas a través de la aplicación de los pesos.

Un enfoque de integración de las relaciones espaciales y mediciones de similitud semántica entre diferentes conceptos geoespaciales considerando que las relaciones espaciales son partes fundamentales de la descripción semántica de los geo-datos se presentan en (Schwering, 2005). Se seleccionan un conjunto de relaciones espaciales formalizadas en lenguaje natural según el modelo computacional de Shariff et al. El trabajo está enfocado en la medición de las distancias semánticas en el nivel conceptual. Un ejemplo de integración de información geográfica a nivel de sistema son los SIG gobernado por ontologías (SIGGO – del inglés ODGIS *Ontology Driven Geographic Information System*) que actúan como un sistema integrador independientemente del modelo (Fonseca, 2001).

El concepto de ontología ha atraído una atención creciente en la comunidad de las ciencias de la información debido a su capacidad para lograr una representación del conocimiento compartido. El uso de ontologías en la información geográfica tiende a ser diferente dependiendo de la perspectiva y objetivos de los usuarios (Winter, 2001).

Las ontologías tienen una gran importancia en la creación y uso de las normas de intercambio de datos, así como en la solución de problemas derivados de la heterogeneidad y poca interoperabilidad de los datos geográficos. Las ontologías pueden ser usadas como una alternativa para representar los datos y, de forma explícita, el conocimiento acerca de ellos.

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 2, 2014
ISSN: 1994-1536 | RNPS: 0547
Pág. 132 - 143
http://rcci.uci.cu

Para el procesamiento semántico del conocimiento de los datos geoespaciales almacenados en las ontologías es fundamental el cálculo de la similitud semántica, la cual es esencial en el procesamiento de las consultas de datos de los usuarios y es la base para la recuperación e integración de información semántica (Schwering, 2008). Para determinar la similitud entre dos entidades se analizan dos nociones fundamentales: las características comunes y diferentes y la distancia semántica (Schwering, 2008).

Metodología computacional

La metodología propuesta se basa en un procesamiento semántico que permite localizar un determinado tipo de suelo con determinadas características. Además de devolver la ubicación del suelo recupera imágenes de estos suelos permitiéndole al usuario una visualización de las características y propiedades de los suelos.

Aplica medidas de similitud semánticas y emplea un vocabulario estructurado que incluye la información geográfica referida al Recurso Suelo pero los resultados, desde el punto de vista metodológico, pueden ser válidos para cualquier otro tipo de IG. El área seleccionada para el estudio pertenece al municipio de San José de las Lajas ubicado en la Provincia Mayabeque en la región occidental de Cuba. Los suelos son clasificados de acuerdo con su estructura y composición. El método está compuesto por cinco etapas: Conceptualización, Síntesis, Procesamiento de la solicitud, Recuperación y Ordenamiento.

Marco de trabajo

La Figura 1 muestra cómo se lleva a cabo el proceso de gestión de los recursos para localizar un suelo con determinadas características realizando un procesamiento semántico para esto. Inicialmente se realiza la **Conceptualización** del dominio de trabajo describiéndolo a través de una jerarquía de objetos con sus relaciones y características.

Las ontologías son estructuras que pueden crecer, integrarse con otras ontologías, reutilizarse en la construcción de ontologías de otros dominios, estas características constituyen beneficios de las geoontologías como estructura de integración de datos y conocimiento espacial. El uso de las ontologías, como mecanismos para representar conocimiento de un dominio concreto, puede ayudar a un sistema de gestión a focalizar las consultas de los usuarios y a llegar a sitios donde la comparación sintáctica entre cadenas de caracteres no es suficiente.

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 2, 2014
ISSN: 1994-1536 | RNPS: 0547
Pág. 132 - 143
http://rcci.uci.cu

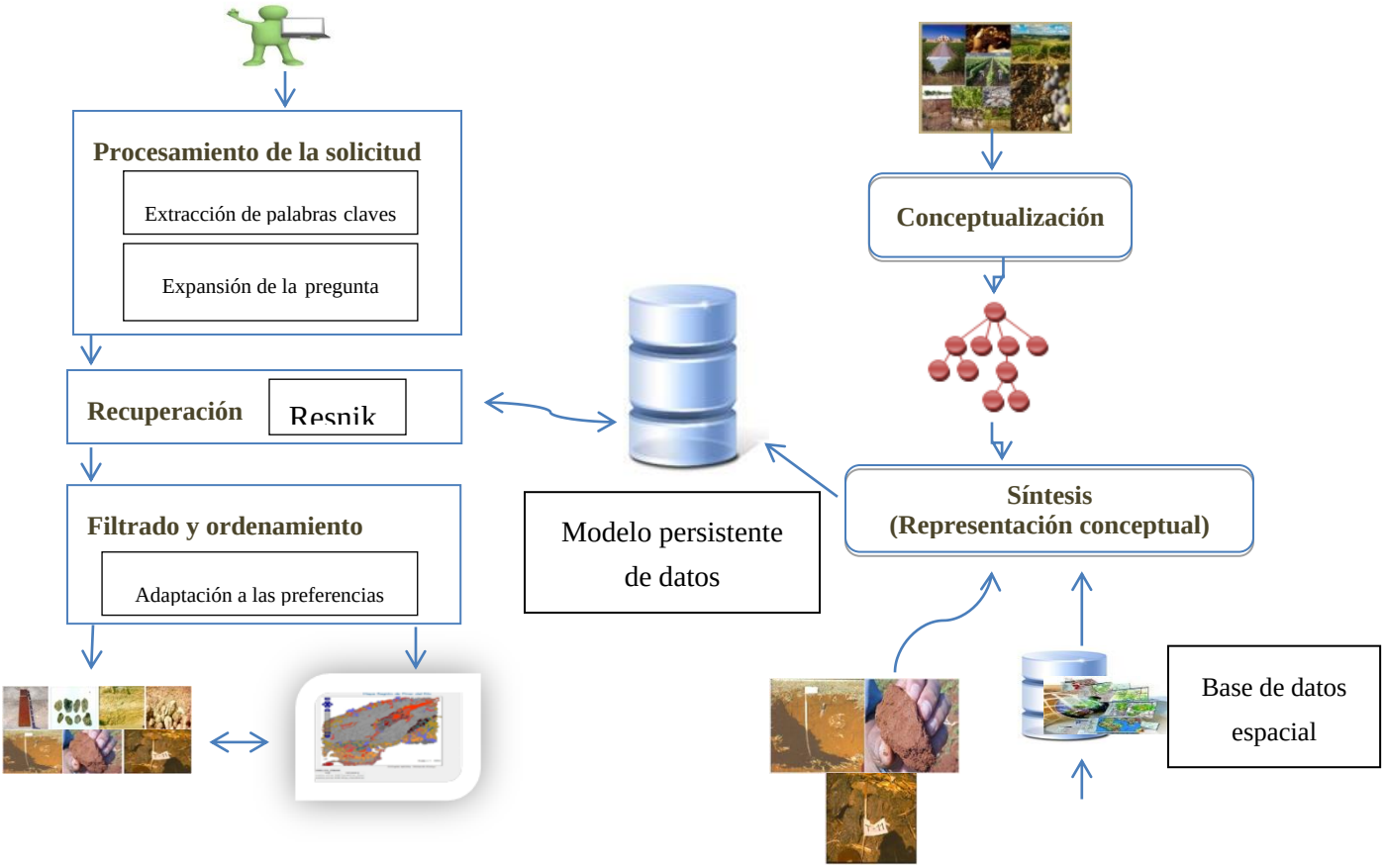

Figura 1. Arquitectura del modelo de gestión.

Estas estructuras proporcionan la vía para representar conocimiento. Son más que una taxonomía, en ellas los significados de los términos no son ambiguos, lo que los hace semánticamente independientes del usuario y del contexto. Identifican clases de objetos, sus relaciones y las jerarquías de conceptos dentro de un dominio específico.

El propósito de las ontologías no es servir de vocabulario o taxonomía sino el compartimiento y reusabilidad del conocimiento entre aplicaciones. En este trabajo se presenta una ontología que describe las características de los suelos de la Provincia de Mayabeque en la región occidental de Cuba siguiendo la Nueva Versión de Clasificación Genética de los Suelos (ver Figura 2).

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 2, 2014
ISSN: 1994-1536 | RNPS: 0547
Pág. 132 - 143
http://rcci.uci.cu

Figura 2. Fragmento de la ontología.

En la **Síntesis** se instancia la ontología con los datos almacenados en PostGis para establecer una relación de las imágenes que muestran los suelos según la clasificación con los mapas y los conceptos de la ontología. La integración de datos espaciales en ontologías debe realizarse mediante un proceso previo de anotación de los datos. La forma más simple de integrar datos espaciales a ontologías es almacenando completamente el dato en la ontología, es decir, los *bytes* que representan al dato espacial estarán embebidos en la estructura ontológica en algún apartado previsto para ello. Pero no basta con embeber el dato íntegramente en la ontología sino que es necesario hacer una correspondencia entre los objetos geográficos presentes en el dato y los conceptos vinculados con ellos.

Dentro del **Módulo de Procesamiento de la solicitud** del usuario se enriquece la consulta del usuario convirtiendo la consulta sintáctica en una representación semántica equivalente, a partir de las relaciones entre los conceptos de una ontología. El primer paso que sufre la cadena de búsqueda introducida por el usuario básicamente consiste en simplificar y normalizar las palabras de la consulta, eliminando palabras no relevantes (*stop words*), singularizando las palabras, etc. Para este paso se utiliza el analizador morfológico *Freeling* de la Universidad de Barcelona.

El segundo paso tiene por objetivo reconocer y activar en la ontología los conceptos subyacentes o similares a los términos de búsqueda introducidos por el usuario. Para identificar los conceptos similares en la ontología se utiliza el algoritmo de *Resnik* (1993). Este algoritmo es uno de los más destacados en el cálculo de la similitud semántica, el

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 2, 2014
ISSN: 1994-1536 | RNPS: 0547
Pág. 132 - 143
http://rcci.uci.cu

cual propone que la similitud entre dos conceptos c1 y c2 de una estructura taxonómica, puede ser obtenida mediante la ecuación (1).

$$sim\ (c1, c2) = \max_{c \in S(c1,c2)} (-\log p(c)) \quad (1)$$

Donde S (c1, c2) representa el conjunto de conceptos de los cuales tanto c1como c2 descienden. Mientras que p(c) es la probabilidad del concepto c. El término concepto hace referencia al conjunto de términos que apuntan a una misma idea. Ahora bien, para estimar la probabilidad de un concepto c, *Resnik* utiliza la frecuencia de aparición de los términos de ese concepto en el *Brown Corpus of American English* (Francis, 1982) pero debido a que era necesario utilizar un *corpus* que estuviera relacionado con los términos que se manejan en la ontología se decidió utilizar como corpus el libro "El suelo y su fertilidad" del Dr. Nelson J. Martín Alonso, del Departamento de Riego, Drenaje y Ciencias del Suelo de la Facultad de Agronomía.

Posteriormente con todos los conceptos obtenidos (iguales y similares) se ejecuta el **Módulo de Recuperación** realizando una búsqueda en la base de datos geográfica, las imágenes recuperadas están relacionadas semánticamente con los conceptos de la base de conocimientos y en los mapas se detalla su ubicación espacial. Al usuario se le muestran las imágenes del suelo que buscaba, o alguno con características similares, con su respectiva ubicación en el mapa. Si desea ver la imagen ampliada, la selecciona y además se le brinda la descripción de la imagen.

Sobre este listado de las imágenes ordenadas por la relevancia de los términos se aplica el **Módulo de Filtrado y Ordenamiento**. Cuando una imagen es adicionada a un repositorio sus características son extraídas y relacionadas con los conceptos de la ontología. Una vez que un usuario utiliza una imagen, estas características son almacenadas como preferencias del usuario y a partir de estas el usuario puede solicitar una nueva búsqueda, mostrándose finalmente las imágenes relacionadas con las preferencias y otras que están relacionadas a estas, pero que no están dentro de las preferencias del usuario. Finalmente, el sistema muestra listas ordenadas de imágenes con el mapa donde se encuentran localizados los suelos que cumplen los requerimientos del usuario. Además de mostrarle otros conceptos similares que pueden resultar de interés.

El usuario del sistema, por otra parte, tendrá un perfil con sus datos y preferencias almacenadas, obtenidas como resultado de la búsqueda de imágenes de suelos, que se irán almacenando según las búsquedas más frecuentes del usuario. Si ya ha realizado varias búsquedas y tiene preferencias en su perfil, entonces podrá hacer una búsqueda

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 2, 2014
ISSN: 1994-1536 | RNPS: 0547
Pág. 132 - 143
http://rcci.uci.cu

personalizada, donde al seleccionar las imágenes relacionadas semánticamente con esta se le mostrarán otras relacionadas con la preferencia que no estén en el perfil.

Resultados y discusión

Se experimentó con las herramientas desarrolladas, inicialmente fue necesario implementar una ontología de la Nueva Versión de Clasificación Genética de los Suelos de Cuba y realizar la anotación semántica de los mapas. El sistema de gestión resultante permite realizar búsquedas de información de los diferentes tipos de suelos. Al seleccionar la opción buscar, por ejemplo, si se lanza una búsqueda de determinado tipo de suelo "tipo ferralítico rojo", se recupera un mapa de la región en el cual se visualiza la localización de este tipo de suelo (ver Figura 3), además, se mostrarán todas las imágenes relacionadas con la solicitud y si el usuario lo desea al dar clic en alguna de estas imágenes esta se mostrará con una vista más amplia (Ver Figura 4), mostrando la descripción de la misma y los conceptos de la ontología que tienen relación semántica con la misma.

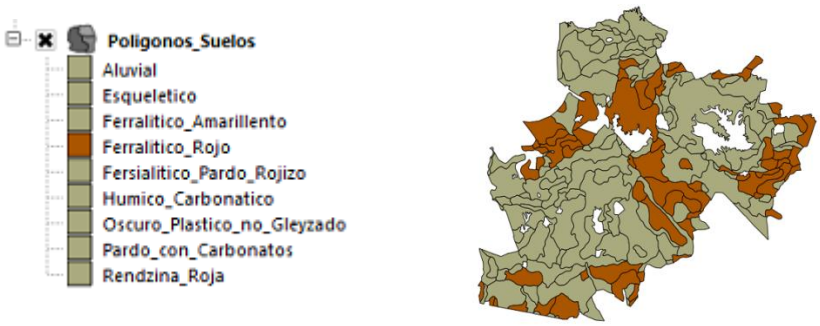

Figura 3. Visor del mapa de San José de las Lajas resaltando los suelos Ferralíticos Rojos.

Figura 4. Imagen recuperada del sistema una vez lanzada la búsqueda de suelos Ferralíticos Rojos.

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 2, 2014
ISSN: 1994-1536 | RNPS: 0547
Pág. 132 - 143
http://rcci.uci.cu

Tabla 1. Comparación de tiempos promedios de procesamiento de la información.

Acciones	Tiempo consumido con datos no anotados (mlseg)	Tiempo consumido con datos anotados (mlseg)
Recuperación de suelos con determinado porciento de humificación.	450	335
Recuperación de imágenes de diferentes tipos de suelos.	470	280
Recuperación del mapa con filtro Tipo de Suelo	520	300

Con la anotación semántica de los datos geográficos se obtuvieron mejores tiempos de recuperación y con las respuestas esperadas por el usuario. Este experimento muestra la factibilidad de la implementación y ejemplifica su aplicación con datos geográficos reales. La aplicación de la medida de similitud semántica de *Resnik* reduce el "silencio" o resultados nulos en el sistema de recuperación.

Conclusiones

En el presente trabajo se ha propuesto un sistema de gestión semántica de información geoespacial aplicando medidas de similitud en el proceso de recuperación. La aplicación de la medida de similitud semántica de *Resnik* hace que se evite retornar resultados vacíos, pero hay que ser cuidadosos con los resultados que pueden estar completamente alejados de lo que busca el usuario.

Al usar el análisis semántico es posible aproximarse a una forma de procesamiento inteligente, el cual no obedece exclusivamente a una coincidencia sintáctica pero sin olvidar las ventajas del enfoque geoespacial clásico.

Con este trabajo se contribuye a disminuir la carencia de sistemas que integren aspectos semánticos conjuntamente con aspectos espaciales, de allí que el modelo de integración presentado en este trabajo, expone el análisis semántico como un excelente complemento mutuo para las técnicas tradicionales de análisis geoespacial.

En futuros trabajos se experimentará la arquitectura propuesta en grandes volúmenes de datos y se aplicaran nuevas medidas de similitud para identificar los objetos geográficos. Es necesario encontrar elementos que permitan una mayor vinculación entre la semántica formal y la información necesaria para la visualización gráfica de la información geográfica.

Referencias

- AERTS, K., K. MAESEN, and A. van ROMPAEY, *A Practical Example of Semantic Interoperability of Large-Scale Topographic Databases Using Semantic web technologies. Proceedings of the AGILE'06, Visegr, Hungary* (2006) 35-42, 2009.
- AGARWAL, P., *"Ontological Considerations in GIScience." International Journal of Geographical Information Science* Vol. 19, No. 5, May 2005, 501–536
- FONSECA, F., *Ontology-Driven Geographic Information Systems. Ontology-Driven Geographic Information Systems*, 2001.
- HAKIMPOUR, F. and A. GEPPERT, *Resolution of Semantic Heterogeneity in Database Schema Integration Using Formal Ontologies. Information Technology and Management*, 2005. 6(1): p. 97-122.
- HESS, G.N. and C. IOCHPE, *Ontology-Driven Resolution of Semantic Heterogeneities in gdb Conceptual Schemas. Proceedings of the* GEOINFO'04: VI Brazilian Symposium on GeoInformatics, 2004: p. 247-263.
- JANOWICZ, K., RAUBAL, M., SCHWERING, A., and Kuhn, W. *Semantic Similarity Measurement and Geospatial Applications. In: Workshop at COSIT* 2007. Disponible en: [http://www.blackwell-synergy.com].
- KAVOURAS, M., M. KOKLA, and E. TOMAI, *Comparing Categories Among Geographic Ontologies. Computers and Geosciences*, 2005. 31(2): p. 145-154.
- LI, B. and F. T. FONSECA. "TDD - *A Comprehensive Model for Qualitative Spatial Similarity Assessment." Spatial Cognition and Computation* 6(1): 31-62. 2006.
- RESNIK, P., *Selection and Information: A Class-Based Approach to Lexical Relationships, Ph.D Dissertation*. 1993.
- ROSSITER, D., *Digital Soil Mapping as a Component of Data Renewal for Areas with Sparse Soil Data Infrastructures. In: Second Global Workshop on Digital Soil Mapping, July* 4-7, 2006, Rio de Janeiro, Brazil, 2006.

- SOTNYKOVA, A., *et al., Semantic mappings in description logics for spatio-temporal database schema integration, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), E. Zimanyi*, Editor. 2005. p. 143-167.
- STOIMENOV, L., A. STANIMIROVIC, and S. DJORDJEVIC-KAJAN, *Discovering Mappings Between Ontologies in Semantic Integration Process. Proceedings of the AGILE* 2006, *Visegrad*, 2006: p. 213-219.
- SHARIFF, A., EGENHOFER, M., MARK, D., 1998. *Natural-Language Spatial Relations Between Linear and Areal Objects: the Topology and Metric of English-Language Terms. International Journal of Geographical Information Science* 12 (3), 215–245.
- SCHWERING, A. *Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. Transactions in GIS*, 2008, 12(1): 5–29
- SCHWERING, A., RAUBAL, M. Spatial *Relations for Semantic Similarity Measurement. In: Proceedings of the ER'05: 24th International Conference on Conceptual Modeling. Lecture Notes in Computer Science*, vol. 3770. *Springer, Berlin, Heidelberg*, p. 259–269. 2005.