Revista Internacional de Ciencia y Tecnología de la Información Geográfica
International Review of Geographical Information Science and Technology

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

APROXIMACIÓN GEOSEMÁNTICA PARA DETECTAR INCONSISTENCIA EN LOS METADATOS DE SERVICIOS WEB GEOESPACIALES

WALTER RENTERIA AGUALIMPIA[1], FRANCISCO J. LOPEZ-PELLICER[1], JAVIER LACASTA[1], PEDRO R. MURO-MEDRANO[1], F. JAVIER ZARAZAGA-SORIA[1]
[1] Dpto. de Informática e Ing. de Sistemas Universidad de Zaragoza
C/ María de Luna 1, 50.018 Zaragoza, España
{walterra, fjlopez, jlacasta, prmuro, javy}@unizar.es

RESUMEN

La validación de metadatos asociados a servicios *web* geoespaciales es uno de los problemas más relevantes relacionados con la preservación de la información en catálogos de Servicios *Web* Geoespaciales. Este trabajo presenta un nuevo método automático para detectar inconsistencias en los registros de metadatos que asegura la coherencia de la descripción espacial contenida en los registros de metadatos. Este método está basado en la combinación de técnicas de agrupación espacial y métodos estadísticos de análisis geográfico. Los resultados permiten mostrar la necesidad de sistemas que detecten inconsistencias con el fin de validar la integridad semántica de los metadatos de recursos geoespaciales.

Palabras clave: validación, metadatos, catálogos geoespaciales, servicios geoespaciales, agrupaciones espaciales, geocodificación.

GEOSEMANTIC APPROACH TO DETECT INCONSISTENCY IN GEOSPATIAL WEB SERVICES METADATA

ABSTRACT

The validation of the metadata associated with Geospatial Web Services is one of the main problems related to the preservation of the information in the Geospatial Web Service Catalogs. This work describes a new automatic method for detecting outliers in metadata registers in order to ensure coherence of the spatial description contained in metadata. The method is based on the combination of spatial clustering techniques and geographical statistical analysis. The results allow

Recibido: 26/10/2012
Aceptada versión definitiva: 2/4/2013

© Los autores

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

showing the need for systems to detect inconsistencies in order to validate the semantic integrity of geospatial metadata resources.

Keywords: validation, Geospatial catalogues, metadata, spatial clustering, geographic statistical analysis, geocoding

1. Introducción

En el contexto de la recuperación de información geográfica (*Geographic Information Retrieval* o GIR) relacionada con Servicios *Web* Geoespaciales es de suma importancia validar los metadatos que describen dichos servicios. La validación permite, por ejemplo, garantizar la veracidad, credibilidad y utilidad de los resultados de las consultas a catálogos que contengan dichos metadatos, y los análisis que se lleven a cabo sobre dichos resultados. Los enfoques tradicionales de validación se basan principalmente en verificar la conformidad de los metadatos con un determinado estándar. Esta conformidad no es suficiente para sistemas de recuperación de información avanzados que deben su eficacia a la veracidad, exactitud e integridad de los metadatos (Hartmann y Stuckenschmidt, 2002). Por tanto, también es necesario comprobar que su contenido sea coherente. En este artículo analizamos y proponemos una solución para el siguiente problema:

> *¿Sigue siendo válida la descripción de un servicio si la extensión descrita en el contenido estructurado (por ejemplo, en las cajas de las capas de un servicio de mapas) no concuerda con el ámbito descrito en el contenido no estructurado (por ejemplo, en el título y en la descripción del servicio)?*

En términos generales esta pregunta requiere un tipo diferente de validación que verifique la consistencia interna entre las propiedades de cada registro de metadatos. Las propiedades con contenido espacial son susceptibles de ser validadas en busca de posibles inconsistencias usando diferentes puntos de vista: *sintáctico*, *geométrico*, *topológico* y *semántico* (Yu-hong y Feng-yuan, 2008). Aunque analizamos algunas de estas propiedades, en este trabajo nos centramos en la validación de la consistencia semántica, que consiste en usar modelos o Sistemas de Organización del Conocimiento para comparar las referencias espaciales indirectas del contenido no estructurado y las referencias espaciales directas del contenido estructurado. La semántica es vista aquí como el empleo de modelos de representación del conocimiento humano, junto con modelos estadísticos para el uso de palabras que permiten comparar las representaciones del conocimiento entre piezas de información textual. Este tipo de enfoque se conoce en la literatura como *semántica latente de contenidos no estructurados* (Jones *et al.*, 2008).

Las incoherencias en las descripciones de los recursos geográficos sobre los cuales se aplican las consultas espaciales pueden ocasionar desde resultados discrepantes o mal ponderados, hasta la omisión de resultados que podrían satisfacer la consulta. Si los metadatos contienen alguna inconsistencia en la descripción espacial se crean distorsiones que entorpecen la visualización, localización e interpretación de los recursos geográficos. Por ejemplo, un servicio cuya extensión corresponde a España pero su descripción textual se refiere a Polonia generará resultados de

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

búsqueda incoherentes. Un caso práctico de este tipo de problema se presenta en la figura 1. Algunos de estos problemas de inconsistencias de tipo espacial en la descripción de recursos geográficos son recogidos por Monmonier (1991) y Hill (2006, 156-161). Por estas razones, los metadatos, además de describir o documentar qué información proporcionan los recursos geoespaciales, como se menciona en Zarazaga-Soria *et al.* (2003), deben contener descripciones coherentes desde varias perspectivas, entre ellas la textual y la espacial.

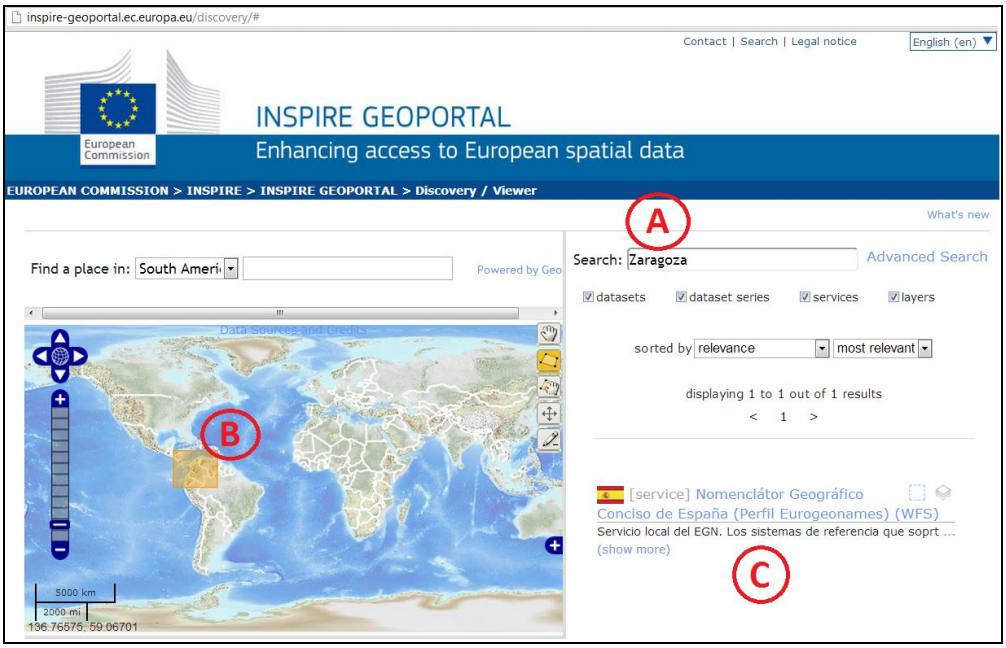

Figura 1. Diferentes métodos de búsqueda espacial (que podrían englobarse en sistemas del tipo *GeoNetwork*, uno de los más populares catálogos de recursos geográficos).
Al realizar la consulta **(A)**: Zaragoza en la zona geográfica de Colombia **(B)**, la respuesta es un servicio sobre nombres geográficos de España **(C)**. Ante resultados como estos los usuarios pueden experimentar sensaciones de frustración.

Un nuevo concepto surge aquí: *visibilidad de un registro de metadatos*. En el contexto de recuperación de información espacial, un registro de metadatos será *invisible* cuando se realiza una búsqueda espacial por topónimos y el registro no contiene el topónimo que se esperaría que tenga dada su ubicación o extensión espacial. Esto es, si se quisiera recuperar un servicio que está en Barcelona pero su descripción apunta a Madrid, su inconsistencia le impedirá ser "visto" por el sistema de recuperación. Un recurso geográfico será visible cuando, ante una consulta espacial, tenga consistencia entre sus diferentes propiedades espaciales y pueda ser hallado al preguntar indistintamente por cada una de sus propiedades espaciales o por todas a la vez.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

Para abordar de manera automática el problema de detección de inconsistencias, el cual conlleva a potenciar la invisibilidad espacial de registros de metadatos de Servicios *Web* Geoespaciales, proponemos el uso combinado de técnicas de *agrupación espacial* (o *spatial clustering*) y *métodos estadísticos de análisis geográfico*. El caso de estudio presentado se centra en registros de metadatos de Servicios *Web* Geoespaciales de España. Para su análisis se ha utilizado como sistema de organización del conocimiento la ontología para la representación de dominios jurisdiccionales, descrita en López-Pellicer *et al.* (2012a), que formaliza la información administrativa de España, complementado con el modelo "*Global Administrative Areas*" (Hijmans *et al.*, 2012).

El análisis de los clústeres, además de revelar una relación de proximidad espacial, podría revelar el consenso compartido por los expertos que han documentado las referencias textuales de los recursos geográficos. Bajo esta hipótesis, un recurso con descripción espacial incoherente con el consenso puede ser visto como una inconsistencia y, consecuentemente, con potenciales problemas de visibilidad, o como un recurso con necesidad de revisión.

Para cumplir con los objetivos propuestos, el artículo se estructura en cinco apartados. Tras esta introducción, se aborda el estado de la cuestión. Luego se presentan la metodología utilizada, basada en técnicas de cómputo de clústeres espaciales y análisis estadístico geográfico. En el cuarto apartado se muestra la implementación de la metodología. En el quinto se describen y analizan los resultados obtenidos, seguido de una valoración y discusión de los mismos. Finalmente, se enuncian las conclusiones y posibles líneas de trabajo futuro.

2. Estado de la cuestión

En la literatura se pueden encontrar por lo menos cuatro aproximaciones de validación de metadatos desde el punto de vista espacial: *sintáctica, geométrica, topológica y semántica* (Yu-hong y Feng-yuan, 2008), dentro de los cuales, los más cercanos a este trabajo son el aspecto sintáctico y el semántico, y serán mirados con más detalle en las secciones siguientes.

La validación sintáctica consiste en verificar que no halla inconsistencias o errores en los caracteres, datos o formatos en los metadatos. La detección de este tipo de inconsistencia es sencilla y está bastante desarrollada. Algunos ejemplos relevantes tienen que ver con enfoques que realizan trabajos de validación de registros de metadatos de acuerdo a estándares internacionales ISO (ISO, 2003; ISO, 2007) o a las guías técnicas de la directiva de Infraestructura de Información Espacial de la Unión Europea (INSPIRE); algunos de los aspectos validados por estos enfoques son, por ejemplo: esquemas, nombres, tipos, cardinalidad, estructuras. Varios trabajos abordan este enfoque de validación (Kliment *et al.*, 2012; Ford *et al.*, 2011; Capdevila *et al.*, 2012).

La validación semántica puede ser abordada desde varias perspectivas (Svilia *et al.*, 2007; Bruce y Hillmann, 2004), las cuales pueden ir desde enfoques semánticos simples hasta complejos sistemas de reglas semánticas lógicas. Un primer enfoque es verificar si es posible establecer un

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

mapa de equivalencias entre la terminología empleada en un registro de metadatos y los conceptos de una o varias ontologías. Un trabajo relevante en este sentido es abordado en Raatikka y Hyönen (2002). Una segunda aproximación verifica la consistencia interna del contenido del metadato. Por ejemplo, Yue *et al.* (2010) realizan un procedimiento de validación de metadatos sobre Servicios *Web* Geoespaciales; su trabajo establece una comparación entre los metadatos de entrada y metadatos generados de acuerdo a condiciones descritas mediante una ontología. Una tercera categoría es la presentada por Hartmann y Stuckenschmidt (2002); su enfoque analiza metadatos de sistemas de información basados en *web*, en particular, analizan portales *web* partiendo de los metadatos extraídos de sus páginas de contenidos.

Nuestra aproximación de validación presenta algunas similitudes con los trabajos de Hartmann y Stuckenschmidt (2002) y Yue *et al.* (2010), ya que nuestro enfoque comprueba si el contenido se ajusta a un modelo de organización del conocimiento dado. Las diferencias principales radican en el método y el objeto de estudio. Por ejemplo, el trabajo de Hartmann y Stuckenschmidt (2002) se basa en reglas obtenidas por medio de aprendizaje de máquinas (*machine learning*) y el objeto de estudio son los metadatos de portales *web*. Nuestro enfoque se centra en analizar y detectar inconsistencias en registros de metadatos de Servicios *Web* Geoespaciales por medio de un nuevo enfoque que combina técnicas de inteligencia artificial de agrupamiento y *métodos estadísticos de análisis geográfico*. En el trabajo de Yue *et al.* (2010) la validación es un proceso más complejo, en el cual se generan metadatos que luego son comparados con precondiciones dadas a través de sentencias basadas en modelos de conocimiento como las ontologías, por ejemplo: un sistema de referencia de coordenadas específico o un formato de archivo particular. Nuestro trabajo también hace uso de modelos de organización del conocimiento, pero para asegurar que las referencias espaciales directas son coherentes con las referencias espaciales indirectas.

3. Metodología

En esta sección abordaremos el agrupamiento de metadatos de Servicios *Web* Geoespaciales, partiendo de la hipótesis de que dichas agrupaciones revelarán algún consenso espacial sobre las descripciones de expertos en documentación que han descrito los registros de metadatos de cada grupo o clúster. Los clústeres de registros cuyas referencias espaciales directas difieren del consenso serán señalados como inconsistentes. Para determinar el consenso utilizaremos técnicas de agrupación o *clustering*. Esta técnica es propia del área de minería de datos e inteligencia artificial y permite formar agrupaciones de elementos en función de algún criterio o medida de similitud (Sugihara *et al.*, 2011).

Existen al menos dos formas en las que se pueden agrupar espacialmente los registros de metadatos de Servicios *Web* Geoespaciales usando la información explícita en su descripción:

- Agruparlos aplicando alguna medida de distancia según la posición geográfica recogida en el metadato (referencia geográfica directa).
- Agruparlos de acuerdo a los topónimos utilizados en el metadato asociados a su ubicación (referencias geográficas indirectas).

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

Para poder detectar inconsistencias en los metadatos asumimos que detrás de una agrupación o clúster geográfico subyace una propiedad, característica o fenómeno que podría estar directamente asociado con el motivo de la agrupación. Desde el punto de vista espacial textual dicha propiedad debería ser el topónimo del lugar donde está el clúster. En la línea de este planteamiento, algunas de las preguntas que surgen son las siguientes:

- ¿Qué porcentaje de elementos del clúster hacen referencia textual a la ubicación donde se encuentra el clúster?; ¿es significativo dicho porcentaje?.
- Al analizar espacialmente las divisiones creadas por los clústeres, ¿qué propiedad es resaltada? o ¿se podría informar cuál es el motivo de la agrupación?.
- ¿Bajo qué propiedad toma sentido o "es más significativo" el análisis del grupo?.

Con miras a resolver estas preguntas, se plantea la metodología de trabajo descrita en la figura 2. Esta metodología tiene las siguientes etapas:

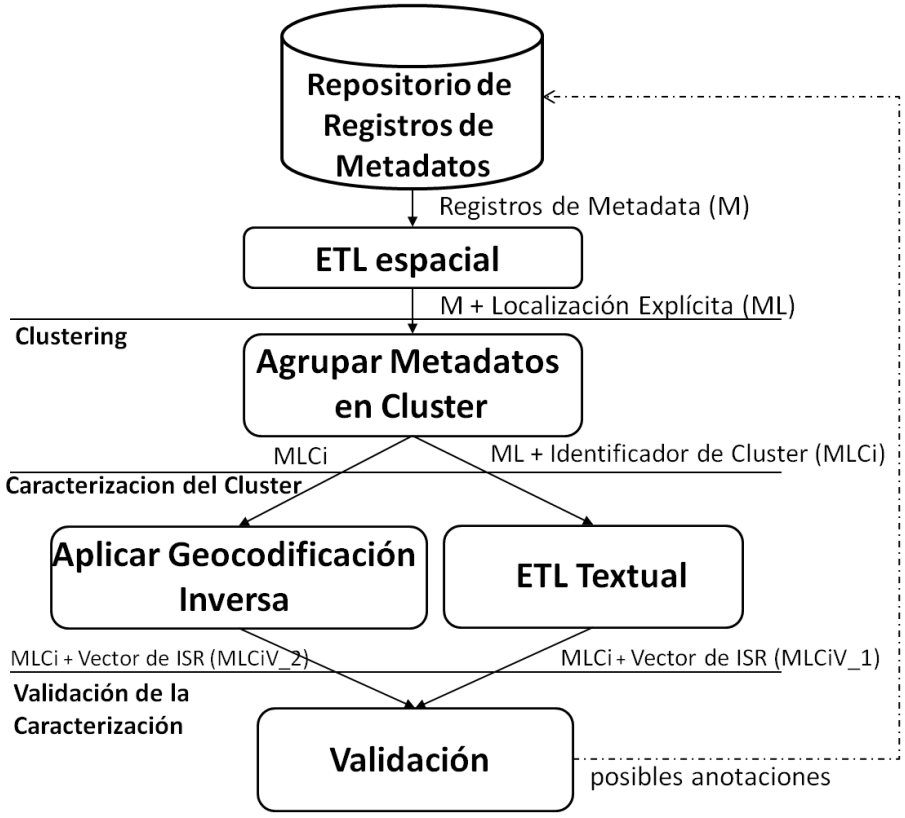

Figura 2. Método de detección de inconsistencias basado en clústeres.

© Los autores

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

- **Pre-procesamiento de los metadatos:** compuesta por los módulos de Repositorio de Metadatos y ETL espacial. Para crear el repositorio de registros de metadatos se ha utilizado el *crawler* descrito en López-Pellicer *et al.* (2012b). Partiendo del Repositorio de Registros de Metadatos (M) de Servicios *Web* Geospaciales, se pre-procesa la información para acondicionarla en el sub-módulo de Extracción, Transformación y Carga (ETL). Esto permite homogeneizar las coordenadas de los registros a un mismo sistema de coordenadas, obteniendo registros de metadatos con su localización explícita (identificado como ML en la figura 2).
- **Agrupación espacial de metadatos**. A partir de la localización explícita y homogeneizada de los metadatos se pueden generar los grupos o clúster, obteniendo como salida metadatos anotados con la etiqueta del clúster al que pertenece.
- **Geocodificación inversa**. A partir de la localización de cada clúster se aplica un sub-proceso que asocia a cada clúster nombres de lugar (*topónimos*). El *topónimo,* o posiblemente conjunto de *topónimos,* pasan a ser los identificadores que caracterizan el clúster. Este proceso genera un vector de topónimos implícito (identificado como MLCiV_2 en la figura 2).
- **Extracción de información espacial textual**. Otra hipótesis se centra en que la co-ocurrencia espacial de registros de metadatos no se da por casualidad, sino que estará reflejada en la información textual o descriptiva de carácter espacial. En otras palabras, los registros de metadatos que concurren en un mismo lugar deberían tener topónimos en común, los cuales serán extraídos de las descripciones que aparecen en el cuerpo del metadato del servicio, para generar un vector de topónimos explícitos (identificado como MLCiV_1 en la figura 2).
- **Validación**. Esta fase está dirigida a verificar los siguientes aspectos:
    - La correspondencia entre los topónimos presentes en los registros de metadatos y los topónimos de la extensión que cubre el clúster con base en un modelo de organización del conocimiento geográfico.
    - La significancia o relevancia de los topónimos respecto a la extensión del clúster.

El *aspecto* semántico de la validación que se propone hace referencia a la parte conceptual que está asociada a toda extensión geográfica, es decir, las referencias geográficas indirectas que describen una extensión territorial. Estas extensiones están concebidas por organizaciones humanas que compilan su estructura mediante el uso de sistemas de organización del conocimiento, tales como taxonomías y ontologías. Partiendo de esta concepción, un clúster geográfico es semánticamente válido si posee una correspondencia mínima con las referencias geográficas que aparecen en el sistema de organización del conocimiento usado como base para el análisis. La componente semántica es aportada por sistemas de organización del conocimiento de base geográfica, que sirven para validar si lo que entendemos y decimos de los nombres de un territorio es coherente con la descripción anotada en el metadato del recurso asociado al mismo territorio. Si existe coherencia, entonces la visibilidad del recurso estará garantizada a la hora de hacer

© Los autores

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

búsquedas espaciales por una u otra propiedad espacial. A partir de estos análisis se propondrán posibles correcciones o anotaciones de topónimos ausentes que enriquezcan los registros de metadatos, a través de un proceso de realimentación con el repositorio, señalado con la línea punteada en la figura 2.

4. **Implementación**

4.1. ETL espacial

Este proceso filtra y prepara los metadatos para poder aplicar el algoritmo de cómputo de clústeres espaciales sobre su localización. Las funciones principales de esta fase son:

- Extracción de las geometrías: identificación de coordenadas o caja rectangular (*Bounding Box*)
- Verificación del orden de entrada de las coordenadas: (Latitud, Longitud) o (Longitud, Latitud).
- Transformación y homogeneización de geometrías: convertir a un mismo sistema de coordenadas.
- Simplificación de las geometrías a centroides.

4.2. Agrupamiento de metadatos en clústeres

El proceso de detección de metadatos inconsistentes por medio del cómputo de clústeres espaciales utiliza el algoritmo DBSCAN (Ester *et al.*, 1996) para identificar agrupaciones de metadatos geográficos de acuerdo al criterio de densidad espacial. Entre sus principales ventajas están las siguientes:

- Es un método que no requiere supervisión humana para determinar los clústeres.
- Es capaz de detectar clústeres de recursos de formas geométricas diversas.
- Permite identificar recursos espacialmente anómalos (*outlier*).
- Se ha comprobado la eficiencia del DBSCAN para gestionar bases de datos espaciales de gran tamaño como es el caso de las Bibliotecas de Mapas Digitales actuales (Sander *et al.*, 1998).

En síntesis, este algoritmo genera agrupaciones de metadatos de Servicios *Web* Geoespaciales identificando las zonas donde hay mayor densidad o concentración según un valor de radio pre-establecido. Además, este algoritmo identifica como inconsistentes los metadatos aislados desde el punto de vista de densidad espacial.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

Una vez se han calculado los clústeres para un radio en particular, es necesario asegurar que los clústeres formados son los más óptimos, es decir, comprobar que la distancia entre los elementos del clúster y entre los clúster es la mejor posible. Esto se logra con índices de validación de algoritmos de cómputo de clústeres, tales como los índices *C-Index* y *Goodman-Kruskal* (Hubert y Schultz, 1976; Goodman y Kruskal, 1954). La filosofía detrás de este tipo de índices es determinar el radio que produce mejores clústeres según una función de distancia entre pares de elementos. Aunque inicialmente el índice *GK* se ha usado en las primeras pruebas, se ha preferido utilizar el *C-Index* en este trabajo por su velocidad de cómputo (Milligan, 1981). *C-Index* mide cuán buenas son las agrupaciones resultantes en relación a otros resultados con diferentes valores de un radio (*épsilon*) usado por el algoritmo de *clustering*. La expresión para este índice es:

$$C_{i,epsilon} = \frac{S - S_{\min}}{S_{\max} - S_{\min}}$$

Donde *S* es la suma de las distancias de todos los pares de elementos del mismo clúster para un *épsilon*.

De la expresión se deduce que *C* será pequeño si los pares de elementos con pequeñas distancias están en el mismo clúster. Por lo tanto, cuanto más cercano a 0 sea *C*, mejores son los clústeres resultantes porque la distancia entre los pares de elementos será pequeña. A partir de todos los $C_{i,epsilon}$ de cada cluster se obtiene el valor promedio, que sirve como indicador global del radio de clustering apropiado para el caso estudiado.

4.3. Aplicación de geocodificación inversa

En esta fase del proceso se requiere saber cuáles *topónimos* están asociados a cada clúster; para suplir esta necesidad se usa la geocodificación inversa. Esta fase consiste en asignar nombres de lugares (*topónimos)* a unas coordenadas geográficas conocidas (Turner, 2006; Florczyk *et al.*, 2009).

Dado que un *clúster* puede tener una forma geométrica compleja, en este punto surge la necesidad de desarrollar un nuevo geocodificador inverso que pueda asociar *topónimos* a geometrías más complejas que la de un punto o una caja rectangular. El requerimiento mencionado no es soportado por los geocodificadores estudiados (GeoNames, 2012; Google Geocoding API, 2012; Geocoder.ca, 2012). Por esta razón se implementó uno propio. Sus principales características son dos:

- Permite asociar topónimos a clústeres de metadatos de servicios con formas geométricas diversas o complejas.
- Además de los topónimos, proporciona el porcentaje de área del clúster asociada a la entidad del topónimo. Esta información puede ser usada como ponderación en sistemas de recuperación.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

Sumado a estos aspectos, en el geocodificador inverso desarrollado se puede especificar el nivel de división política en el que se desea establecer la operación de geocodificación (por ejemplo, país, estado o comunidad y provincia). La aclaración hecha surge dado que un geocodificador inverso podría retornar como respuesta para una misma coordenada o región, tanto el nombre de una calle como un barrio, una ciudad, un estado o país. El geocodificador inverso toma como dato de entrada la extensión geográfica del *clúster* (*la geometría*) y regresa los topónimos asociados a la ontología de dominios jurisdiccionales (López-Pellicer *et al.*, 2012a) o a las unidades administrativas proporcionadas por la base de datos de áreas administrativas globales GADM (Hijmans *et al.*, 2012). En resumen, el procedimiento de Geocodificación Inversa *GI* opera como describe la siguiente expresión:

$$F_{GI}(geometria, nivel) = \langle toponimo_1, toponimo_2, ..., toponimo_m \rangle, \langle w_1, w_2, ..., w_m \rangle$$

Donde *geometría* es la geometría del clúster de metadatos de servicios; *nivel* es un parámetro opcional que especifica el nivel de división política/administrativa en el que se desea realizar la geocodificación (si no se especifica, se regresan todos los topónimos de todos los niveles que intercepten con la zona que cubre la geometría del cluster); $toponimo_i$ es el vector de topónimos asociados a dicha geometría con sus respectivos valores de ponderación w_i, que corresponden al porcentaje de área interceptada entre la zona que representa el topónimo y la extensión del clúster.

Dado que a un mismo punto o región geográfica pueden estar asociados más de un topónimo, se propone crear un vector de topónimos relevantes, donde la inclusión de un topónimo en el vector o, en otras palabras, la relevancia del topónimo puede estar determinada por varios factores, por ejemplo:

- *Área*: un clúster puede cubrir varias entidades geográficas, entonces será más relevante un topónimo asociado a lugares con mayor área.
- *Nivel administrativo*: en un clúster de pequeño tamaño (parques, municipios, etc.) los topónimos de entidades locales tienen mayor relevancia que los topónimos de entidades regionales, nacionales, ya que los topónimos locales son más informativos y discriminantes en una búsqueda.

La idea es crear un conjunto o vector de topónimos relevantes que podrían caracterizar cada clúster. Por ejemplo, un clúster cuya extensión geográfica sea la comunidad autónoma de Aragón podría estar identificado por el vector de referencias geográficas indirectas (topónimos) y sus valores de ponderación o de relevancia de la siguiente manera:

$$a_4 = \langle Aragón, Huesca, Teruel, Zaragoza \rangle, \langle 1, 0.33, 0.31, 0.36 \rangle$$

Donde los valores de ponderación de cada topónimo corresponden al porcentaje de superficie cubierta con respecto la extensión de la comunidad de Aragón.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

4.4. ETL de información textual espacial

La información espacial de carácter textual generalmente no está estructurada en campos definidos, sino en texto libre. Esto hace necesario implementar módulos de reconocimiento espacial. Esta fase consiste en identificar y extraer a partir de la información geoespacial textual no estructurada (títulos, palaras clave, resumen, etc.) los nombres de lugares presentes en los metadatos de los servicios de cada clúster. Estos nombres de lugares, conocidos como *topónimos*, pueden requerir una transformación o adaptación. La transformación podría abarcar varios aspectos, desde homogeneizar los acentos hasta traducir de un idioma a otro u ofrecer una equivalencia entre varias formas en las que se puede encontrar un topónimo, como el caso del topónimo "*A Coruña*" o "*La Coruña*". El proceso extrae los topónimos de los metadatos y genera un vector de topónimos relevantes ponderado según la frecuencia de aparición de los topónimos en los diferentes metadatos del clúster.

Este proceso de filtrado y corrección sintáctica garantiza poder realizar comparaciones de topónimos de manera más flexible. Cabe mencionar que esta fase en sí misma presenta otros retos, ya que no es fácil identificar los *topónimos* que no se encuentran documentados como tal, sino que están como texto libre. Sumado a esto se ha observado que, posiblemente debido a irregularidades en diversas prácticas de documentación, los *topónimos* no siempre están asociados a la posición real del servicio, sino al proveedor, al publicador o, incluso, siendo un único proveedor, poseen servidores en diferentes lugares que también se incluyen como topónimos en el metadato. La falta de claridad de la proveniencia de estos *topónimos* y su pobre descripción aumentan el reto de la identificación, extracción, transformación, carga y limpieza de *topónimos*.

4.5. Validación

Para valorar la importancia de validar la consistencia de los metadatos Bruce y Hillmann (2004) planteaba la siguiente situación: un usuario que busca información geográfica espera recuperar colecciones de objetos espaciales similares utilizando criterios similares. El cumplimiento de esta situación depende de la coherencia y consistencia de los metadatos. La validación debe ayudar a certificar el consenso bajo el cual los datos están agrupados y deberían ser referenciados y buscados. En términos generales, esta fase, denominada *validación geo-semántica*, se vale de modelos de organización del conocimiento espacial para comprobar la correspondencia entre la descripción espacial del recurso y la información espacial contenida en el modelo, relacionada con la localización del recurso. Mientras los índices de validación en el algoritmo de *clustering* verifican que los clústeres formados son los mejores, esta validación verifica si un alto porcentaje de registros de metadatos de servicios del clúster poseen en sus descripciones los topónimos relevantes de la zona geográfica en la que se encuentran. En otras palabras, los topónimos de los metadatos serán comparados con los topónimos generados por la función de geocodificación inversa F_{GI}, de acuerdo al modelo de organización del conocimiento espacial de unidades administrativas ya mencionado.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

Esta correspondencia puede ser cuantificada usando el modelo de espacio vectorial (*Vector Space Model* VSM) (Salton *et al.*, 1975). Este modelo genera un valor o índice que cuantifica el número de elementos (topónimos) coincidentes entre los vectores. El índice puede ser calculado a través de la siguiente expresión:

$$\cos(\theta_i) = \frac{F_{GI}(ExtenC_i) \cdot TopoC_i}{\|F_{GI}(ExtenC_i)\| \|TopoC_i\|}$$

Donde $ExtenC_i$ es la extensión geográfica del clúster i-ésimo y $TopoC_i$ es el vector de topónimos de los registros de metadatos de los servicios *web* geoespaciales del clúster *i* ordenados según aparecen en el vector $ExtenC_i$. Aplicando el proceso de geocodificación inversa F_{GI} sobre $ExtenC_i$ se obtendrá un vector de topónimos α que caracteriza la extensión geográfica del clúster; los valores en cada posición del vector contienen el porcentaje de intercepción entre la extensión del clúster y la extensión espacial descrita en el metadato. Un vector β contendrá en cada posición el número de veces que un topónimo aparece en los registros de metadatos del clúster C_i. La expresión quedará así:

$$\cos(\theta_i) = \frac{\alpha_i \cdot \beta_i}{\|\alpha_i\| \|\beta_i\|} = W_i$$

Donde $||V||$ es la norma del vector. Al valor de este índice le llamaremos, de aquí en adelante, índice de consistencia *W*.

Un valor alto del índice *W* indica que el clúster *i* es consistente con las descripciones o referencias geográficas que aparecen en el sistema de organización del conocimiento *geográfico* usado por la función de geocodificación inversa F_{GI}. De esta manera se facilitará la detección de aquellos registros de metadatos inconsistentes con los registros de metadatos adyacentes en el clúster. Habrá mayor certeza al señalar un registro como inconsistente cuando éste pertenece a un clúster con índice de consistencia alto.

Un valor bajo del índice *W* indica que el clúster es inconsistente con la descripción que aparece en el sistema de organización del conocimiento *geográfico* usado por la función de geocodificación inversa F_{GI}. Pero también puede indicar que el sistema contra el cual se realiza la validación del clúster en particular no es el más apropiado. Incluso *W* puede generar una indeterminación cuando no existe ninguna correspondencia. Ahora bien, esto no representa una deficiencia del método propuesto, sino una alerta que podría usarse para revelar problemas de invisibilidad espacial de los metadatos, ya que en las búsquedas espaciales que usan los topónimos de esa área los resultados no regresarán estos recursos. Esto se debe a que sus metadatos no asocian el recurso con el lugar donde se realiza la búsqueda. Esta discrepancia semántica entre las referencias espaciales directas e indirectas es sacada a la luz cuando los valores de *W* no son altos. Valores de *W* muy bajos señalaría la presencia de posibles inconsistencias, implicando la

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

inexistencia de consenso entre el conjunto de topónimos de los registros y los topónimos característicos de la extensión representada por el clúster.

Un valor intermedio del índice implica que entre los registros de metadatos pertenecientes al clúster no hay acuerdo o consenso acerca de los topónimos que deberían caracterizar la extensión geográfica donde están agrupados. Otra posible explicación del desacuerdo de topónimos puede implicar la necesidad de usar otros sistemas para validar los registros, por ejemplo, una ontología de cuencas hidrográficas. Finalmente la falta de consenso entre los topónimos de un clúster podría producirse por la heterogeneidad de los recursos espaciales.

5. Descripción y análisis de resultados

Aplicando el método sobre un repositorio de 3.994 registros de metadatos de Servicios *Web* Geoespaciales de España se ha conseguido como resultado 84 clústeres usando un radio de 0.02 grados.

Los experimentos mostraron que para este caso de estudio este radio produce los mejores clústeres según el valor del índice y la filosofía de distribución de densidades del algoritmo DBSCAN. Estos clústeres se ilustran en la figura 3 como polígonos que cubren un grupo de puntos que representan registros de metadatos. Un acercamiento a uno de los clústeres se muestra en la figura 4, representando los registros de metadatos de la comunidad autónoma de La Rioja y su polígono envolvente que representa el clúster.

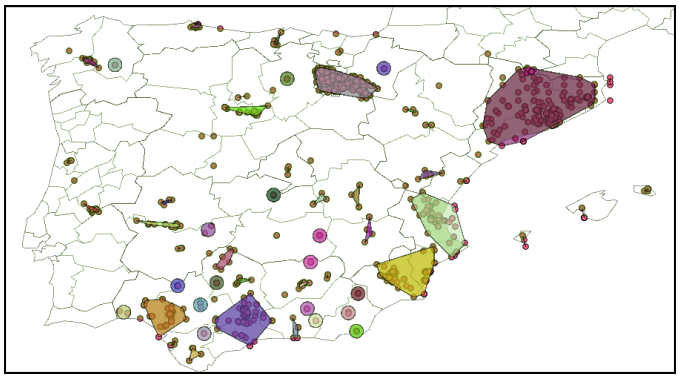

Figura 3. Clústeres de registros de metadatos detectados en España.
(Los polígonos envolventes representan clústeres de registros de metadatos, los registros son representados por los puntos, algunos clústeres solo poseen un punto porque sus elementos son concéntricos o muy próximos).

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

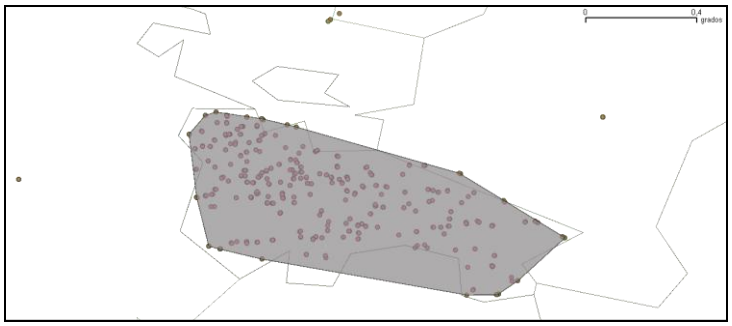

Figura 4. Clúster de los registros de metadatos de la comunidad de La Rioja.

La consistencia de un conjunto de topónimos que caracterizan un clúster puede definirse por medio de un umbral sobre el índice *W*; por ejemplo, hemos considerado que clústeres con valores del índice de consistencia $W > 0.5$ son vistos como consistentes. Partiendo de esta definición, un clúster es consistente cuando al menos el 50% de sus topónimos están bien asociados (mapeados) con la zona geográfica que cubren. Dichos topónimos característicos pueden ser útiles para enriquecer o corregir registros deficientes, incluso para anotar o sugerir topónimos para nuevos registros de metadatos cubriendo la misma zona.

Los resultados arrojaron que un 29,2% de clústeres de registros de metadatos poseen registros con topónimos efectivamente asociados a la zona geográfica de su clúster (véase figura 5). Un análisis detallado de cada clúster y de los valores del índice de consistencia *W* se presenta en el Apéndice 1. Los resultados muestran que un clúster tendrá mayor consistencia cuanto mayor sea el número de registros con topónimos coherentes o efectivamente asociados a la zona geográfica donde co-ocurren. Similarmente, cuanto mayor sea este número de registros coherentes dentro de un clúster más confiable será su caracterización. En la figura 6 se observa la línea de corte que señala los clústeres que poseen al menos un 50% de registros de metadatos consistentes (llamados aquí clústeres bien mapeados, valores del índice de consistencia *W* mayores que 0,5), es decir, clústeres con topónimos propios de la extensión geográfica cubierta. De forma similar, en este caso de estudio el 70,2% de los clústeres son inconsistentes (índice de consistencia $W < 0,5$); estos clústeres representan el 84,2% de los registros, equivalentes a 3.217 registros (véase la figura 7). Explorando los datos se observa que pocos clústeres agrupan la mayoría de los registros inconsistentes, por lo tanto, aprovechando el análisis y la caracterización espacial de pocos clústeres se pueden corregir muchos recursos, por ejemplo, los clústeres 17, 22 y 32 (ver Apéndice 1) son los más numerosos y representan más del 48% de los recursos. Esta mejora se logra gracias a la caracterización provista por los clústeres, ya que esta caracterización brinda pautas para asesorar la calidad.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

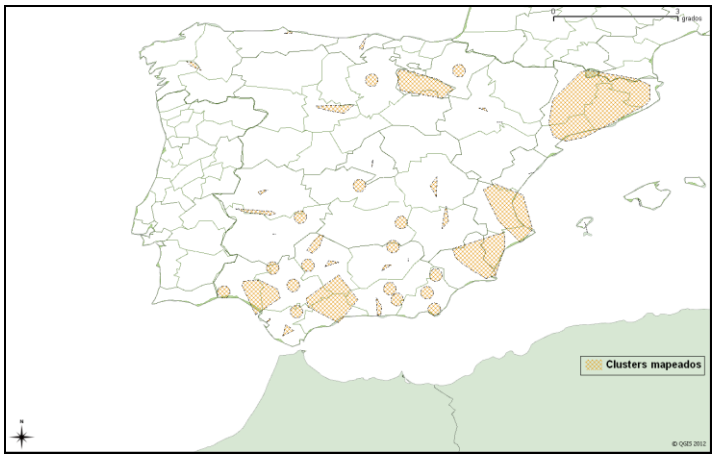

Figura 5. Clústeres consistentes.
(Estos clústeres poseen topónimos que se pudieron asociar a los topónimos del lugar).

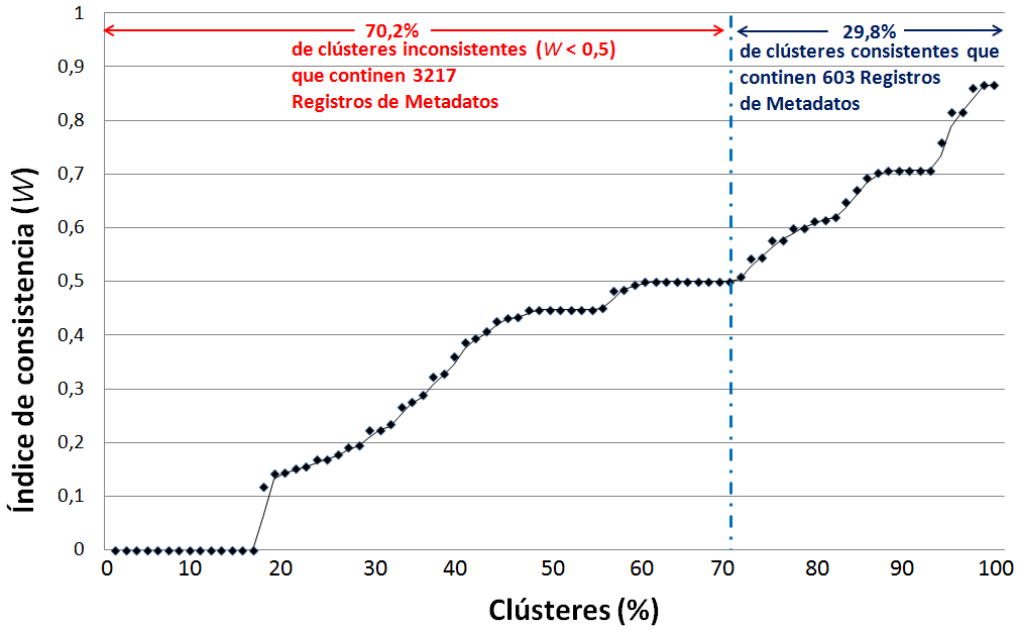

Figura 6. Gráfico de análisis de consistencia por clústeres.
(70,2% de los clústeres comprenden 3.217 registros de metadatos. Gracias a la caracterización provista por los clústeres se podría dar una asesoría para mejorar un gran número de registros. Mejorando pocos clústeres se consigue corregir muchos recursos, por ejemplo: los clústeres 17, 22 y 32. Ver Apéndice 1).

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

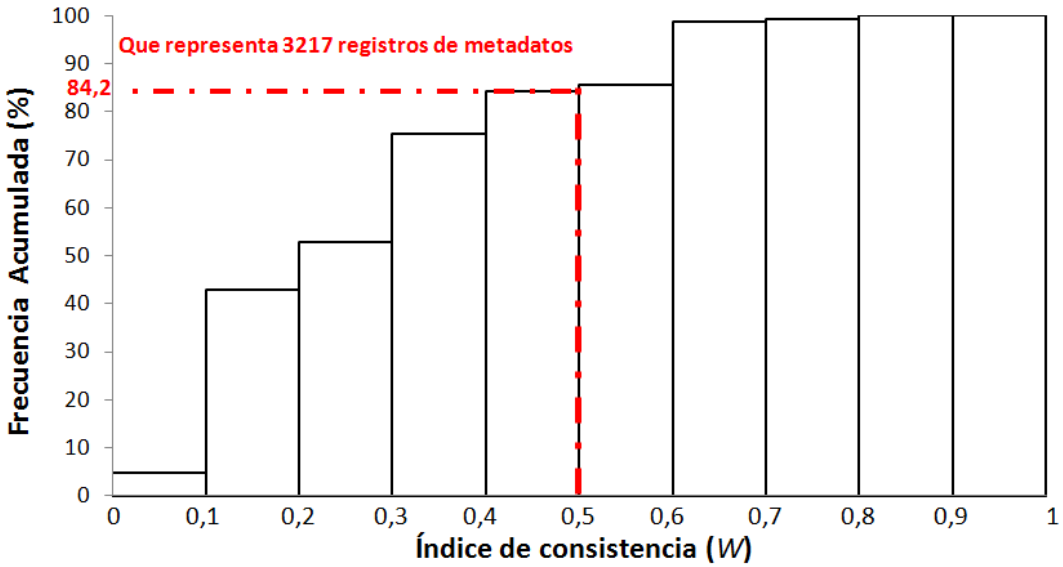

Figura 7. Diagrama de frecuencia del índice de consistencia de los registros de metadatos geoespaciales. Se consideran consistentes cuando el índice de consistencia ($W > 0,5$).

Algunas de las razones más comunes por las que el valor de consistencia *W* es tan bajo se deben a que existen registros con problemas de correspondencia entre sus topónimos y sus referencias espaciales directas. La segunda causa se centra en que los topónimos en los metadatos son muy generales respecto a la extensión específica que cubren. Adicionalmente, los metadatos incluyen muchos topónimos que se refieren a otros territorios que no están contenidos; dichos topónimos tendrá baja ponderación, esto disminuirá su valor de consistencia. Situaciones como las anteriores pueden llevar a potenciales problemas de invisibilidad.

En general, estos clústeres no pudieron ser asociados por el método, dado que no se encontró ninguna coincidencia de los topónimos procedentes de los registros y los topónimos asociados por el *geocoder* a la misma zona geográfica de su respectivo clúster. Al analizar los registros de metadatos se encontró que una de las causas por las cuales no se pudo establecer asociaciones radica en que los registros no tenían realmente topónimos, sino términos como "comunidad autónoma", "provincia" o "calle". De forma similar sucede con los registros de metadatos que no usan topónimos oficiales, los que usan nombres de lugar en otros idiomas o dialectos (ejemplo: mallorquín), clústeres que cubren zonas fronterizas (ejemplo: Andorra), clúster sobre zonas marítimas costeras. Ejemplos de estas situaciones son ilustrados por la figura 8. Un ejemplo más muestra clústeres cuyas descripciones señalan explícitamente que son metadatos de servicios correspondientes a una zona específica, pero su extensión geográfica cubre otra zona; el clúster A en la figura 8 contiene topónimos de entidades administrativas dentro de la comunidad autónoma de Extremadura, pero su extensión geográfica está sobre el Océano Atlántico.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

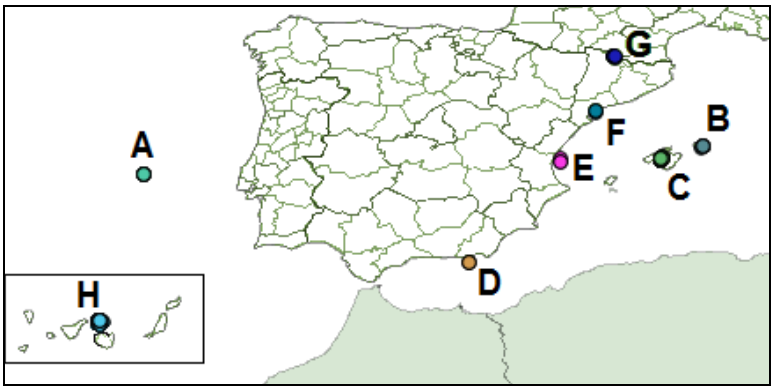

Figura 8. Caracterización de clústeres inconsistentes.
(A) con topónimos de entidades administrativas dentro de la comunidad autónoma de Extremadura, pero su extensión geográfica sobre el Océano Atlántico; **(B y C)** con topónimos en dialecto mallorquín; **(D, E, F)** clúster sobre zonas marítimas costeras; **(G)** cubriendo zonas fronterizas; **(H)** en medio de las islas.

Los valores del índice de consistencia pueden ser usados como indicadores de calidad para anotar a los registros de metadatos. También podrían contribuir a enriquecer registros sin descripción que cubren la misma extensión, para mejorar su visibilidad ante búsquedas espaciales de tipo textual. Esta función es llevada a cabo por el proceso de realimentación con el repositorio, señalado con la línea punteada en la figura 2. Esto puede constituir una herramienta útil en tareas de asesoramiento de la calidad de los recursos geográficos.

6. Discusión y valoración de resultados

Hay al menos seis puntos que merecen ser discutidos en relación al proceso de detección de inconsistencia en los registros de metadatos:

- Análisis individual de los registros o análisis grupal.
- Repercusión de la escasez de descripciones en los registros.
- Cuestiones que se pudieron resolver con la validación semántica.
- Tratamiento de servicios como una representación puntual.
- Enriquecimiento del sistema de organización del conocimiento.
- Enfoques que no usan clústeres sobres enfoques que los usan.

Respecto al primer punto, existen enfoques que apuestan por la validación manual elemento a elemento, ya que la verificación del experto permite elevar el nivel de consistencia (Lahesmaa-Korpinen *et al.*, 2010). Los enfoques que usan validación automática en el ámbito general de metadatos apuestan por sistemas que hagan frente a los grandes volúmenes de información y a la eficacia en los tiempos de análisis. Cuando se calculan los clústeres sobre los registros de metadatos de los servicios, las agrupaciones aportan señales de los topónimos

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

relevantes según el conocimiento de los expertos del dominio que previamente los documentaron, manifestando con esto un consenso respecto a los topónimos más representativos del lugar. Este consenso es sacado a la luz por medio del clúster que los agrupa. El análisis individual de registros sin usar clústeres no emplea la información de contexto y consenso que aporta el clúster, esto reduce la capacidad de validar con respecto al conocimiento adyacente de la comunidad de expertos. El contexto revelado por el clúster es el que permite identificar un elemento que se aparta del consenso para señalarlo como una posible inconsistencia.

En relación al segundo punto, en el experimento se encontró un 49% de registros de metadatos poco descriptivos, con mezcla de topónimos de lugar, de proveedor y publicador. A pesar de este porcentaje, con el método propuesto se consiguió identificar registros con inconsistencias cuando el valor del índice fue al menos del 50%. Esto también implica que cuando el porcentaje de registros carentes de descripción supera la mitad de los elementos del clúster se hace difícil sugerir correcciones y anotaciones confiables.

Respecto al tercer punto, la validación semántica permite responder preguntas como las siguientes:

- ¿Qué porcentaje de registros de metadatos pueden ser recuperados a través de una consulta espacial textual usando los topónimos de una zona geográfica específica?. La respuesta es el conjunto de registros de metadatos que superen un "mínimo valor" de consistencia con respecto al clúster al que pertenecen. Este valor mínimo puede ser fijado, por ejemplo, alrededor del 50%.
- ¿Qué porcentaje de clústeres presentan inconsistencia entre su extensión geográfica y sus referencias geográficas?. La respuesta viene directamente indicada por la suma de todos aquellos clúster que superen un valor de umbral dado, por ejemplo, clústeres que tienen al menos un 50% de registros bien anotados. Esto indicará todos aquellos clústeres cuya extensión geográfica es consistente con sus referencias geográficas de acuerdo al sistema de organización del conocimiento geográfico.

Referente al cuarto punto, la falta de distinción entre agrupaciones de servicios micro-locales, locales, regionales, y nacionales es un inconveniente detectado en los enfoques convencionales de obtención de clústeres, ya que hacen poco uso de factores de ponderación para agrupar estos tipos de servicios, distinguiendo así recursos geográficos que se encuentra representados por puntos y recursos que son representados por superficies (que sería el caso de los servicios).

En relación al quinto punto, los resultados muestran la necesidad de enriquecer el sistema de organización del conocimiento con información adicional, por ejemplo, incluyendo descripciones multilingües, información espacial fronteriza, información sobre zonas costeras y marítimas asociadas a las entidades territoriales, información de los topónimos alternativos de los territorios, además de los nombres oficiales. El enriquecimiento puede incluir sistemas de organización del conocimiento adicionales de otros dominios para solventar situaciones que no son precisamente inconsistencias.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

Finalmente, los enfoques basados en la obtención de clústeres presentan una ventaja adicional sobre enfoques que no lo usan. Un nuevo servicio cubriendo la misma zona del clúster podría documentarse aprovechando las descripciones espaciales que representan el clúster en su conjunto, agilizando las labores de documentación. Además, un nuevo servicio ya documentado podría validarse rápidamente con respecto a la documentación del clúster. Los clústeres proporcionan una vía para caracterizar espacialmente recursos que co-ocurren; la presencia de un consenso significativo en la caracterización puede ayudar a un posterior proceso de asesoramiento y curación de los datos.

7. Conclusiones y trabajo futuro

Debido a que los sistemas de búsqueda, en parte, deben su fiabilidad a la veracidad con la que los metadatos describen los recursos, es importante desarrollar nuevos mecanismos automáticos que asistan los sistemas de búsqueda espaciales para validar, corregir, curar y administrar los metadatos asociados a los Servicios *Web* Geoespaciales.

En este trabajo se han presentado las primeras fases de investigación de un método orientado a asegurar una validación semántica de los registros de metadatos asociados a los Servicios *Web* Geoespaciales, que detecta automáticamente posibles inconsistencias en la descripción espacial de los registros. Su detección y posterior enmienda puede mejorar la consistencia y, por ende, la confiabilidad de los análisis y las consultas que se realizan sobre la información de los sistemas de búsquedas, convirtiéndose en una herramienta de asesoramiento de la calidad de recursos geoespaciales.

El método propuesto se sustenta en la combinación de técnicas de agrupación espacial con métodos estadísticos para advertir tanto la presencia de inconsistencias como la ausencia de información espacial descriptiva relevante. Aplicado a los sistemas de búsqueda espaciales, este método podría mejorar la precisión de los resultados de búsquedas, ya que un recurso que no posee referencias espaciales indirectas (topónimos) de su localización puede pasar desapercibido ante consultas espaciales basadas en texto.

Han sido señalados algunos de los retos del tratamiento y preservación de la información en sistemas que usan registros de metadatos. Entre estos resulta de interés aquellos que están vinculados con la validación de la información de carácter espacial de los metadatos asociados a los recursos geográficos. Los retos mencionados indican la necesidad de trabajos de investigación enfocados a desarrollar mecanismos que asistan los procesos de documentación de registros de metadatos de Servicios *Web* Geoespaciales y recursos geográficos digitales en general.

Los resultados de investigación abren la puerta a futuros trabajos, principalmente en tres áreas: ampliar la gama de inconsistencias geosemánticas que pueden ser detectadas; explorar el uso de sistemas de organización del conocimiento de otros dominios que permitan enriquecer y validar las descripciones de los clúster de Servicios *Web* Geoespaciales; adicionalmente se plantea la

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

necesidad de validar los resultados contra otras pruebas. Una línea interesante a explorar consiste en usar técnicas de agrupación espacial que empleen representaciones espaciales superiores a una dimensión. Esta modificación permitiría capturar mejor la naturaleza bidimensional de las coberturas de Servicios *Web* Geoespaciales y aumentar la precisión de los resultados de los sistemas de búsqueda que emplean consultas espaciales.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Gobierno de España, a través del proyecto TIN2012-37826-C02-01, el Instituto Geográfico Nacional (IGN) y de GeoSpatiumLab S.L. El trabajo de Walter Rentería Agualimpia ha sido cofinanciado por el Gobierno de Aragón a través de la beca B181/11.

Referencias bibliográficas

Bruce, T.R. y Hillmann, D. (2004): "Metadata in practice, chap". *The continuum of metadata quality: defining, expressing, exploiting*, ALA Editions, Chicago, IL, pp. 238–256.

Capdevila, J.; Agudo, J.; Zarazaga-Soria, F.; Barrera, J.; Sánchez, A.; Soteres, C.; Criado, M. y Crespo, M. (2012): "Gateway MARC21-ISO19115: Definition and reference implementation", en *7th International Workshop on Digital Approaches to Cartographic Heritage*. Barcelona, (Spain) April, (1), pp. 66-73.

Ester, M.; Kriegel, H.P.; Sander, J. y Xu, X. (1996): "A density-based algorithm for discovering clusters in large spatial databases with noise", en *Proc. of the 2nd Intl Conf. on Knowledge Discovery and Data Mining KDD96*, AAAI Press, pp. 226-231.

Florczyk, A.; López-Pellicer, F.; Gayán-Asensio, D.; Rodrigo-Cardiel, P.; Latre, M. y Nogueras-Iso, J. (2009): "Compound Geocoder: get the right position", en *GSDI 11 World Conference and the 3rd INSPIRE Conference 2009*, Rotterdam 15-19 June 2009.

Ford, M.; Martirano, G.; Schleidt, K. y Vinci, F. (2011): "An INSPIRE validation briefcase for nature conservation and beyond", en *INSPIRE Conference 2011*, Edinburgh, Scotland. 2011.

GeoNames (2012): "GeoNames geographical database". [Consulta: 1-10-2012]. Disponible en http://www.geonames.org/.

Geocoder.ca (2012): "Reverse geocoder". [Consulta: 5-10-2012]. Disponible en http://geocoder.ca/?reverse=1.

Goodman, L. y Kruskal, W. (1954): "Measures of associations for cross-validations", en *J. Am. Stat. Assoc*. 49, pp. 732-764.

Google Geocoding API (2012): "Google Maps API Web Services". [Consulta: 15-10-2012]. Disponible en https://developers.google.com/maps/documentation/geocoding/?hl=en.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

Hartmann, J. y Stuckenschmidt, H. (2002): "Automatic Metadata Analysis for Environmental Information Systems", en *Proceedings of Environmental Informatics 2002*. Metropolis Verlag, Marburg (Germany)

Hijmans, R.; Garcia, N. y Wieczorek, J. (2012): "GADM database of Global Administrative Areas. GIS Shapefile". [Consulta: 15-5-2012]. Disponible en http://www.gadm.org/.

Hill, L.L. (2006): *Georeferencing: The Geographic Associations of Information*. Digital Libraries & Electronic Publishing, Cambridge, MA, The MIT Press.

Hubert, L. y Schultz, J. (1976): "Quadratic assignment as a general data-analysis strategy", en *British Journal of Mathematical and Statistical Psychologie*. 29, pp. 190-241.

International Organization for Standardization-ISO (2003): *International Standard: Geographic information-Metadata. ISO 19115:2003*. Technical Committee 211.

International Organization for Standardization-ISO (2007): *International Standard: Technical Specification: Geographic information-Metadata-XML Schema Implementation. ISO 19139:2007*. Technical Committee 211.

Jones, C. B.; Purves, R. S.; Clough, P. D. y Joho, H. (2008): "Modelling vague places with knowledge from the Web", en *International Journal of Geographical Information Science*, vol. 22, nº. 10, pp. 1045–1065.

Kliment, T.; Tuchyňa, M. y Kliment, M. (2012): "Methodology for conformance testing of spatial data infrastructure components including an example of its implementation in Slovakia", en *Slovak Journal of Civil Engineering*, Vol. 1, pp. 10-20.

López-Pellicer, F.J.; Lacasta, J.; Florczyk, A.; Nogueras-Iso, J. y Zarazaga-Soria, F.J. (2012a): "An Ontology for the representation of Spatio-Temporal Jurisdictional Domains in Information Retrieval Systems", en *International Journal of Geographical Information Science*. 2012, vol. 26, nº 4, pp. 579-597. doi:10.1080/13658816.2011.599811

López-Pellicer, F.J.; Renteria-Agualimpia, W.; Béjar, R.; Muro-Medrano, P.R. y Zarazaga-Soria, F.J. (2012b): "Availability of the OGC geoprocessing standard: March 2011 reality check", en *Computers and Geosciences*. 2012, vol. 47, pp. 13-19. Disponible en: http://dx.doi.org/10.1016/j.cageo.2011.10.023.

Lahesmaa-Korpinen, A.M.; Carlson, S.M.; White, F.M. y Hautaniemi, S. (2010): "Integrated data management and validation platform for phosphorylated tandem mass spectrometry data", en *Proteomics*, 10, pp. 3515-3524.

Milligan, G.W. (1981): "An algorithm for generating artificial test clusters", en *Psychometrika*, 50(1), pp. 123–127.

Monmonier, M. (1991): *How to Lie with Maps*. University of Chicago Press, Chicago, IL

Raatikka, V y Hyvönen, E. (2002): "Ontology-based semantic metadata validation", en *HIIT-Helsinki Institute for Information Technology Publications*, Number 2002-03, pp. 28-40. Disponible en http://www.hiit. /publications/.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

Salton, G.; Wong, A. y Yang, C. (1975): "A vector space model for automatic indexing", en *Communications of the ACM* 18, 11, pp. 613–620.

Sander, J.; Ester, M.; Kriegel, H. y Xu, X. (1998): "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications", en *Data Mining and Knowledge Discovery*, Springer Netherlands, Vol 2, 2. (1999), pp. 169-194. Disponible en doi 10.1023/A: 1009745219419.

Stvilia, B.; Gasser, L. y Twidale, M. (2007): "Metadata quality problems in federated collections", en L. Al-Hakim (Ed.): *Challenges of Managing Information Quality in Service Organizations*, IGI Global, pp. 154-186.

Sugihara, K.; Okabe, A. y Satoh, T. (2011): "Computational method for the point cluster analysis on networks", en *GeoInformatica* 15, pp. 167-189. Disponible en doi 10.1007/s10707-009-0092-5.

Turner, A. (2006): *Introduction to Neogeography*. O'Reilly Media (O'Reilly Short Cuts series).

Yu-hong, W. y Feng-yuan, W. (2008): "A Schema-matching-based Approach to Propagating Updates between Heterogeneous Spatial Databases", en *Proceedings of the SPIE*, Volume 7146, Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Advanced Spatial Data Models and Analyses, 714605-714605-10.

Yue, P.; Gong, J. y Di, L. (2010): "Augmenting Geospatial Data Provenance Through Metadata Tracking in Geospatial Service Chaining", en *Computers & Geosciences* 36 (3), pp. 270-281.

Zarazaga-Soria, F.J.; Lacasta, J.; Nogueras-Isso, J.; Torres, M.P. y Muro-Medrano, P.R. (2003): "A Java Tool for Creating ISO/FGDC Geographic Metadata", en *Geodaten- und Geodienste-Infrastrukturen - von der Forschung zur praktischen Anwendung*. Beiträge zu den Münsteraner GI-Tagen. IFGI prints. 2003, vol. 18, pp. 17-30.

© Los autores
www.geo-focus.org

Renteria-Agualimpia, W., López-Pellicer, J., Lacasta, J., Muro-Medrano, P., Zarazaga-Soria, F. (2013): "Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales", GeoFocus (Artículos), nº 13-1, p. 154-176. ISSN: 1578-5157

APÉNDICE 1
Clústeres de registros de metadatos de servicios *web* asociados a topónimos

Clúster Nº	Índice *W*	Total de registros		Clúster Nº	Índice *W*	Total de registros		Clúster Nº	Índice *W*	Total de registros
1	0	11		29	0,28	131		57	0,5	4
2	0	73		30	0,29	82		58	0,5	4
3	0	23		31	0,32	238		59	0,5	3
4	0	6		32	0,33	496		60	0,51	25
5	0	7		33	0,36	51		61	0,54	7
6	0	3		34	0,39	40		62	0,55	6
7	0	5		35	0,39	41		63	0,58	3
8	0	3		36	0,41	4		64	0,58	3
9	0	3		37	0,43	47		65	0,6	6
10	0	30		38	0,43	6		66	0,60	4
11	0	10		39	0,43	5		67	0,61	5
12	0	6		40	0,45	50		68	0,61	32
13	0	5		41	0,45	29		69	0,62	440
14	0	2		42	0,45	15		70	0,65	4
15	0,12	5		43	0,45	5		71	0,67	6
16	0,14	2		44	0,45	4		72	0,69	6
17	0,14	446		45	0,45	2		73	0,7	5
18	0,15	3		46	0,45	2		74	0,71	3
19	0,16	6		47	0,45	69		75	0,71	3
20	0,17	3		48	0,48	34		76	0,71	3
21	0,17	56		49	0,49	13		77	0,71	3
22	0,18	895		50	0,49	20		78	0,71	3
23	0,19	12		51	0,5	4		79	0,76	8
24	0,2	25		52	0,5	4		80	0,82	10
25	0,22	2		53	0,5	3		81	0,82	3
26	0,22	2		54	0,5	4		82	0,86	5
27	0,24	51		55	0,5	3		83	0,87	5
28	0,27	106		56	0,5	3		84	0,87	5

© Los autores
www.geo-focus.org