Sistema de acceso, catalogación y publicación de servicios OGC ofertados a través de las IDEs

QUINTANILLA, Antonio; CIFUENTES, David; SÁNCHEZ, Javier; MÁRQUEZ, Javier

La multitud de servicios e información espacial ofertados a través de las Infraestructuras de Datos Espaciales (IDEs) presentan dificultades para ser encontrados y utilizados. Con el fin de salvar esas dificultades y conseguir un uso generalizado de la información geográfica, se propone crear un repositorio de servicios OGC que permita consultas ordenables por criterios conceptuales y geográficos.

Se ha desarrollado un sistema software basado en servicios Web que cataloga de forma automática los servicios OGC disponibles en la red a nivel mundial, satisfaciendo la necesidad de inventariar y explotar de forma sencilla, mediante una API, la extracción de información de servidores y capas de los servicios OGC delimitada a un ámbito geográfico.

Operando de forma semejante a un buscador Web, proporciona como resultado fuentes de servicios que pueden ser explotados por aplicaciones y/o visores. De esta manera, no solamente se genera un repositorio central de servicios, sino que se contribuye al uso generalizado de la información espacial. Creando herramientas de verificación, de control de calidad de servicios y la oportunidad de generar nuevas líneas de negocio en el desarrollo de sistemas basados en SIG.

PALABRASCLAVE

Catálogo, servicios, OGC, WMS, IDE, búsqueda, repositorio, GetCapabilities

INTRODUCCIÓN

El Instituto de Desarrollo Regional de la Universidad de Castilla-La Mancha viene trabajando desde hace años en el diseño y desarrollo de sistemas que faciliten el uso de la gran cantidad de información geográfica publicada a través de las IDEs o de servicios de mapas comerciales. Debido principalmente a la gran cantidad de servicios OGC, la falta de descripción detallada y la heterogeneidad en los datos, se considera necesario la puesta en marcha de herramientas de usabilidad y explotación.

Se pretende ofrecer una herramienta capaz de facilitar esta búsqueda al igual que la realizan los buscadores Web convencionales a los que estamos acostumbrados. Para ello se ha planteado un sistema con una estructura que permite catalogar la información suministrada por la descripción de los servicios OGC, mediante peticiones "*GetCapabilities*", consiguiendo un data warehouse que sea explotado mediante un motor de búsqueda. Dicho almacén de datos está diseñado para favorecer el análisis y la divulgación eficiente de capas, features y servicios.

Las peticiones al sistema son de la forma <*términos, criterios [ámbito, SRS]*>. Se realizan consultas mediante términos que coincidan con la descripción de los servicios y puedan delimitarse para un determinado ámbito espacial para mejorar la búsqueda de registros. Obteniendo como respuesta <*capas/features, servidor, Spatial Reference System (SRS)*> representada según la importancia de un PageRank. Esta puntuación será calculada según la calidad de la información de los servicios.

RASTREO, REGISTRO Y PUBLICACIÓN

A continuación se detallan los procesos internos necesarios para recopilar la información en el sistema y su publicación.

Rastreo

El proceso más complejo es el rastreo de servidores [1]. Consiste en un programa que ejecuta un algoritmo que descubre enlaces correspondientes a servicios geoespaciales estandarizados por la OGC [2]. Cuando se encuentra un servicio válido se obtienen los metadatos mediante solicitudes *GetCapabilities* que describen la información del servicio. El robot o araña registra los metadatos del servicio disponible en un sistema (servidores y capas de información) de información según preponderancias. El robot puede procesar muchos tipos de contenido, pero no todos. Por ejemplo, no puede procesar el contenido de una serie de archivos multimedia o páginas dinámicas.

El proceso de rastreo comienza con una lista de URL de páginas Web generadas a partir de anteriores procesos de rastreo o URL's definidas previamente. Estas se amplían con los datos obtenidos en las Web (IDE's, Administraciones, Sitios especializados…) o mediante nuevos servicios que dan de alta los usuarios.

Indexación

A continuación el sistema procesa todas las URLs rastreadas para elaborar un índice completo de todas las palabras detectadas, e identificar la ubicación de cada servicio. El parseo de información obtenida de "*GetCapabilities*" recupera etiquetas y atributos de contenido clave ("Title", "Name", "Abstract", "CRS", "BoundingBox", etc), con el objetivo de ser registrados en el sistema data warehouse e indexados para mejorar el acceso.

El módulo de indexación es el subsistema encargado de manejar la estructura de índices asociados a contenidos. Gestiona la lectura/escritura de los datos indexados, estableciendo una correspondencia entre el contenido y los índices. Esto permite acceder a la información de carácter alfanumérico y establecer la relación entre un sistema de información y la ubicación del servicio.

A.1) Index Manager

La estructura de indexación consiste en un fichero invertido que cubre el ámbito geográfico del motor de búsqueda [5]. Asociados a cada registro, se mantiene información de los elementos de los servicios OGC (WMS, WFS…) e información que enlaza con el data warehouse del sistema. Esto permite un sistema híbrido que agiliza la búsqueda de conceptos y mantiene toda la información de los servicios registrados en el catálogo.

Los registros indexados en el fichero invertido *f* tienen la forma $\{c: l_i, b_i, s_i\}$, donde *c* es un concepto representado por una o más palabras (información extraída de las peticiones *GetCapabilities*); l_i es el identificador de una capa o registro cuyo título contiene el concepto *c*, pudiendo darse que *c* no tenga contenido el título pero si otro concepto; b_i es el bounding box de c_i; y finalmente, s_i es una puntuación cuyo rango es [0, 1] que mide el nivel de fiabilidad de la información ofrecida por la capa l_i sobre el concepto *c* y *b*.

El proceso de actualización de la estructura de índice se lleva a cabo en tiempo de indexación, esto permite realizar otras operaciones [3] que agilizan posteriormente la búsqueda. Este proceso se inicia después de encontrar un nuevo servicio WMS, WFS…. El procedimiento se divide en tres partes principales: parseo de información del servicio, indexación en el sistema de ficheros invertidos y puntuación "PageRank" para la recuperación de resultados.

A.2) Indexación de Capas

La primera parte del algoritmo es la actualización de la estructura de índices. En él se actualizan los ficheros invertidos una vez ha sido analizada la respuesta "*GetCapabilities*" del servicio. Por simplicidad se detallan los pasos que deben realizarse para cada capa l_i de los servicios WMS:

1. Obtener los metadatos del servicio mediante petición "*GetCapabilities*".
2. Puntuar la información según las variables de criterio establecidas para asignar el PageRank; {Tiempo de respuesta, concurrencia de palabras, SRS, BoundingBox…}

3. Registrar la información de cada capa *c* en el data warehouse y establecer los índices en los ficheros de indexación.
4. Registrar los términos conceptuales de valor del paso 1 {Nombre, Título, Abstract, CRS…} en los ficheros de indexación para agilizar la búsqueda.

Content Repository (Manejador de Consultas)

Capaz de navegar la estructura de índice con el fin de responder apropiadamente a las consultas. Las respuestas tienen un conjunto de resultados que se ordenan utilizando las puntuaciones s_i previamente calculadas. El manejador de consultas accede a los datos del sistema de información por cada capa l_i, obteniendo datos sobre título, abstract, escala máxima-mínima, bounding box, etc., y servicios (propietario, URL, forma de respuesta, etc.). El módulo gestiona los resultados según la calidad del servicio, ofreciéndola en forma de listado en base a los criterios de filtros aceptados por el API de recuperación de información.

Publicación

Cuando un usuario introduce una consulta {término-1, término-2…}, el sistema navega a través de la estructura de índices controlado por "Index Manager". Las coincidencias en términos y ámbito son recuperadas según relevancia s_i. La relevancia se determina mediante el PageRank de los registros, valorados por multitud de parámetros de calidad de los metadatos.

• Los ***términos*** de búsqueda son palabras descriptivas para recuperar la información de las capas $c_{i..n}$ registradas en el "catálogo".

• Los ***criterios*** son delimitadores de búsqueda que permiten obtener resultados óptimos. Se establecen criterios como zona de actuación (delimitada espacialmente mediante BoundingBox), sistema de referencia espacial (SRS), formatos de información u otros parámetros. Los *criterios* son especialmente útiles para recuperar registros (información de los servidores) con determinados pautas, que ayudan a generar aplicaciones y obtener datos de calidad. La información del "catálogo" está disponible de forma pública debido a que ha sido recuperada de servicios OGC disponibles en Internet.

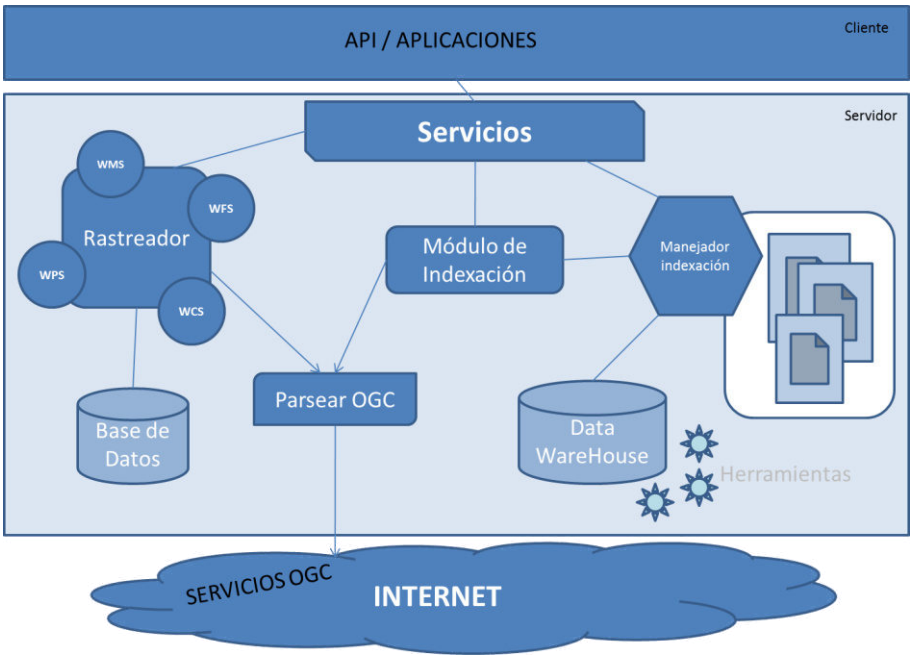

Figura 1: Arquitectura y comunicación entre los módulos del Sistema de Catálogo.

ANÁLISIS Y PROBLEMAS

La utilización de este tipo de herramientas permite extraer información y analizar los resultados de la calidad de los datos. Este artículo se centra en servicios WMS [4] debido a que se tratan de los más populares en las IDEs. Se han analizado un conjunto de estos servicios, a nivel estatal, regional y local con el propósito de focalizar la atención en los errores más comunes en su definición, así como sugerir buenas prácticas, lo que permitiría solventar las dificultades de descubrir servicios y recuperar las capas según la descripción de sus términos [5]. La calidad de los metadatos tiene un impacto directo en la recuperación de la información (Comprovts 2004) [6].

La aplicación de servicio de catálogo permite realizar consultas para obtener información válida de los datos indexados. Por lo que no solamente se trata de un repositorio de servicios OGC para recuperar capas de información por ámbito y descripción, sino de una potente herramienta de análisis de generación de servicios para detectar errores.

De los metadatos de los servicios, en analogía a la definición del estándar ISO 19115 y la directiva INSPIRE, se han seleccionado un conjunto de registros esenciales con el fin de facilitar la recuperación de los registros de información a nivel de servicio. Por tanto es de gran importancia el enriquecimiento de los metadatos ofertados por el servicio *GetCapabilities*, ayudando a conseguir resultados más precisos ya que de otro modo, posiblemente, no podrían ser encontrados.

Los principales metadatos evaluados y registrados en el catálogo han sido:

- Tittle (Obligatorio ISO 19115 e INSPIRE)
- Abstract (Obligatorio ISO 19115 e INSPIRE)
- Resource language (Obligatorio ISO 19115 e INSPIRE)
- Metadata date (Obligatorio ISO 19115 e INSPIRE)
- Topic category (Obligatorio ISO 19115 e INSPIRE)
- Responsible (party information or role) (Obligatorio INSPIRE, parcial ISO 19115)
- Geographic bounding box (Obligatorio INSPIRE)
- Keyword value (Obligatorio INSPIRE)
- Reference System Info

Los campos con mayor peso de actuación en la recuperación de información son “Title”, “Abstract”, “Geographic Bounding Box” y “Reference System Info”.

Los principales problemas encontrados han sido la falta de definición de los campos “Title” y/o “Abstract”. Se pueden encontrar nombres genéricos, títulos incompletos, “Absctract” sin especificar o sin aporte de información respecto a “Title”.

Otro de los parámetros analizados de importancia es el “Geographic Bounding Box”, que determina el ámbito de actuación de la información. Este parámetro se ha considerado muy interesante a nivel de aplicación debido a que permite realizar búsquedas espaciales de la información y devolver al usuario aquellos registros de la zona de actuación de trabajo. El principal problema radica en una mala definición del BoundingBox (BBOX), ocasionando que el sistema sea incapaz de recuperarla adecuadamente para determinados casos puntuales. Se han encontrado circunstancias donde el BBOX no corresponde con la zona geográfica o se encuentra definido a nivel mundial, cuando únicamente se trata de una zona mucho más restringida.

La ausencia del campo ScaleHint es otro de los problemas, ya que este permite identificar a que escala de visualización se encuentra la información de la capa. Se ha detectado que la mayoría de los servicios no lo especifican o se encuentra mal definido.

El último parámetro de estudio se refiere al “Reference System Info” en el que se ofertan los servicios. Gran parte de los servidores no soportan la proyección Google Mercator, muy habitual en servicios WEB-Mapping de uso muy popular; Google, Yahoo, OpenStreetMap, MapQuest, Bing. Ofertar los servicios en Google Mercator (EPSG:900913) supondría un gran revulsivo en la utilización por parte de los usuarios.

Véase la Tabla 1 donde se detallan algunos los servicios WMS analizados y los servidores que ofertan la proyección Google Mercator. El estudio se ha centrado en las siguientes IDEs. Comunidad autónoma de Andalucía que presenta un amplio servicio de mapas desde sus IDEs, el estudio del servicio del Instituto Geográfico Nacional (IGN), una IDE local como Zaragoza y algunos de los servicios WMS ofrecidos por el Ministerio de Agricultura, Alimentación y Medio Ambiente (MAGRAMA).

	Servicios Indexados	EPSG:4326 Coordenadas Geográficas WGS84	EPSG:23030 Proyección UTM ED50 Huso 30 N	EPSG:25830 Proyección UTM ETRS89 Huso 30 N	EPSG:900913 Proyección Google Mercator
ANDALUCIA*	580 (3705 capas)	578	579	571	488
IGN	12 (49 capas)	12	12	12	1
IDE ZARAGOZA	4 (42 capas)	4	4	4	0
MAGRAMA	29 (29 capas)	29	29	29	0

* Servicios IDEs analizados en Andalucía:
http://www.juntadeandalucia.es,
http://www.ideandalucia.es,
http://siggra.dipgra.es/siggra/mapservers,
http://mapserver.eprinsa.es/cgi-bin/eiel?,
http://www.idejaen.es/wms?,
http://www.idemap.es/jac/ArcGIS/services/Ortos5000/MapServer/WMSServer

Tabla 1: Proyecciones soportadas por los distintos WMS analizados.

ACTUALIZACIÓN DE LA INFORMACIÓN

La actualización de la información es clave para ofrecer un servicio de calidad. El sistema está definido a tres niveles.

Automático: Un sistema *cron* actúa cada cierto tiempo comprobando la disponibilidad de los servicios que se tienen registrados. Esta labor es importante para disponer de un catálogo de registros actualizados y ofrecer un servicio de calidad basado en PageRank.

Manual: A nivel manual, es el propietario el que introduce el servicio en el catálogo, o bien el técnico-administrador del sistema, el que puede editar la información. Esto permite enriquecer la descripción de las capas.

Asistido: Mediante un sistema de incidencias que son recibidas por el técnico-administrador del sistema, notificándole las solicitudes de usuarios y valoraciones que se realicen sobre un determinado registro. Consiguiendo una mejora de la información por parte de una comunidad, ya que la información puede ser enriquecida o pudiéndose establecer un contacto con la persona encargada de la publicación del servicio.

En el desarrollo del sistema software se han evaluado los distintos niveles planteados. La alternativa más sostenible es la de un sistema *cron* automático de verificación de la información. Esto convierte al sistema en un repositorio de registros donde el usuario y/o aplicaciones externas puedan obtener resultados según los criterios de calidad de los metadatos. Esta alternativa es la que aplican los motores de búsqueda en analogía a las páginas Web. El diseñador de la Web es el encargado de establecer los parámetros necesarios para aparecer en los primeros puestos del buscador. La idea es

análoga, pero a nivel de capas/registros de servicios OGC, esto puede ayudar a que los propietarios de los servicios de información realicen definiciones más completas.

CONSULTA/API

El sistema implementa una fachada de servicios REST, permitiendo generar aplicaciones que realicen consultas al repositorio de una forma sencilla mediante URLS. Los principales parámetros de petición son: {N-términos} palabras que describan el elemento a recuperar y criterios de filtro {BBOX, SRS, FORMAT...}.

http://<catalogo_servicios>/SearchLayers?query=termino&bbox=boundingbox

APLICACIONES Y USOS DE LA HERRAMIENTA

El sistema software planteado no es únicamente un repositorio de servicios WMS para acceder a la información. Abre un abanico de líneas de explotación debido a su integración con otros sistemas. Actualmente, a nivel de servicio WMS lo más parecido es la Web de la IDEE [7], pero esto no permite obtener las capas de forma rápida, únicamente las URL de los servicios.

A partir del software planteado, se ha creado la aplicación WizardGIS [8], un sistema que permite crear visores personalizados con una cartografía base. Esto se realiza basándose en servicios proporcionados por Google, Bing, OpenStreetMap y muchos otros, así como cualquier servicio Web de mapas (WMS) disponible en Internet. El usuario no tiene necesidad de conocer las direcciones URL de los servicios WMS que desea incorporar a su mapa. Realizando una consulta mediante una herramienta estándar de búsqueda, selecciona de un listado de capas, las más adecuadas para incorporar al visor de mapas que está creando. WizardGIS es una herramienta muy sencilla y útil para crear y publicar visores de mapas temáticos (estudios urbanísticos, medio ambientales, prototipos, impactos, callejeros, etc.). En su primera versión se encontró con la dificultad de ofertar un servicio de catálogo de capas eficiente que facilitase a cualquier usuario, experto en SIG o no, construir su visor de mapas. El sistema de catálogo planteado pretende solventar esa dificultad. WizardGIS es una alternativa a GeoNode [9], una plataforma para compartir datos y mapas. Ambas nacen con la misma filosofía, con la ventaja que WizardGIS permite realizar búsquedas acotadas a nivel espacial y definir el tipo de proyección, algo muy interesante para entornos profesionales.

Figura 2: Paso 4 de WizardGIS. Selección de capas a incorporar en el mapa. Utilización del servicio de catálogo descrito en el artículo.

Dejando atrás las bondades de emplear el sistema de catálogo para construir aplicaciones como WizardGIS, la información almacenada puede ser explotada mediante herramientas de análisis de datos. Esto contribuye a la optimización de servicios, validación y calificación para estos según los estándares ISO 19115 e INSPIRE. Tanto los "publicadores" como los usuarios pueden usar la aplicación propuesta para verificar la interoperabilidad de la información publicada y así mejorar los servicios ofertados.

CONCLUSIÓN

El sistema de catálogo es una herramienta para "centralizar" servicios OGC y recuperarlos por palabras claves de forma sencilla y rápida. Pero, además, es una buena herramienta que permite analizar la gran cantidad de información publicada por servidores de mapas, sirviendo de *tester* de los servicios ofertados.

La primera conclusión de los tests realizados en cuanto a la calidad de los metadatos es la falta de homogeneización de estos y la existencia de notables carencias en su definición. Esta heterogeneidad, dificulta enormemente la posibilidad de utilizar la información, crear mapas enriquecidos y publicarlos.

Los Datum más usuales en España (UTM ED50, UTM ETRS89 y UTM WGS84) no siempre son ofrecidos en su conjunto en los servicios analizados.

Para poder mostrar información disponible en servicios OGC junto a otra cartografía propietaria como Google Maps, Yahoo, Bing, etc, es necesario que las diferentes administraciones publiquen sus servicios en la proyección Spherical Mercator (EPSG:900913). Actualmente, muy pocos servicios, a excepción de los provenientes de Andalucía cumplen dicho requisito.

La asociación de los metadatos a las capas es una deficiencia generalizada, por lo que se concluye la necesidad de trabajar en la línea de crear herramientas que ayuden a garantizar los estándares de publicación de servicios. Sería de gran interés para el caso de MapServer establecer herramientas que garantizaran la creación de ficheros .map de forma visual, asegurando la definición de metadatos y el empleo de la etiqueta "wms_inspire_capabilities", cumpliendo con los campos obligatorios en la especificación del estándar.

Para aprovechar al máximo el gran esfuerzo realizado por parte de la OGC en la definición de estándares y, por parte de administraciones públicas, la publicación de servicios, es necesario disponer de herramientas y metodologías que permitan crear servicios con metadatos de calidad. De esta manera, se podrán crear aplicaciones sobre dichos servicios realmente útiles y con un nivel de usabilidad aceptable que consiga llegar, de una vez por todas, al ciudadano medio.

El catálogo de servicios propuestos podría contribuir, no sólo como un repositorio central y buscador de información espacial, sino como una herramienta para optimizar estos servicios. Crear un conjunto de datos de valor para establecer un sistema que contribuya al uso de aplicaciones SIG y establecer buenas prácticas de actuación en la definición de los metadatos.

El objetivo final es favorecer la mejora los servicios publicados, su homogeneidad y usabilidad para conseguir nuevas líneas de negocio.

REFERENCIAS

[1] Li, Wenwen, Yang, Chaowei and Yang. An active crawler for discovering geospatial Web services and their distribution pattern – A case study of OGC Web Map Service. International Journal of Geographical Information Science. 2010, Vol. 24, Nº8 , 1127-1147.

[2] N. Chen, J. Gong, and Z. Chen. A high precision ogc web map service retrieval based on capability aware spatial search engine. In Proceedings of the 2nd international conference on Advances in computation and intelligence, ISICA'07, pages 558-567, Berlin, Heidelberg, 2007. Springer-Verlag.

[3] J. Márquez, J. E. Corcoles, and A. Quintanilla. A semantic index structure for integrating ogc services in a spatial search engine. In IEEE International Conference On Open Systems, Kuala Lumpur, Malaysia. 2010.

[4] Francisco J. Lopez-Pellicer, Rubén Bejer, Aneta J. Florczyk, Pedro r. Muro-Medrano, F. Javier Zarazaga-Soria. A review of the implementation of OGC Web Services across Europe. International Journal of Spatial Data Infrastructures Research, 2011, Vol. 6, 168-186

[5] Francisco J. Lopez-Pellicer, Aneta J. Florcyk, Rúben Béjar, Javier Nogueras-Iso, F. Javier Zarazaga-Soria, Pedro R. Muro-Medrano. State of Play: Spain and Portugal. SDI services' state of play in autumn 2010. I Jornadas Ibéricas de Infra-estructuras de datos Espaciales.

[6] J. Compvoets,A. Breat, A. Rajabifard, I.Willioamson, (2004). Assesing the worldwide development of nacional spatial data clearinghouses. International Journal of Geographical Information Science, 18:7, 665-689.

[7] IDEE. *Infraestructura de Datos Espaciales de España* (http://www.idee.es) Servicios obtenidos del directorio WMS.

[8] Antonio Quintanilla, Javier Márquez, Carlos Baños, Javier Sánchez. WizardGIS, un asistente de generación de visores cartográficos personalizados. II Jornadas Ibéricas de Infra-estructuras de datos Espaciales. 2011.

[9] GeoNode. Plataforma de código libre para compartir datos y mapas. (http://www.geonode.org).

AUTORES

Antonio QUINTANILLA
Antonio.quintanilla@uclm.es
Universidad de Castilla-La Mancha
IDR

David CIFUENTES
cifu15@gmail.com
Universidad de Castilla-La Mancha
IDR

Javier SANCHEZ
jsanc70@hotmail.com
Universidad de Castilla-La Mancha
IDR

Javier MÁRQUEZ
javimarkez@gmail.com
Universidad de Castilla-La Mancha
IDR