Metadatos de servicios estándares. Cómo compartirlos y gestionarlos.

Alejandro Guinea de Salas[1], Anja Ludewig[2] .

[1]Geograma SL
Castillo Lantaron, 8 Vitoria-Gasteiz
Tel: 902 99 55 84,Fax: +34945230340,
geograma@geograma.com

[2] University of Applied Sciences
Friedrich-List-Platz 1, D-01069 Dresden

Resumen

Actualmente los servidores de catálogos han llegado a un estado de madurez importante. Gracias a las herramientas opensource, implementar un servidor de catálogo es relativamente fácil. Por otro lado, la comunidad SIG ha puesto a disposición de cualquiera herramientas que permiten generar metadatos según los estándares aprobados y más utilizados. Cualquier empresa o institución productora de información geográfica puede ya introducir en sus procedimientos habituales la creación de metadatos y, en teoría, aprovechar todas las ventajas que proporcionan los estándares para compartir, explotar y gestionar los catálogos de información. De hecho, los estudios demuestran que ya existen en España decenas de miles de metadatos creados. Sin embargo, a pesar de que la teoría lo permite, existen todavía ciertas dificultades para que la aplicación práctica de las herramientas sea un hecho.

El artículo trata sobre la aplicación de las herramientas a la gestión (no a la producción) de los metadatos, y cómo aplicar las recomendaciones del SGT de catálogo de una forma práctica, con ejemplos y comentarios sobre las recomendaciones. Además, se explicará una solución desarrollada para la búsqueda de servicios estándares mediante harvesting o recolección de los metadatos contenidos en los propios servicios OGC.

Palabras clave: Metadatos, OCG, Servicios OGC, servidores de metadatos, WMS, búsqueda de servicios de mapas.

1 Introducción

La expansión del protocolo Web Map Service (WMS) ha marcado un hito. Las aspiraciones de INSPIRE [1] eran utópicas en una época en la que no existían apenas mapas en internet, las webs de descarga de cartografía eran inexistentes, y los formatos propietarios eran llamados estándares y adquirían esta condición gracias a su uso y no a una norma determinada.

El Open Geospatial Consortium (OGC) [2] comenzó a definir unos estándares ambiciosos, como primer paso para la interoperabilidad entre sistemas de información geográficos.

El acrónimo GIS evoluciona hacia IDE y éste le mira por encima del hombro transmitiendo interoperabilidad, difusión, eficacia y resultados. Todo ello ha culminado con el WMS mencionado al inicio, como primera evidencia de que las piezas OGC – Inspire comienzan a encajar.

1.1 Los metadatos en este nuevo escenario

Los metadatos, antes desestructurados e incluso olvidados, se han ido homogeneizando hasta tal punto de cumplir rigurosos estándares como pieza clave de las futuras Infraestructuras de Datos Espaciales. La pieza que debería ayudar a encontrar y descubrir los servicios que necesitamos, en un momento dado y para un uso determinado. No hay un mapa o un servicio bueno o malo. Hay un mapa o un servicio que me sirve o no me sirve. Y esto no lo puedo determinar sin metadatos. Sin embargo, la creación de metadatos aún no es una actividad asentada, y su publicación aún menos. La consecuencia es que tenemos servicios estándares, pero los metadatos que nos permiten encontrar y evaluar estos servicios no resultan accesibles.

2 Compartir metadatos

Afortunadamente, existe un Subgrupo de trabajo de la IDEE, el Subgrupo de Trabajo CAT [3] que ha acometido la tarea de estudiar este nicho concreto, intentar

explicar las causas esta situación y proponer soluciones. Una primera conclusión es que, además de la dificultad que puede tener la elaboración de metadatos, los trabajos necesarios para la explotación, gestión y publicación de los mismos son bastante costosos.

Por otro lado, y a pesar de los estándares, los catálogos distribuidos adolecen de una serie de matices aún por resolver. El Subgrupo de trabajo de la IDEE está realizando unos estudios en relación a esto, y propone una serie de recomendaciones para facilitar la creación de catálogos distribuidos. Los problemas se centran, sobre todo, en los grados de libertad que dejan los estándares, que dificultan la comunicación entre los servidores que alojan los catálogos distribuidos, así como el protocolo o mecanismo de acceso a la información.

En este sentido, se plantean una serie de recomendaciones diferenciadas, en un principio, en dos frentes claramente diferenciados: el mecanismo y el formato de acceso.

1.1 Formato de acceso

El primer punto de las recomendaciones de refiere al formato concreto de acceso a la información. Basado en XML, y en los estándares ISO 19115 y Dublin Core. Debido a que OGC no ha llegado a desarrollar un test de conformidad de catálogo para las ninguna de las interfaces propuestas en el estándar, es necesario contemplar ciertas restricciones que pueden afectar a la viabilidad de la interoperabilidad. Se remite al documento elaborado por el SGT CAT para profundizar en estos aspectos propios del interface.

2.1 Mecanismo de acceso

El otro frente es el protocolo o mecanismo de acceso a la información, que podrá ser mediante un interfaz OGC apoyado en un servidor de servicios de catálogo, o bien el acceso mediante protocolos más básicos de publicación en web (http, ftp, etc). Esta segunda opción resulta muy interesante para entidades de mediano pequeño tamaño, por lo ágil y económica que puede resultar su publicación.

Los servidores de catálogo poseen una gran madurez y existen numerosas soluciones, muchas de ellas de software libre, que permiten desplegar un servidor de metadatos. Uno de estos ejemplos, quizás el más utilizado, sea Geonetwork, un desarrollo Open Source en Java distribuido, que integra las siguientes aplicaciones:

- Un editor de metadatos que soporta los principales estándares de metadatos (Dublín-Core, ISO19115, FGDC)
- Un servidor de catálogo que permite publicar los metadatos.
- Un motor de búsqueda que permite realizar búsquedas de información geoespacial tanto en el servidor de catálogo local como en servidores de catálogo externos.
- Un módulo de administración para gestionar los usuarios y permisos del servidor de catálogo.
- Se puede usar Geonetwork para configurar servicios (WMS, WFS, WCS, SLD etc), crear metadatos, entidades, etc.

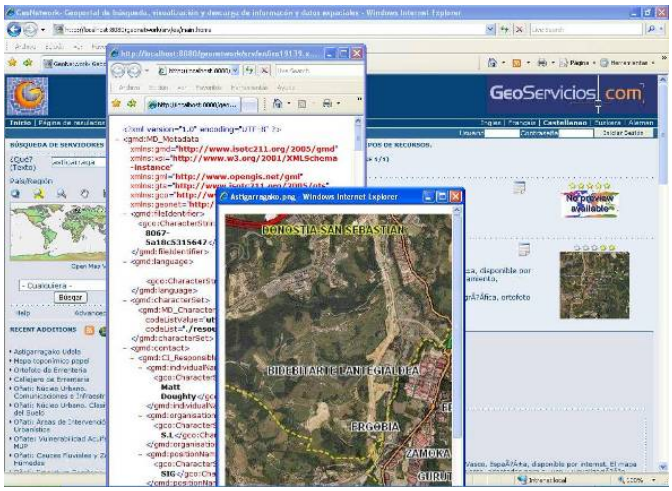

Figura 1. Publicación de un catálogo de metadatos con GeoNetwork

El problema es que estos servidores de catálogo, aunque cumplan los estándares definidos por el OGC, como se indica en el informe final del proyecto "Software for Distributed Metadata Catalogue Services to Support the EU Portal" [JRC 2006], son muchos los grados de libertad que dejan los estándares OGC de catálogo. Esto implica la necesaria toma de decisiones a la hora del diseño e implentación es necesario tomar muchas decisiones, que no necesariamente serán las mismas en todos los casos, con lo que se impide una comunicación fluida que posibilite el desarrollo de un servidor de catálogo distribuido.

Sin seguir profundizando en los servidores de catálogo como vía para la publicación de metadatos, existe otra vía ya mencionada, que es muy interesante desde el punto de vista tecnológico, por lo sencillo, ágil y económico que resulta el método.

Esta vía consiste en publicar los ficheros XML directamente en un servidor web, con los metadatos en formato ISO 19139, la extensión de los fichero sería .xml, y estos ficheros estarían acompañados por un fichero "capabilities.xml"

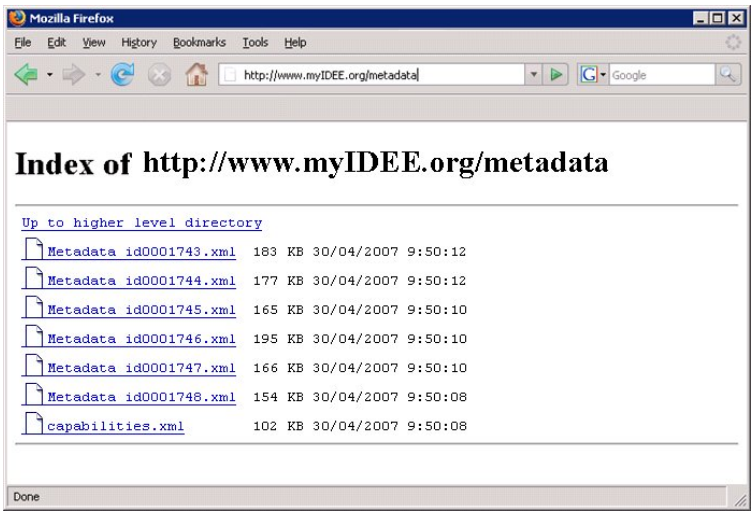

Figura 2: Ejemplo de publicación de Metadatos mediante Directorio Web

Este fichero capabilities.xml recoge los datos generales del servicio de catálogo según la especificación definida por el SGTCAT. De esta forma tan sencilla se estaría publicando los metadatos de una forma interoperable. Con sólo dar a conocer la URL del directorio de publicación, los metadatos estaría accesibles, y sería posible comenzar a construir un catálogo distribuido con los nodos de la IDEE. La publicación de estas URL's se haría a través de e-mail o formulario directamente en la IDEE, incoporando poco a poco estos nodos.

3 Harvesting de metadatos

Paralelamente a los avances y las conclusiones desarrolladas por el SGTCAT, se ha desarrollado una aplicación de Harvesting de los metadatos que están recogidos de forma implícita en los propios servicios OGC, concretamente en los servicios WMS y WFS. En definitiva, se trataba de explorar los dos caminos para encontrar y utilizar servicios concretos: a partir de los metadatos encontrar los servicios, y a partir de los servicios encontrar los metadatos. El proyecto se ha desarrollado conjuntamente con estudiantes de la University of Applied Sciences de Dresden (Alemania).

Los servicios WMS, como la mayoría de servicios OGC, tienen un método llamado GetCapabilities, que permite conocer las características (capacidades) de ese servicio. Todavía estas capacidades no están enlazadas con los metadatos, pero La especificación WMS 1.3 define un elemento MetadataURL para cada capa del WMS en la que se supone que puedes especificar donde están los metadatos de esa capa. Hasta que esta especificación esté definida, no se conocen más metadatos, que los que se pueden ver en el método GetCapabilities. Sin embargo, estos datos son más relevantes y tienen más posibilidades de lo que puede parecer en un principio:

Visualizando la respuesta al método GetCapabilites de cualquier servicio WMS, se pueden observar datos sobre el proveedor, descripción del servicio, descripción de las capas, ámbito que cubre el servicio, sistemas de coordenadas que soporta, palabras clave, derechos de uso. En definitiva, en la práctica podríamos obtener prácticamente todos los datos necesarios para evaluar si el servicio es el que buscamos o no. Podemos incluso obtener imágenes de la leyenda, que sin duda contribuye a conocer mejor tanto el servicio como la información que proporciona, como se puede ver en la siguiente imagen.

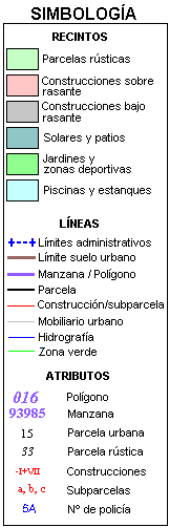

Figura 3: Imagen proporcionada por el servicio WMS de catastro

Si fuéramos capaces de recopilar de forma indexada aquellos campos que resultan importantes para localizar un servicio (no para su consumo), podríamos facilitar el acceso a los servicios.

El proyecto, realiza una recopilación de metadatos mediante dos aplicaciones diferenciadas, el Spider o araña, que rastrea la web para localizar los servicios, y el motor de indexación que realiza las tareas necesarias para facilitar la búsqueda.

3.1 Desarrollo de un spider o araña web

Un web crawler (o araña de la web) es un programa que inspecciona las páginas del World Wide Web de forma metódica y automatizada [4]. Los Web crawlers se utilizan para crear una copia de todas las páginas web visitadas para su procesado posterior por un motor de búsqueda que indexa las páginas proporcionando un sistema de búsquedas rápido.

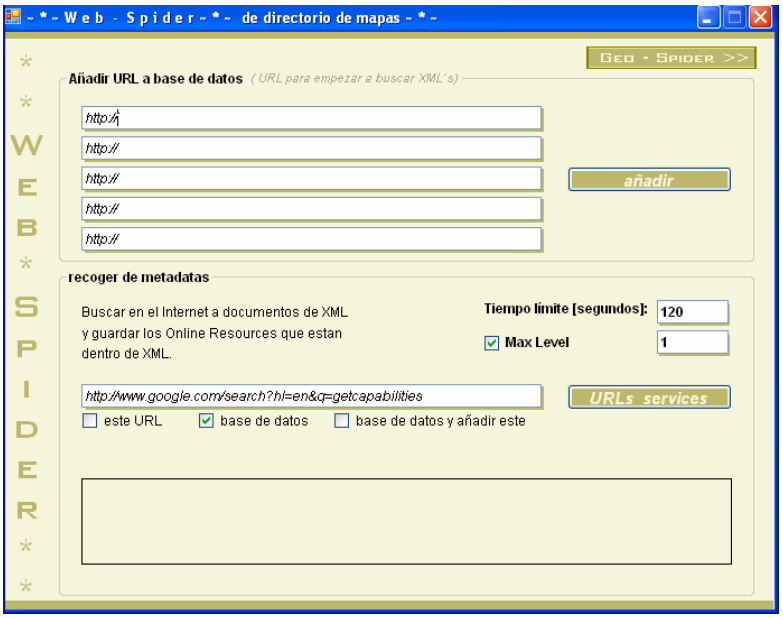

Figura 4: Interface del Web Crawler

Para el proyecto de harvesting de metadatos, Una araña rastrea la web en busca de resultados xml devueltos por urls que contienen el comando getcapabilities. De esta forma se intenta resolver uno de los principales problemas del acceso a los servicios y/o metadatos, que es conocer la URL de acceso a los mismos. Asimismo, es posible introducir URLs de servicios directamente en la base de datos mediante un formulario.

Se trata de desarrollar una aplicación que recoge los datos relativos a la URL que devuelve el método GetCapabilities y los almacene de manera indexada en una base de datos. La ventaja de utilizar una araña diseñada específicamente, además de automatizar el proceso, permite guardar las informaciones más relevantes, desechando las demás.

En el proyecto piloto se han conseguido algo más de 2.000 URL's de todo el mundo, con unas 10.000 capas asociadas.

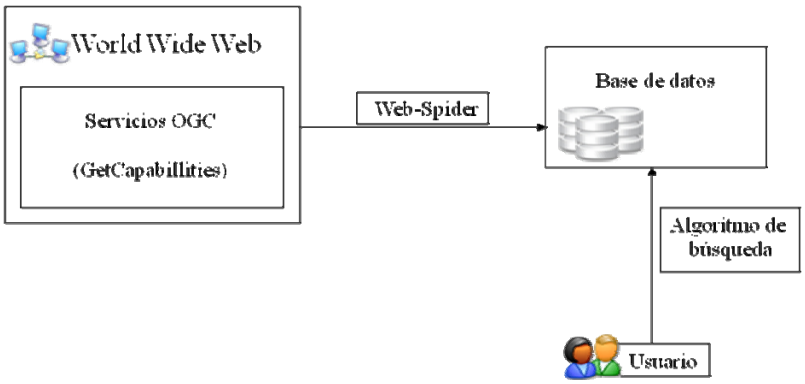

Figura 4: Esquema del sistema de Harvesting

La ejecución de la araña no estuvo exenta de dificultades durante el lanzamiento del proyecto piloto. Saturación de la conexión y protestas de los usuarios, limitación del router de conexiones simultáneas, colapso del ancho de banda, modificación del comportamiento de los motores de búsqueda debido al número de peticiones, necesidad de limitación en los niveles de profundidad de búsqueda del web crawler son sólo algunas de ellas. En definitiva, el sólo despliegue de la araña de búsqueda tuvo una magnitud considerable.

La búsqueda en la web está complementada con la posibilidad de volcar URLs de servicios directamente en la base de datos, facilitando y acelerando la inclusión de los mismos en el motor de búsqueda.

4.1 Motor de indexación

Una vez recopiladas las direcciones de los servicios que devuelven un xml con el formato getcapabilities, una aplicación paralela recorre periódicamente cada uno de los servicios recopilando aquellos metadatos que aportan valor para la realización de búsquedas. Por ejemplo, a pesar de que las proyecciones disponibles puede ser un dato importante para su utilización, no es un parámetro que se utilice normalmente para la búsqueda de un servicio, por lo que no se recopila.

Con el fin de acelerar los procesos de indexación, registro y búsqueda, se limitaron los campos a almacenar, quedando el modelo de datos como se refleja en la siguiente figura:

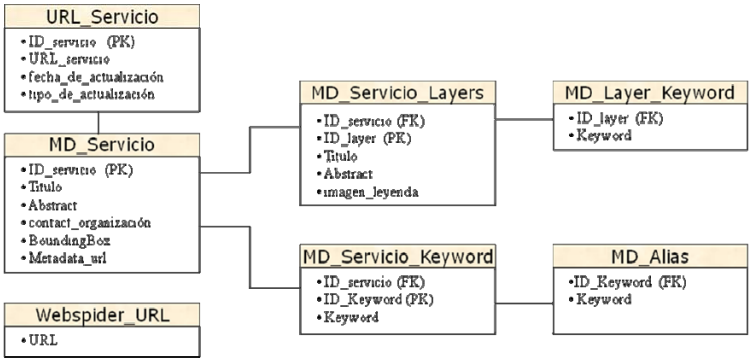

Figura 5: Modelo de datos

En el proyecto piloto realizado se ha detectado la poca importancia que se le está dando a los campos que permiten describir los servicios, a pesar de que pueden servir de gran ayuda tanto a la búsqueda de los mismos como a su utilización, y durante la recopilación se ha detectado también un porcentaje relativamente alto de servicios que no devuelven un XML correctamente construido. La definición de los límites del servicio también ha permitido obtener conclusiones interesantes, que han de tratarse con especial atención.

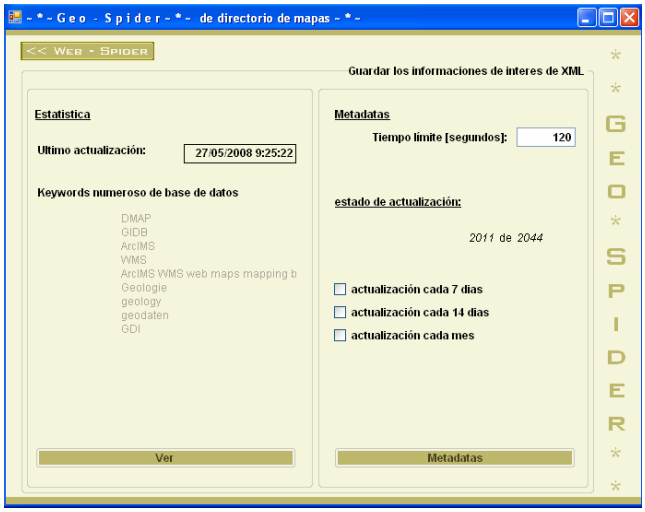

Figura 6: Interface del motor de indexación

Los servicios OGC permiten, además, acceder a los metadatos del servicio mediante una etiqueta específica, lo que permite, ya de forma on line, ampliar la información de cada servicio o incluso de cada una de las capas de cada servicio.
Gracias a la recopilación de los metadatos, ya es posible realizar una serie de interesantes búsquedas, que facilitan el descubrimiento de los servicios y su utilización. La utilización conjunta de la base de datos recopilada junto con servicios WMS que permitan publicar los límites de los servicios (BBOX) cierra un círculo que abre muchas oportunidades.

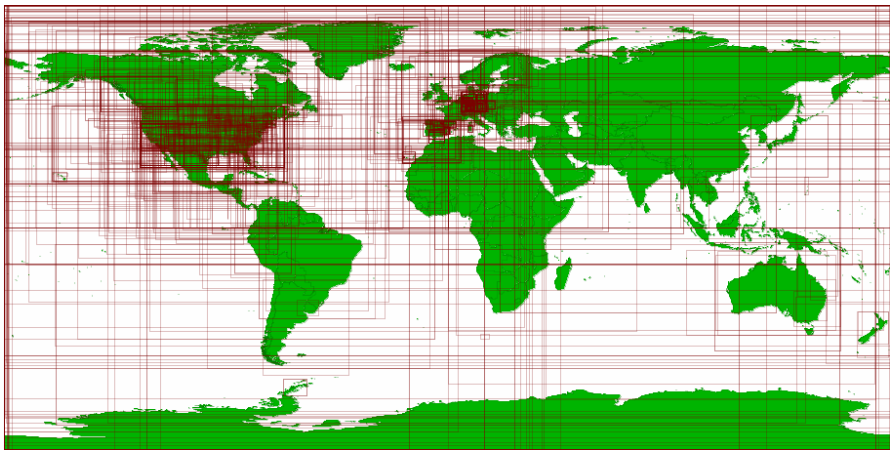

Figura 7: Representación gráfica de los límites de los servicios encontrados

4 Conclusiones y propuestas de actuación

Hay metadatos, hay catálogos, hay servicios que contienen metadatos, pero aún no hay herramientas que exploten estos datos. El desarrollo de un catálogo distribuido es una opción, que puede ser complementada con desarrollos específicos de harvesting de metadatos.

Los metadatos intrínsecos al servicio puede proporcionar una valiosa información que no se está aprovechando. Los propietarios de servicios WMS no son conscientes aún de la importancia que tienen estos metadatos asociados al servicio más allá del propio servicio, como se observa por la ausencia de keywords y otros datos de interés que podrían facilitar su localización y su uso.

Teniendo en cuenta el indicador del número de servicios publicados como grado de madurez en los mismos, España tiene una posición destacada en Europa y en el mundo, lo que es una oportunidad de la que empresas e instituciones deberían ser conscientes para trabajar conjuntamente de forma que la oportunidad se consolide realmente.

Dados los interesantes resultados que se están obteniendo, se invita a cualquiera que pueda estar interesado a sentar las bases para crear un equipo de trabajo conjunto (instituciones, empresas, particulares o universidad) y seguir desarrollando el proyecto de forma colaborativa.

Agradecimientos. Nos gustaría hacer especial mención al Subgrupo de Trabajo de Catálogo de la IDEE [3] tanto por la importancia del documento de recomendaciones como paso hacia la consolidación de un catálogo en la IDEE, como por la propuesta para simplificar la publicación de metadatos.

5 Referencias

[1] Directiva INSPIRE. http://inspire.jrc.ec.europa.eu/

[2] OGC. Open Geospatial Consortium. http://www.opengeospatial.org/

[3] F.J. Zarazaga, C. Laborda, J.M. Agudo, A.F. Rodríguez: Recomendaciones sobre interfaces de catálogo de datos. Subgrupo de Trabajo CAT SGTCAT20061116.

[4] Wikipedia. http://www.wikipedia.org