10

Integración de información geográfica mediante el uso de técnicas de alineamiento de ontologías

Integración de información geográfica mediante el uso de técnicas de alineamiento de ontologías

William Ernesto Guerrero Rodríguez[1]

Resumen

En los últimos años la integración de información geográfica ha cobrado especial relevancia, la idea de tener datos geográficos eficientes, sin redundancia e interoperables ha sido asumida como una forma de ampliar el potencial de los datos vistos como un capital de las organizaciones. Algo que evidencia esta preocupación es el desarrollo creciente de iniciativas como las IDES, la creación de la WEB-Semántica, la generación de estándares en los procesos de manipulación o administración de la información (ISO,OGC), entre otros. El problema de esta integración radica en que las fuentes de información geográfica cuentan con un alto nivel de heterogeneidad sintáctico, semántico y estructural. Actualmente, el uso de ontologías, entendida, como la especificación explícita de una conceptualización, se ha convertido en una herramienta eficaz que intenta unificar la expresión de un contexto en un determinado dominio. Comúnmente al integrar ontologías o elementos de distinta ontología es posible encontrar discrepancias que impiden su interrelación. Para hallar equivalencias entre elementos de una ontología, se ha planteado el uso de las denominadas técnicas de alineamiento de ontologías que pretenden definir métricas de similaridad entre objetos con el fin de establecer paridades entre los conceptos. La investigación se ha orientado a identificar las diversas técnicas de alineamiento de ontologías haciendo especial énfasis de su uso en el dominio geográfico, y la forma en que estas técnicas pueden aportar a la integración de información en los niveles de instancia y concepto en los SIG o en las BDE.

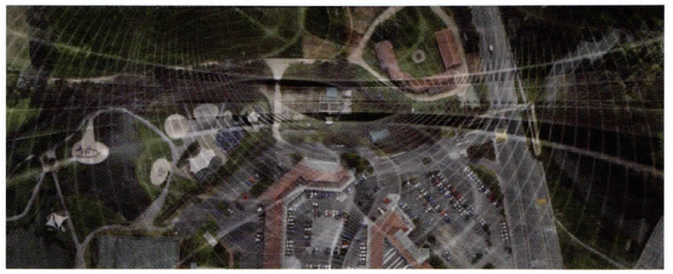
Palabras Clave:

Heterogeneidad, Integración, Ontologías geográficas, Alineamiento de Ontologías, Información Geográfica.

1 guerrero.william@gmail.com, Universidad Distrital Fransisco Jose de Caldas.

Abstract
In recent years the integration of geographic information has had special relevance, the idea of spatial data efficient, non-redundant and interoperable, has been assumed as a way to expand the potential of data seen as a capital of organizations. Something that shows this concern is the increasing development of initiatives such as the spatial data infrastructures SDI, the creation of the Semantic Web, the generation of standards in the process of handling or information management (ISO, OGC), among others initiatives. The problem of this integration is that geographic information sources have a high level of syntactic, semantic and structural heterogeneity. Currently the use of ontologies understood as the explicit specification of a conceptualization has become a powerful tool that attempts to unify the expression of a domain in a given context. Usually when you run the integration process of ontologies is possible to find differences in the elements that prevent the creation of relationships between ontologies. To find equivalences between elements of an ontology, have been defined ontology alignment techniques that aim to define a similarity metric between objects in order to establish parity between the concepts. The investigation has focused on identifying the various techniques of ontology alignment with particular emphasis on its use in the geographical domain, and how these techniques can contribute to the integration of information at the instance and concept in geographic information systems (GIS) or in spatial databases.

Keywords:
Heterogeneity, Integration, Geographical ontologies, Ontology Alignment, Geographical information.



Introducción

El problema de la heterogeneidad de la información ha sido tratado con gran interés por los investigadores de las ciencias de la información, esto se debe a la necesidad que ha surgido de integrar datos derivados de distintas fuentes con una alta heterogeneidad sintáctica, estructural y semántica. El problema se ha abordado desde distintas perspectivas y las soluciones han sido desarrolladas paralelamente a la evolución de las estructuras y dispositivos de administración de la información. En la actualidad el uso de mecanismos de representación de conocimiento como las ontologías son usados para servir como intermediario debido a su rica estructura semántica basada en teorías lógicas, que provee definiciones formales explícitas de entidades y sus relaciones, que pueden facilitar las definiciones de métodos para proyectar, trasladar e integrar información. A su vez se han desarrollado diversas técnicas que permiten obtener correspondencias conceptuales entre distintas ontologías con el objetivo de crear comunicación entre las fuentes de información, a dichas técnicas se les han llamado métodos de alineamiento de ontologías. Cuando se desean establecer correspondencias conceptuales en la heterogeneidad de conceptualización comentada, es posible abordar el problema desde distintas perspectivas; en el presente documento pretende mostrar cómo establecer estas correspondencias, aplicando técnicas de alineamiento de ontologías (tanto a nivel de instancia como de esquema) en las bases de datos geográficas desde un marco de comparación semántica.

Este documento tiene dos objetivos el primero mostrar las generalidades de las ontologías en el dominio geográfico y su uso en la integración de información, y el segundo mostrar un modelo comparativo a nivel semántica. Para cumplir con estos dos objetivos el documento se organizó así: En la primera parte se menciona el problema de heterogeneidad de la información, en la segunda se define una ontología haciendo especial énfasis de su uso en el dominio geográfico, en la tercera parte se muestran las principales técnicas y métodos de alineamiento, en la cuarta se muestran los trabajos más destacados en el área de alineamiento de ontologías a nivel geográfico centrándose en mostrar uso de instancias para la identificación de correspondencias semánticas, posteriormente se muestra algunos escenarios de comparación de información a nivel conceptual y, finalmente, se propone un modelo de comparación aplicándolo en BDE.

Aproximación al problema de heterogeneidad de la información geográfica

El modelamiento de cualquier estructura de organización de la información parte de la caracterización de cada uno de los elementos y relaciones percibidas de la realidad de acuerdo con un contexto específico, de esta forma la abstracción de un mismo fenómeno puede cambiar de acuerdo a la "visión" sobre la cual se construye el modelamiento, los objetivos y los requerimientos para la elaboración de la base de datos. De-

bido a la propia naturaleza del proceso de conceptualización antes mencionado, cada uno de los elementos y estructuras en general, mostrados en los esquemas de las bases de datos, presentan propiedades muy particulares.

En un ambiente de transferencia de información o de conocimiento, estas particularidades de diseño e implementación de los repositorios de datos, se perciben como un problema de heterogeneidad, en donde la particularidad de la estructura de los objetos, impide la comunicación entre elementos de distintas bases de datos. El problema de la heterogeneidad en bases de datos geográficas puede verse desde dos puntos de vista, uno derivado del proceso de conceptualización y otro asociado a la implementación.

La heterogeneidad asociada a la implementación se ha abordado desarrollando estándares que faciliten la transferencia y comunicación de información. Organizaciones como la ISO la OGC o la W3C (para datos y repositorios web) han tenido especial relevancia, pues han definido procedimientos y formatos que permiten la comunicación entre distintas fuentes de información geográfica o los distintos repositorios, en este sentido formatos como GML (Geography Markup Language) para datos geográficos y esquemas de bases de datos geográficas, XSD (XML Schema Definition) o XMI (XML Metadata Interchange) para esquemas en general, permiten un lenguaje común que vuelve transparente el intercambio de información desde el punto de vista de la implementación.

La heterogeneidad asociada a la conceptualización hace referencia a las diferencias en el modelado y/o abstracción de la realidad, es decir, a las discrepancias en la forma en que un mismo concepto puede ser modelado. En términos de ingeniería de dominio esto hace referencia a medir las diferencias semánticas entre dos conceptos afines en esquemas conceptuales distintos [1], esto implica determinar los "rasgos semánticos" y/o características que especifican un concepto y que hacen que sea interpretado como un determinado elemento en una base de datos. Como se puede observar el problema se ha afrontado desde una perspectiva de las ciencias cognitivas y desde allí se ha desarrollado.

En la web el problema de la heterogeneidad es el principal inconveniente a la hora de consultar y recuperar información, por este motivo actualmente una gran parte de la comunidad investigativa de las ciencias de la información ha venido orientando sus trabajos al desarrollo de mecanismos que aumenten la calidad de las consultas en la web, con el objetivo de desarrollar lo que se conoce como web semántica. Para facilitar este proceso de consulta se han desarrollado estructuras de modelado de información conocidas como ontologías que manejan una estructura jerárquica entre conceptos que puede llegar a ser muy útil a la hora de acceder a la información. Las ontologías como elementos orgánicos de la realidad modelada son los elementos fundamentales que representan las unidades conceptuales.

Cuando son utilizadas dos o más ontologías (o geo-ontologías en un ámbito geográfico) para describir a un objeto o dominio geográfico, la información contenida en estas puede representar a dicho objeto de manera diferente en cada una de las geo-ontologías, generando la posibilidad de que esas representaciones no sean reconocidas como equivalentes, lo que implica nuevamente enfrentarse al problema de la heterogeneidad asociada a la conceptualización. Con el fin de establecer un conjunto de objetos equivalentes pertenecientes a diferentes ontologías, se han desarrollado procedimientos que faciliten relacionar estos objetos y encontrar correspondencias entre los conceptos pertenecientes a ontologías diferentes; a dichos procedimientos comúnmente se les llama alineamiento de ontologías.

Existen diversas técnicas que persiguen un alineamiento de ontologías dependiendo de las características semánticas que se estudien; sin embargo, de forma general, el alineamiento de geo-ontologías se trabaja en dos categorías [2]: A nivel de concepto y a nivel de instancia. Como se ve el problema de determinar conceptos semejantes ya sea en esquemas o en ontologías es el mismo; en este sentido, podríamos afirmar que las técnicas de alineamiento de ge-ontologías pueden ser usadas para encontrar elementos correspondientes en bases de datos geográficas.

Ontologías en el dominio geográfico – Geontologías

Aunque aún no se tiene una definición común para esta expresión, en el ámbito de administración de la información, de forma general el término ontología hace referencia a una organización conceptual detallada dentro de uno o varios dominios, con el objetivo de facilitar la comunicación de la información entre diferentes sistemas o conjuntos de datos. La palabra ontología en este contexto en realidad hace referencia a una definición filosófica, en donde ontología se entiende como el estudio de la esencia o sustancia de los fenómenos, ocupándose de la definición del ser y de establecer las categorías fundamentales a partir de sus propiedades, estructuras y sistemas. En ciencias de la información se han acuñado otras definiciones como la dada por Gruber [3] quien define que: "*Una ontología es una especificación explícita de una conceptualización*"; como vemos, esta definición presenta un alto nivel de abstracción. De manera formal podríamos decir que una ontología se define como una tupla de la forma $O = < C, R, I, A >$ donde C es el conjunto de conceptos, R es el conjunto de relaciones, I el conjunto de instancias y A el conjunto de axiomas.

En una ontología, un concepto se asocia a una clase, las clases están organizadas en una taxonomía, generalmente en una jerarquía de subsunción de conceptos mediante relaciones semánticas. Una ontología en el dominio geográfico se caracteriza porque los conceptos contenidos se presentan en una posición espacial; un ejemplo de un concepto geográfico podría ser un bosque, un río o un volcán; estos conceptos, además de asociar características temáticas, siempre se asocian a un espacio geográfico. En la figura 1 puede verse un ejemplo de una ontología geográfica presentada en un diagrama.

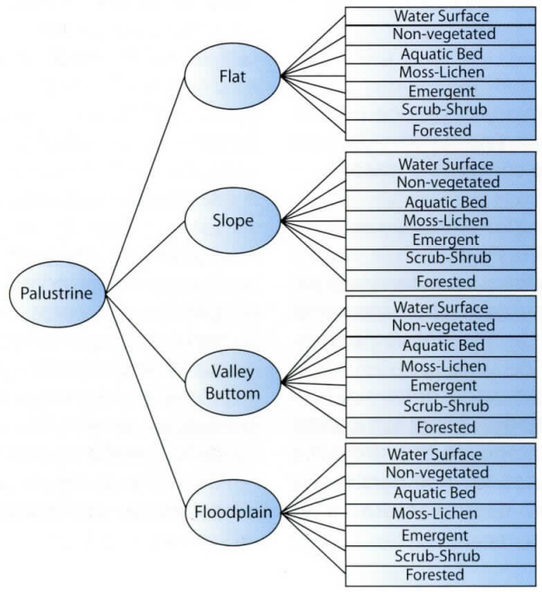


Figura 1.
Ontología geográfica de las áreas pantanosas.

Fuente:
Extraído de Cruz, Sunna y Makar, 2007.

Las relaciones son asociaciones binarias entre conceptos, estas relaciones binarias son las que les dan sentido al mundo abstraído en el modelado ya que a partir de ellas se realizan inferencias del comportamiento entre los conceptos. En las ontologías geográficas existe una clase especial de relación llamada relación espacial, estas relaciones especifican cómo están ubicados los elementos en el espacio en relación con los otros objetos. Desde este punto de vista, plantea restricciones de comportamiento de los objetos observados en la realidad. Para construir las relaciones espaciales es necesario obtener las relaciones topológicas, entre las entidades o conceptos geográficos, las cuales se refieren a propiedades tales como la conectividad, la adyacencia y la intersección entre clases geoespaciales.

En un contexto de ontologías, los atributos son vistos como una relación especial en donde su destino es un tipo de dato ya sea string un integer o un double, mientras que en una relación común el rango de relaciones es una clase. En un objeto geográfico se presentan atributos o propiedades muy particulares como la posición espacial y la geometría, las cuales definen de una forma u otra el comportamiento de un elemento geográfico y su abstracción de realidad.

Las instancias son entendidas como elementos o individuos de una ontología, es decir, que representan un caso particular de un concepto o clase.

Los axiomas son afirmaciones que siempre son verdaderas y definen la restricción de las relaciones en el contexto del problema y son usadas para determinar la consistencia de una ontología; desde este punto de vista, pueden verse como restricciones que determinan el comportamiento del objeto. Un axioma común en el campo de la hidrografía podría ser: "Todos los ríos tienen una nacimiento y una desembocadura".

En el dominio geográfico las ontologías son usadas para modelar cada uno de los fenómenos geográficos, representando los conceptos y sus relaciones, provee descripciones formales de los conceptos.

Técnicas de alineamiento de ontologías

La clasificación de los métodos de alineamiento de ontologías están directamente relacionados con los tipos de heterogeneidades que se pueden encontrar en la comparación de dos conceptos, o por la forma en que los mapeos o alineamientos se hacen, según la Knowledge Web Consortium (KWC), el alineamiento de ontologías a nivel local (que será el objeto del presente trabajo) se pueden clasificar en:

Métodos terminológicos

Estos métodos parten de la premisa que si dos clases tienen el mismo nombre hacen referencia a un mismo concepto, de esta forma su objetivo se centra en obtener el significado de la etiqueta que representa un determinado concepto en un nivel de granularidad sin ambigüedades, o por lo menos que las etiquetas provenientes de entidades en ontologías diferentes sean comparables. Estos métodos terminológicos a su vez pueden dividirse en dos: métodos basados en cadenas y métodos basados en el lenguaje.

- ***Métodos basados en cadenas:*** Estos miden las semejanzas entre las cadenas de caracteres de las etiquetas de los nombres que describen los conceptos. Existen muchos caminos para comparar las cadenas, de caracteres dependiendo de cómo la cadena sea vista (como una cadena de letras exacta, un error de secuencia de letras, un conjunto de letras, o como un conjunto de palabras). Los métodos más usados para determinar similaridades entre cadenas son: String equality, Hamming distance, Substring test , Substring similarity, N-gram distance, Edit distance, Jaro similarity. [4]

- ***Métodos basados en lenguaje*:** Los métodos basados en lenguaje usan técnicas de procesamiento del lenguaje natural para encontrar asociaciones entre instancias de conceptos o clases. Estos métodos pueden ser intrínsecos (usando las propiedades lingüísticas de las instancias como la morfología y propiedades sintácticas) o extrínseca (requiriendo el uso de fuentes externas).

Métodos Intrínsecos: Estos métodos se centran en realizar la adecuación terminológica, con la ayuda de análisis morfológicos y sintácticos. Son utilizados con frecuencia en la recuperación de información para mejorar la búsqueda. El estudio morfológico se enfoca principalmente en identificar sufijos, prefijos y pluralidad, identificando declinación o conjugación; el problema al usar algoritmos que tomen en cuenta variaciones morfológicas es que su uso se ve limitado, por el manejo morfológico que se le den a cada una de las frases en cada idioma. Como se ve, los análisis de orden lingüístico intrínseco responden a las variaciones de la forma en que las palabras que representan los términos pueden ser expresadas dependiendo de las conjugaciones o contextos que se presenten.

Métodos extrínsecos: Estos métodos que hacen uso de recursos externos tales como tesauros o diccionarios y/o lexicones o considerando las relaciones semánticas entre palabras como la sinonimia, la hiponimia, meronimia. Generalmente cuando se utilizan estos métodos extrínsecos se analiza con detalle la jerarquía taxonómica que presentan los términos en cada uno de los diccionarios utilizados, es decir, que las medidas de distancia o similitud semántica se calculan en función de la estructura manejada por el diccionario. En la actualidad el diccionario más utilizado para las consultas terminológicas es el WORDNET; sin embargo, en el contexto biofísico ambiental también ha sido usado el GEMET (General Multilingual Environmental Thesaurus).

Métodos estructurales

Los métodos estructurales pueden ser subdivididos en los que consideran la estructura interna y la estructura relacional.

- ***Métodos basados en estructura interna:*** Estos métodos son entendidos como métodos basados en restricciones. Estos usan el nombre de propiedades (los atributos o las relaciones), su cardinalidad, o su tipo de dato para calcular una similitud. Regularmente son combinados con algún método terminológico.

- ***Métodos basados en estructura relacional:*** Comparan la estructura de las entidades que pueden encontrarse. La estructura de ontología puede ser tratada como un grafo etiquetado, donde la etiqueta (una arista) es el nombre de la relación. La comparación de la similitud de los dos nodos que pertenecen a las ontologías diferentes, respectivamente, se basa en su posición dentro del grafo al cual pertenece. Esta idea se basa en que si dos nodos de dos ontologías son similares a sus vecinos y los vecinos son similares, entonces los nodos en cuestión deberían ser similares.

Métodos Extensionales

Los métodos extensionales o basados en instancias calculan la similitud entre individuos (las instancias). Estos méto-

dos intentan inferir la relación entre los nodos de las ontologías analizando sus instancias, ya sea determinando patrones en las instancias o hallando subconjuntos de valores.

Trabajos destacados en la integración de información geográfica desde una perspectiva de alineamiento de ontologías

La alineación de ontologías ha sido estudiada por diferentes autores en la medida en que el problema de la integración de información heterogénea ha cobrado importancia. Como se verán las técnicas de alineamiento no sólo son usadas para alinear ontologías sino que han sido extendidas a la integración de otras estructuras y modelados que soporten información geográfica como los esquemas de bases de datos geográficas o los geoservicios.

En general, los métodos de alineamiento utilizados en ontologías convencionales pueden ser usadas para alinear ontologías geográficas a nivel de conceptos como PROMPT o SMATCH [5] Sin embargo, las principales limitaciones de estos métodos es que no toman en cuenta las características particulares del dominio geográfico. Alrededor de este tema se han desarrollado múltiples, investigaciones, a continuación citaremos las más representativas:

Uno de los primeros trabajos en el área de alineamiento de ontologías en el dominio geográfico fue realizado por Rodríguez, Egenhofer, y Rugg [6], en donde se establece una medida de similaridad semántica asimétrica, el cálculo de similitudes se basa en los rasgos distintivos de las mismas entidades, como las funciones y los atributos. Para definir esta medida de similitud usa relaciones semánticas "parte - todo". Para la construcción de las relaciones semánticas toma en cuenta conceptos lingüísticos entre las palabras y los significados, como la sinonimia y la polisemia (homonimia).

Voltz [7] en su trabajo menciona la importancia que adquiere la integración de bases de datos geoespaciales de distintas fuentes dentro del marco de las infraestructuras de datos espaciales. Divide la integración en dos, una a nivel de esquema y otra a nivel de objeto, en la primera menciona que se fundamenta en la identificación de elementos semánticamente correspondientes en diferentes esquemas, en la segunda se enfoca en aprovechar el nivel de instancia de las representaciones para hallar similitudes entre diferentes esquemas, obteniendo de esta forma medidas de correlación que describen el grado de correspondencia entre los objetos de conjuntos de datos diferentes.

Schwering y Raudal [8] a través de la definición de un espacio geométrico, representan los conceptos geoespaciales cada uno como una región convexa dentro de un espacio vectorial, donde cada una de las características que define un concepto es tomado como una dimensión; de esta forma cada objeto espacial presenta dimensiones. Cada una de las medidas de similaridad son obtenidas a partir del promedio entre cada vector de las dos regiones comparadas; de esta forma, según lo sugieren logran obtener similitudes semánticas con resultados más precisos. Para efectos prácticos cada espacio conceptual es tomado de las instancias de un concepto.

Duckham y Worboys [9]. Su trabajo se centra en el problema de automatización del proceso de fusión de información geográfica, para ello usan las instancias de los conceptos y definen un método algebraico que toma en cuenta la distribución espacial de los datos. En su trabajo asumen que la información temático espacial se comporta como un Semilattice[2] y es comparado con una jerarquía de conceptos, de acuerdo a este planteamiento es posible generar un conjunto combinado producto de la fusión de dos semilattice a partir de la generación de una función que proyecta las relaciones topológicas de contenencia a las relaciones temáticas de los objetos comparados. De esta forma la fusión y las clases resultantes son producto de las proyecciones que se efectúen sobre la asociación de cada elemento en el semilattice y de la relación de contenencia.

En el año 2006 en el VIII simposio Brasilero de Geoinformática, Nudelman, Lochpe, y Castano [10], presentaron un algoritmo (G-Match) para realizar alineamiento de ontologías geográficas, este algoritmo consta principalmente de tres fases en el cálculo de medidas de similitud. En la primera, el nombre de los conceptos y los atributos son comparados usando métodos terminológicos de dos formas una por WordNet y la otra comparando cadenas de caracteres. En la segunda fase, una vez conocidas las medidas de similitud de los nombres, se establece una similitud de la estructura taxonómica de los conceptos dentro de las ontologías, las relaciones entre clases y las relaciones topológicas. En la última fase se establece una función de integración de similaridades a partir de la suma ponderada de las similitudes calculadas con anterioridad. Lo que más se puede resaltar en esta propuesta, es el uso de las relaciones topológicas que enriquece el proceso de comparación entre entidades y que introduce reglas del espacio geográfico dentro de la comparación de conceptos.

Nudelman y Iochpe [11] proponen la utilización de ontologías como mediadores para la integración semántica en esquemas de bases de datos geográficas, en el documento plantean una arquitectura de software que maneja la integración de heterogeneidades sintácticas usando un lenguaje estándar GML. Además, muestran adaptaciones a métodos matemáticos para la medición de conceptos y la aplicación a esquemas conceptuales en bases de datos geográficas (BDG). Proponen un algoritmo determinístico y secuencial que establece medidas de similitud semántica en cada una de las características de las entidades conceptuales. En un inicio se crea una ontología a partir de un esquema conceptual 1, posteriormente cada una de las características de los conceptos del esquema 2 son comparados. El algoritmo primero verifica si el nombre del concepto del esquema comparado se presenta en la ontología, si se establece un concepto afín se crean medidas de similitud a nivel de atributos estableciendo las diferencias, si el nombre del concepto no es encontrado se realizan medidas de similitud semántica asociadas a las relaciones, nombres y atributos, estableciendo los posibles conceptos afines, si la medida de similitud total (es decir la función de similitud que integra todas las medidas) supera un umbral definido, los conceptos afines se establecen, si la medida está cercana al umbral, un experto (usuario) verifica que se encuentren conceptos afines, si por el contrario las medidas de similaridad no superan el umbral, se adiciona el concepto en la ontología. El resultado final del proceso es una ontología que representa la integración de los dos esquemas conceptuales. Con esta ontología se establece la comunicación entre las dos bases de datos.

Navarrete [12] en su tesis doctoral propone una integración de información temática aplicada a un contexto mul-

2 Desde el punto de vista de conjuntos se puede entender como un conjunto parcialmente ordenado en donde cualquiera de los elementos contenidos en él tienen un máximo o un mínimo valor de jerarquía dependiendo de la relación binaria que se señale ya sea subsumsión o agrupación.

timedia. Define un marco de trabajo semántico que le permita efectuar la fusión de información geográfica de manera semiautomática, su aporte más importantes es la generación de una medida de similitud asimétrica entre nombres de entidades a nivel de cadenas de caracteres y la generación de una medida de similaridad basada en las relaciones entre superposiciones de las instancias espaciales[3].

Navarrete y Blat [13] realizan una mezcla de los conjuntos de datos basándose en la distribución espacial de las instancias, plantean que es posible medir un nivel de solapamiento espacial usando un umbral. Un alto solapamiento entre dos instancias u objetos geográficos indicaría que probablemente se refieran a un mismo concepto (relación semántica de equivalencia), Si la extensión espacial del primer valor está contenida en la extensión espacial del segundo valor, probablemente indique la existencia de una relación de contenencia entre las clases (relación semántica parte - todo), es decir, que la clase correspondiente al primer conjunto de datos es un subconjunto de la clase del segundo conjunto de datos comparado.

Kieler [14] en su documento expresa que no es suficiente una integración por aproximación lingüística ya que una interpretación en este nivel no arroja un resultado óptimo debido a la variedad de representaciones en las etiquetas, describe que incluso así se encuentren semejanzas en los términos pueden dar una definición semántica distinta dependiendo del contexto que se trabaje. Por ello, en su trabajo quiere omitir estas comparaciones de etiquetas y plantea identificar relaciones semánticas sólo en función de las características geométricas de las instancias geográficas, parte de la hipótesis que las descripciones de objetos pertenecientes a un mismo fenómeno tienen objetos en una misma posición geográfica o tienen propiedades geométricas similares. Su estudio lo divide en dos escenarios, en el primero muestra como se pueden obtener las relaciones semánticas a partir de la superposición geométrica; sin embargo, resalta cómo al comparar dos conjuntos de datos aunque las entidades representen el mismo concepto puede que no compartan las mismas instancias y por lo tanto, el criterio de superposición no puede ser aplicado. Por lo cual en el segundo escenario se plantea que la obtención de las relaciones semánticas pueden ser inferidas a partir de un patrón geométrico calculado por métodos de minería de datos, en un primer paso construye un modelo para describir el conjunto de clases de datos o conceptos basados en datos de entrenamiento aplicados en un primer conjunto y en la segunda etapa utiliza este modelo, para predecir la clase en la cual se clasificarían los nuevos elementos de datos del segundo conjunto. En la demostración mide la elongación y la rectangularidad de los ríos (elemento de geometría lineal) y lagos (elemento de geometría poligonal) de un grupo de datos y genera árboles de decisión para dichas variables con el algoritmo J48.

Recientemente Vaccaril, Shvaiko y Marchese [15] publicaron un artículo en donde muestran el problema de la heterogeneidad semántica como uno de los principales desafíos que enfrentan las IDE, plantean un contexto de integración de geoservicios basado en una coordinación semántica, implementado bajo el lenguaje LCC (Lightweight Coordination Calculus)[4]. En concreto, la integración de "web services" es realizada a partir de la determinación de concordancias semánticas entre las descripciones de los propios geoservicios. Para cumplir con este objetivo introducen una solución que denominan "structure preserving semantic matching".

3. Basada en la propuesta de Duckham y Worboys 2005.
4. Lenguaje utilizado para describir las interacciones entre los procesos distribuidos, como los agentes y los servicios web.

Escenarios de integración de datos geográficos a nivel semántico

La integración semántica de información geográfica puede realizarse de distintas formas dependiendo de las estructuras de entrada y los objetivos de la comparación que se pretenda realizar. A continuación se presentan 4 escenarios de integración desde una perspectiva de alineamiento de estructuras conceptuales, detectados a partir de la investigación realizada.

Escenario 1: Este escenario es el más común dentro de la integración semántica de datos en la web, plantea la entrada de dos ontologías a un algoritmo de alineamiento y mapeo cuyo producto es una estructura conceptual compartida, a partir de la cual se crea una ontología resultante de un mayor nivel de conceptualización ver figura 2, que comunica las dos ontologías de entrada.

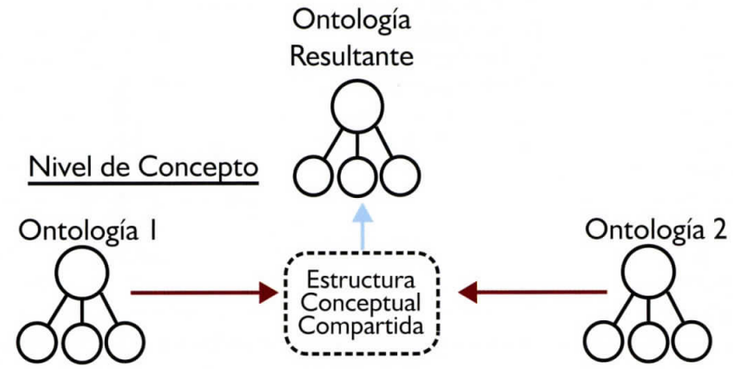


Figura 2. Esquematización del escenario1.
Fuente: Elaboración propia.

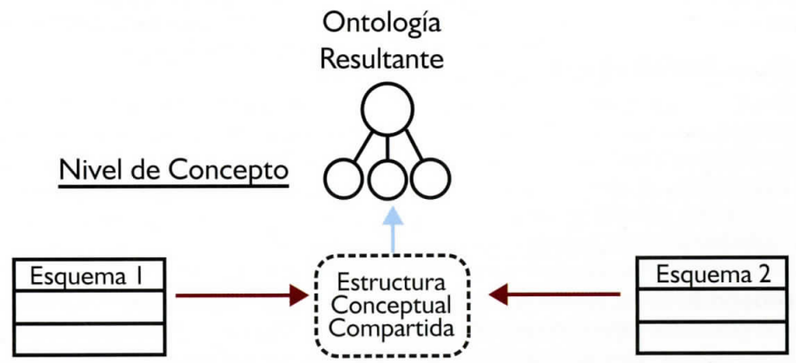
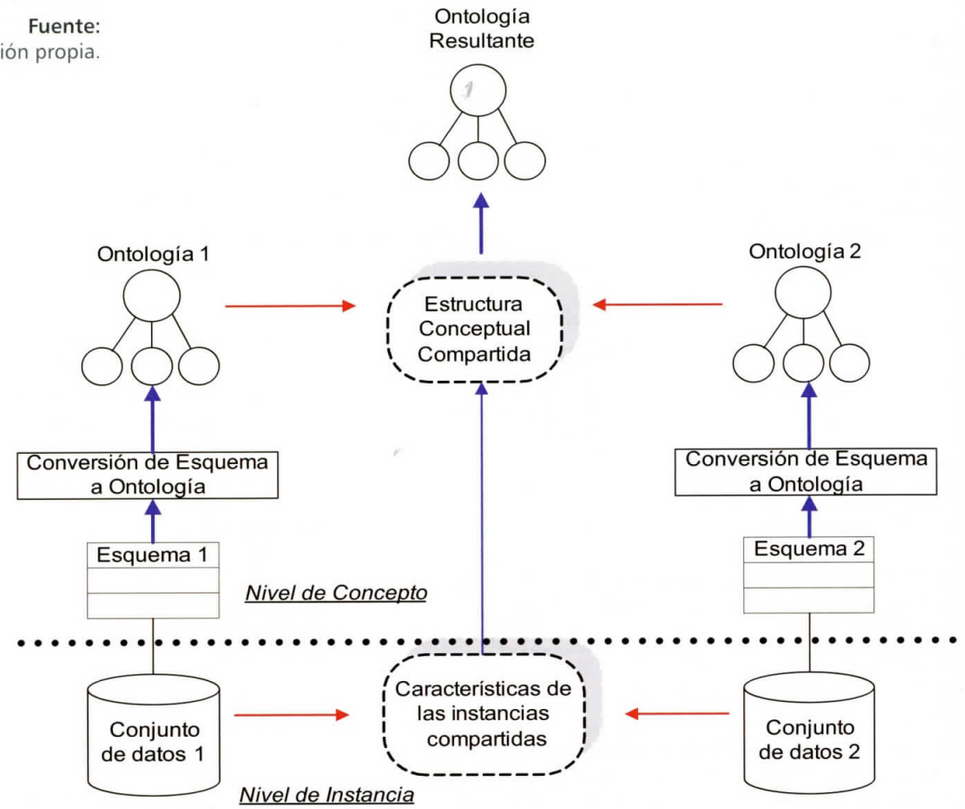


Figura 3. Esquematización del escenario1.
Fuente: Elaboración propia.

Escenario 2: Este escenario muestra el mismo esquema que el anterior, la diferencia radica en los datos de entrada, ya que en cambio de utilizarse ontologías se cuentan con esquemas de bases de datos, esto hace que la implementación del algoritmo de alineamiento varié en función de las características del modelo (ya sea un modelo Objeto Relacional o Entidad Relación) que represente el esquema, al finalizar el proceso se obtiene una ontología que comunica los dos esquemas de entrada y por lo tanto las dos bases de datos, este resultado puede ser de gran utilidad a la hora de resolver conflictos semánticos cuando se construyen bases de datos federadas.

Figura 4.
Diagrama del escenario de integración 3.

Fuente:
Elaboración propia.



Escenario 3: En los escenarios anteriores se planteaba una comparación semántica sólo a nivel de esquemas, sin embargo, como hemos visto también es posible usar las instancias para obtener una estructura conceptual compartida. En este escenario se muestra cómo obteniendo la comparación de las características de las instancias en un conjunto de datos geográficos estructurados (en concreto una BDE) es posible llegar a deducir correspondencias entre conceptos. Además, muestra que es necesario generar una ontología a partir de cada esquema[5] que facilite la generación de la ontología de mayor nivel y por lo tanto la comunicación de las dos bases de datos, al final el algoritmo de alineamiento generará las correspondencias a partir de las similaridades de instancia y de esquema.

5 Este proceso es muy común en el escenario de la web semántica e integración de Sistemas de información ya que generalmente se dé esquemas de bases de datos en la contrucción de ontologías.

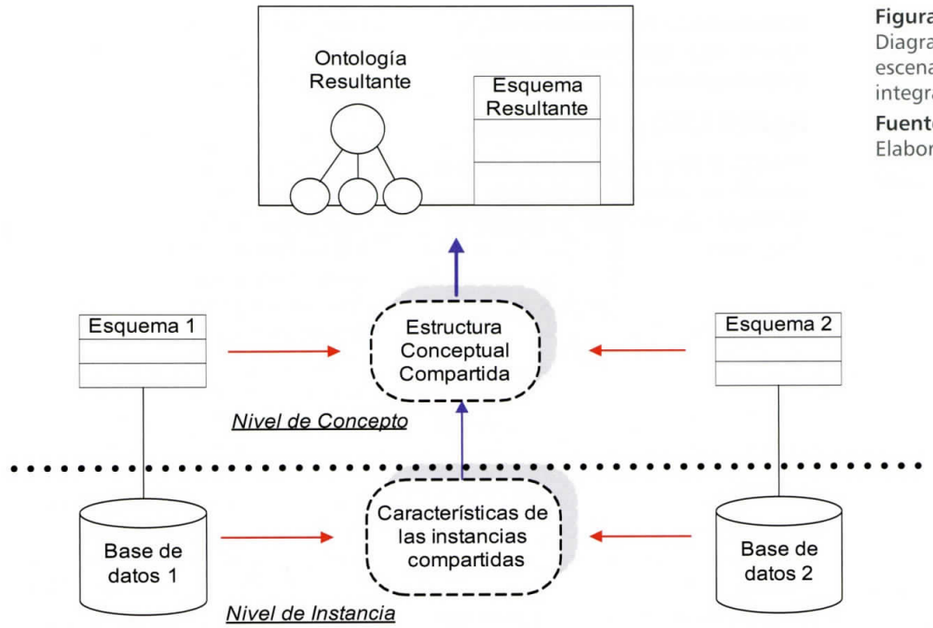


Figura 5.
Diagrama del escenario de integración 4.

Fuente:
Elaboración propia.

Escenario 4: Este ese escenario conserva muchas características del escenario 3, sin embargo, en él se sugiere no generar una ontología por cada esquema de entrada de la integración, sino que la estructura conceptual compartida es obtenida a partir de la propia semántica derivada de los esquemas, ver figura 5. Como resultado de este escenario se puede generar una ontología o un nuevo esquema resultante dependiendo de los objetivos que se planteen.

Modelo de comparación semántica propuesto

A continuación mostramos un marco de comparación semántica basado en el escenario de integración 4. Las entradas en el modelo serán cada una de las entidades de dos bases de datos espaciales y se espera obtener un listado con entidades concordantes y los valores de similaridad calculados entre cada comparación. Con este resultado se facilitará la generación de una ontología compartida que comunique las dos bases de datos, un nuevo esquema de bases de datos si se esta realizando un proceso de reingeniería o un marco de migración automatizado.

Criterios para el cálculo de similaridades

- ***Similaridad entre los nombres de la entidad:*** Este criterio parte de la lógica que si dos elementos comparten el mismo nombre representarán el mismo concepto. Sin embargo, es posible que en esquemas conceptuales modelados con cierta independencia los nombres no sean exactos sino que se compartan ciertas partes del nombre de la cadena, o que aunque representen el mismo concepto tengan nombres sinónimos. Para el cálculo de esta similaridad primero se realizará un método basado en cadenas usando la distancia de Levenshtein como métrica (también llamada distancia de edición), su cálculo establece el número mínimo de operaciones necesarias para transformar una cadena en otra su cálculo se puede ver en [16] para efectos prácticos denotare-

mos a esta métrica como *D(S1,S2)*. Como esta distancia es simétrica y se encuentra acotada por:

$$0 \leq D(S1,S2) \leq \max(|S1|,|S2|)$$

El cálculo de similaridad del nombre usando el método de cadenas con la métrica Levenshtein se puede definir como:

$$SimNombre = 1 - \frac{2*D(S1,S2)}{|S1| + |S2|}$$

Si usando este método no se puede concluir que dichas entidades son concordantes, se usará un método lingüístico conectándose a una base de datos léxica, buscando una relación de sinonimia entre los nombres comparados.

- ***Similaridad entre atributos de la entidad:*** Este criterio se basa en la idea que entidades asociadas a un mismo concepto comparten atributos que identifican las propiedades que a su vez definen la naturaleza del elemento. La comparación se realiza por el nombre o etiqueta del atributo. La similaridad en este caso es calculada en función del número de atributos que comparten las entidades comparadas. Esta similitud será el resultado de la relación entre la suma de las concordancias de atributos y el número total de atributos encontrados en ambas entidades. Esta similaridad será denotada como *SimAt*.

- ***Similaridad entre Relaciones:*** Esta similaridad se calculará como la proporción entre el número de relaciones comunes entre las entidades y el número total de relaciones encontradas en ambas entidades. Para este caso esta similaridad se le designará como: *SimRel*

- ***Similaridad entre las relaciones topológicas:*** Otra forma de comparar entidades geográficas en esquemas es ver sus relaciones topológicas tanto horizontales como verticales[6], analizando cada una de las propiedades que las define y las entidades presentes en la relación. De la misma forma que la similaridad anterior calculando, la proporción entre el número de relaciones topológicas comunes entre las entidades y la unión de relaciones topológicas encontradas en ambas entidades. Para el presente documento esta similaridad se denotará como: *SimTop*. Se debe tener en cuenta que esta relación solo se puede calcular si la relación topológica se encuentra explicita dentro del modelo.

- ***Similaridades entre el componente geométrico de la entidad geográfica:*** Las similaridades nombradas anteriormente son calculadas en un nivel de esquema, sin embargo como se ha recalcado, es posible realizar una comparación a nivel de instancias para deducir una estructura conceptual compartida. En este caso, se aprovechará el componente geométrico de la entidad para calcular la similaridad. Existen distintos criterios a nivel geométrico para realizar estas comparaciones algunos ejemplos son: **Posición de los nodos, Envolvente, Longitud, Sinuosidad, Rectangularidad, Valor Área, Alargamiento:** Aplicado a geometrías línea y polígono. El cálculo de esta similaridad esta en función de la proporción entre el número de instancias detectadas como geométricamente coincidentes y el número menor de instancias de los conjuntos de datos comparados. Esta similaridad será denotada como: *SimGeo*.

- ***Similaridades entre el componente alfanumérico de la entidad geográfica:*** Este criterio propone detectar los registros comunes en las bases de datos comparados, para ello es necesario utilizar métodos de alineamiento terminológicos (ya sean basados en cadenas de caracteres o métodos lingüísticos). El grado de similaridad se calculará como la relación entre el número de registros detectados como semejantes y el número menor de registros en las entidades comparadas. Se denominara *SimAlf*

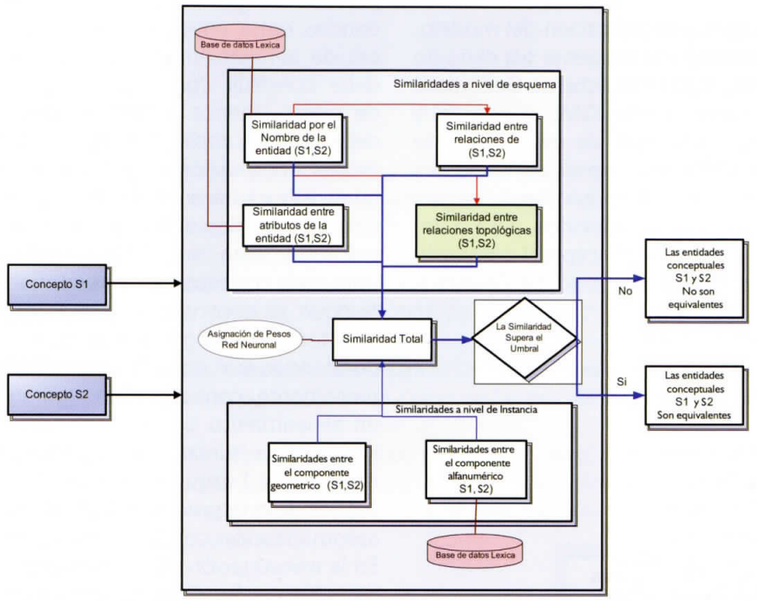


Figura 6.
Diagrama del modelo de integración propuesto.

Fuente:
Elaboración propia.

Estructura del modelo de integración semántica

Una vez se ingresan las entidades de entrada, el modelo calcula similaridades a nivel de instancia y esquema según los criterios definidos con anterioridad. Después de esto se propone una combinación de estas similaridades usando una función que calcule la similaridad total, dicha función vendría determinada por el peso de cada similaridad dentro del modelo y el valor de Similaridad previamente calculado[6]. En la figura 6 se muestra el diagrama del modelo. La función quedaría dada por la siguiente expresión:

$$SimTotal = Wn * SimNom(S1,S2) + Wa * SimAt(S1,S2) + Wr * Sim\,Rel(S1,S2) + Wt * SimTop(S1,S2) + Wg * SimGeo(S1,S2) + Wf * SimAlf(S1,S2)$$

Donde:

$$W\sum = 1$$

Al final la función de similaridad total tendrá valores en un rango de 0 a 1. En donde 0 definirá que no existe similaridad y 1 define que las entidades son iguales. Definiendo un umbral de aceptación podremos definir que los dos términos son equivalentes o no.

Si $SimTotal(S1,S2) \geq \lambda$, Las entidades son equivalentes

6 En este documento topología Horizontal hará referencia a topologías dentro de una misma entidad geográfica y topología vertical describirá las relaciones topológicas entre dos entidades geográficas distintas.

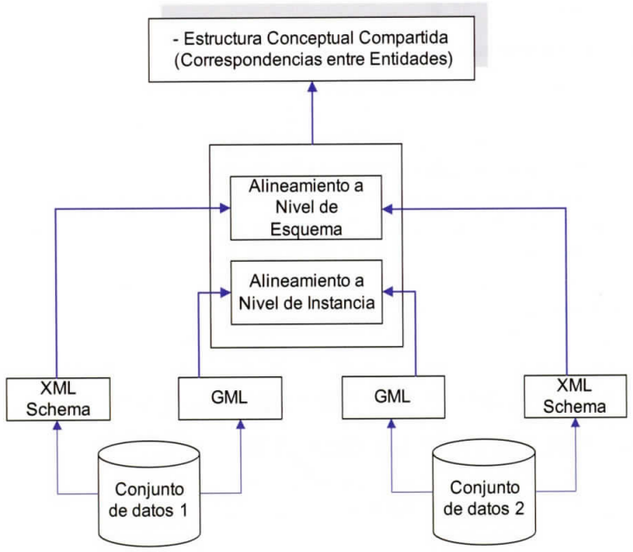
7 Basado en el algoritmo G-Match.

Si *SimTotal(S1,S2)* $<\lambda$, Las entidades no son equivalentes

Marco de implementación

En un marco de aplicación del modelo, se plantea que el esquema sea extraído como un XSD (Xml Schema Definition) y los datos como GML (Geography Markup Language); de esta forma, se facilitará el cálculo de similaridades en el nivel de esquema e instancia según los criterios ya establecidos, para obtener la estructura conceptual compartida. Ver figura 7.

Figura 7.
Marco de implementación.



Fuente:
Elaboración propia.

Conclusiones y trabajo futuro

Como se vio en el documento la generación de un modelo que automatice el proceso de generación de correspondencias entre entidades usando técnicas de alineamiento de ontologías se debe construir con una combinación de varios criterios y métodos, esto se debe a la complejidad que conlleva extraer información semántica de cualquier estructura conceptual, que hace que los criterios aplicados de manera individual sean muy débiles para extraer unas correspondencias confiables. Aunque se encontraron formas de semiautomatizar la generación de correspondencias aún no se tiene un método plenamente confiable que garantice un alineamiento sin la intervención de la experticia humana, por lo cual hace necesario la búsqueda de nuevas alternativas de este proceso que permitan la automatización completa del proceso. En la investigación realizada además se detectó que las diferencias en la temporalidad y granularidad de la información pueden dificultar la aplicación de los alineamientos por lo cual deben incluirse mecanismos que permitan enfrentar esta problemática. Debe resaltarse es como muchos investigadores del área han enfocado sus trabajos al uso de mecanismos exencionales de alineamiento, esto se debe a que han encontrado en las instancias (sobre todo geográfica) una variedad amplia de alternativas para determinar estructuras conceptuales compartidas. Otro punto que se comenta es el problema de usar diccionarios léxicos dentro de la aplicación de métodos terminológicos, esto se debe a la dificultad de encontrar dichos diccionarios aplicados a un contexto particular de trabajo.

En el documento se planteó el marco básico del modelo propuesto y se esquematizó su forma de implementación de manera general el trabajo a corto plazo se orientará a experimentar con datos reales que retroalimenten el modelo. Además, se pretende adelantar la

integración al modelo de un razonador mediante el uso de lógica difusa de los mecanismos de decisión del modelo. También se pretende profundizar en la investigación en el uso de nuevos mecanismos de comparaciones a nivel de instancias geométricas que aumenten la confiabilidad en los resultados de las correspondencias semánticas.

Bibliografía

[1] George,D . "Understanding Structural and Semantic Heterogeneity in the Context of Database Schema Integration". Department of Computing, University of Central Lancashire, Preston UK. 2005.

[2] Nudelman, G. H., Iochpe, C., Ferrara, A., and Castano, S., "Towards Effective Geographic Ontology Matching," en GeoS 2007, pp. 51-65. 2007.

[3] Gruber, T. R., "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," International Journal of Human-Computer Studies, vol. 43, pp. 907-928, 1995.

[4] Euzenat, J., Bach, T. L., Barrasa, J., Bouquet, P., Bo, J. D., Dieng, R., Ehrig, M., Hauswirth, M., Jarrar, M., Lara, R., Maynard, D., Napoli, A., Stamou, G., Stuckenschmidt, H., Shvaiko, P., Tessaris, S., Acker, S. V., and Zaihrayeu, I., "State of the Art on Ontology Alignment,", 2004.

[5] Giunchiglia, F., Shvaiko, P., and Yatskevich, M., "S-Match: an algorithm and an implementation of semantic matching," publicado en Proceedings of ESWS 2004, Heraklion (GR), pp. 61-75, 2004.

[6] Rodríguez, M. A., Egenhofer, M. J., and Rugg, R. D., "Assessing Semantic Similarities among geospatial Feature Class Definitions," in Interoperating Geographic Information Systems, Second International Conference, Interop '99, Zurich, Switzerland, pp. 189-202, 1999.

[7] Volz, S, "Data-Driven Matching of Geospatial Schemas" en Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp 115-132, 2005.

[8] Schwering, A. and Raubal, M., "Measuring Semantic Similarity Between Geospatial Conceptual Regions," in GeoS 2005, pp. 90-106, 2005.

[9] Duckham, M. and Worboys, M., "An Algebraic Approach to Automated Geospatial Information Fusion," Interantional Journal of Geographic Information Science, vol. 19, pp. 537-557, 2005.

[10] Hess, G. N., Iochpe, C., and Castano, S., "An Algorithm and Implementation for GeoOntologies

Integration," in VIII Brazilian Symposium on GeoInformatics, Campos do Jordão, Brazil, pp. 129-140, 2006.

[11] Nudelman G.H., Iochpe C., "Ontology-driven resolution of semantic heterogeneities in GDB conceptual schemas" Universidade Federal do Rio Grande do Sul. Instituto de Informática, 2006.

[12] Navarrete T. "Semantic integration of thematic geographic information in a multimedia context", PhD Thesis, Doctorate in Computer Science and Communication Department of Technology, Universitat Pompeu Fabra, 2006.

[13] Navarrete, T. and Blat, J., "An algorithm for Merging Geographic Datasets Based on the Spatial Distribution of Their Values," en GeoS 2007, pp. 66-81, 2007.

[14] Kieler, B, "Derivation of Semantic Relationships between different Ontologies with the Help of Geometry" en "Semantic Web meets Geopatial Applications", Germany. 2008.

[15] L, Vaccari1,P. Shvaiko y M. Marchese, "A geo-service semantic integration in Spatial Data Infrastructures" in International Journal of Spatial DataInfrastructures Research, Vol. 4, 24-51., 2009.

[16] G. Navarro, "A guided tour to approximate string matching," ACM Computing Surveys (CSUR), vol. 33, no. 1, pp. 31–88, March 2001.