PROYECTO DE RECOPILACIÓN E INDEXACIÓN DE METADATOS PARA FACILITAR EL DESCUBRIMIENTO Y UTILIZACIÓN DE SERVICIOS GEOGRÁFICOS ESTÁNDARES

Alejandro Guinea de Salas
Socio director
Geograma S.L.
Castillo de Lantarón, 8 bajo
01007 Vitoria-Gasteiz (Alava)
www.geograma.com
Diciembre de 2008

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

1 Sumario

El proyecto se presenta como una evolución de un proyecto de desarrollo de un rastreador o spider web para la recopilación de servicio web geográficos, que 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. En el caso de 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.

El artículo explica los avances en el proyecto de harvesting, las posibilidades que brinda para organizar los metadatos, y cómo puede llegar a solucionar los problemas de interoperabilidad actualmente existentes entre los diferentes catálogos de metadatos para ir avanzando en la construcción de un catálogo de metadatos distribuido.

2 El método Getcapabilities de los servicios OGC

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.

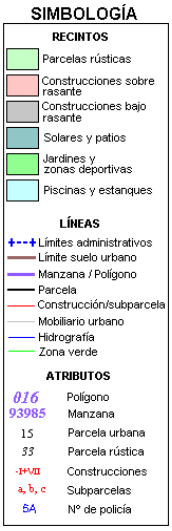


Figura 1: Imagen proporcionada por el servicio WMS de catastro

A continuación se presenta la parte de la respuesta al método GetCapabilities del servicio WMS de catastro:

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>
<!DOCTYPE WMT_MS_Capabilities (View Source for full doctype...)>
- <WMT_MS_Capabilities version="1.1.1" updateSequence="0">
- <Service>
  <Name>OGC:WMS</Name>
  <Title>Cartografía catastral</Title>
   <Abstract>Cartografía Catastral de la Dirección General del Catastro. Este servicio es de uso libre y gratuito. La
      cartografía se actualiza diariamente desde las bases cartográficas del Catastro. No tiene la categoría de cartografía
      oficial, por lo que no debe ser utilizada para ningún tipo de certificado. No está permitida la descarga masiva de
      porciones de cartografía. La D.G. del Catastro se reserva el derecho de restricción del servicio por abuso del
      mismo.</Abstract>
- <KeywordList>
  <Keyword>WMS</Keyword>
  <Keyword>CARTOGRAFIA</Keyword>
  <Keyword>CATASTRO</Keyword>
    </KeywordList>
  <OnlineResource xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="simple" xlink:href="http://ovc.catastro.meh.es"
    />
- <ContactInformation>
- <ContactPersonPrimary>
  <ContactPerson>LINEA DIRECTA DEL CATASTRO, llamando al 902 37 36 35</ContactPerson>
  <ContactOrganization>Oficina Virtual del Catastro DIRECCION GENERAL DEL CATASTRO</ContactOrganization>
    </ContactPersonPrimary>
  <ContactElectronicMailAddress>soporte.ovc@catastro.meh.es</ContactElectronicMailAddress>
    </ContactInformation>
  <Fees>Acceso gratuito.</Fees>
  <AccessConstraints>Acceso libre, pero se prohibe la descarga masiva de porciones de cartografía.</AccessConstraints>
    </Service>
- <Capability>
- <Request>
- <GetCapabilities>
  <Format>application/vnd.ogc.wms_xml</Format>
- <DCPType>
- <HTTP>
- <Get>
              <OnlineResource xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="simple"
    xlink:href="http://ovc.catastro.meh.es/Cartografia/WMS/ServidorWMS.aspx" />
    </Get>
    </HTTP>
    </DCPType>
    </GetCapabilities>
- <GetMap>
  <Format>image/png</Format>
  <Format>image/jpeg</Format>
  <Format>image/gif</Format>
```

```
  <Format>image/bmp</Format>
  <Format>image/tif</Format>
- <DCPType>
- <HTTP>
- <Get>
                <OnlineResource                xmlns:xlink="http://www.w3.org/1999/xlink"                xlink:type="simple"
      xlink:href="http://ovc.catastro.meh.es/Cartografia/WMS/ServidorWMS.aspx" />
    </Get>
    </HTTP>
    </DCPType>
    </GetMap>
- <GetFeatureInfo>
  <Format>text/html</Format>
  <Format>text/xml</Format>
- <DCPType>
- <HTTP>
- <Get>
                <OnlineResource                xmlns:xlink="http://www.w3.org/1999/xlink"                xlink:type="simple"
      xlink:href="http://ovc.catastro.meh.es/Cartografia/WMS/ServidorWMS.aspx" />
    </Get>
    </HTTP>
    </DCPType>
    </GetFeatureInfo>
    </Request>
- <Exception>
  <Format>application/vnd.ogc.se_xml</Format>
  <Format>text/xml</Format>
    </Exception>
  <VendorSpecificCapabilities />
- <Layer queryable="0" opaque="0" noSubsets="0">
  <Title>Cartografía Catastral</Title>
  <SRS>EPSG:23030</SRS>
  <SRS>EPSG:4230</SRS>
  <SRS>EPSG:4326</SRS>
  <SRS>EPSG:4258</SRS>
  <SRS>EPSG:32627</SRS>
  <SRS>EPSG:32628</SRS>
  <SRS>EPSG:23029</SRS>
  <SRS>EPSG:23031</SRS>
  <SRS>EPSG:32629</SRS>
  <SRS>EPSG:32630</SRS>
  <SRS>EPSG:32631</SRS>
  <SRS>EPSG:25829</SRS>
  <SRS>EPSG:25830</SRS>
  <SRS>EPSG:25831</SRS>
  <LatLonBoundingBox minx="-18.409876" miny="26.275447" maxx="5.22598" maxy="44.85536" />
  <BoundingBox SRS="EPSG:4230" minx="-18.409876" miny="26.275447" maxx="5.22598" maxy="44.85536" />
  <BoundingBox SRS="EPSG:4326" minx="-18.409876" miny="26.275447" maxx="5.22598" maxy="44.85536" />
  <BoundingBox SRS="EPSG:4258" minx="-18.409876" miny="26.275447" maxx="5.22598" maxy="44.85536" />
  <BoundingBox SRS="EPSG:32627" minx="770000" miny="3000000" maxx="2700000" maxy="5000000" />
  <BoundingBox SRS="EPSG:32628" minx="180000" miny="3000000" maxx="2170000" maxy="5000000" />
  <BoundingBox SRS="EPSG:23029" minx="-410000" miny="3000000" maxx="1650000" maxy="5000000" />
  <BoundingBox SRS="EPSG:23030" minx="-1050000" miny="3000000" maxx="1150000" maxy="5000000" />
  <BoundingBox SRS="EPSG:23031" minx="-1615000" miny="3000000" maxx="620000" maxy="5000000" />
  <BoundingBox SRS="EPSG:32629" minx="-410000" miny="3000000" maxx="1650000" maxy="5000000" />
  <BoundingBox SRS="EPSG:32630" minx="-1050000" miny="3000000" maxx="1150000" maxy="5000000" />
  <BoundingBox SRS="EPSG:32631" minx="-1615000" miny="3000000" maxx="620000" maxy="5000000" />
  <BoundingBox SRS="EPSG:25829" minx="-410000" miny="3000000" maxx="1650000" maxy="5000000" />
  <BoundingBox SRS="EPSG:25830" minx="-1050000" miny="3000000" maxx="1150000" maxy="5000000" />
  <BoundingBox SRS="EPSG:25831" minx="-1615000" miny="3000000" maxx="620000" maxy="5000000" />
- <Layer queryable="1" opaque="0" noSubsets="0">
  <Name>Catastro</Name>
  <Title>Catastro</Title>
- <Style>
  <Name>Default</Name>
  <Title>Default</Title>
- <LegendURL width="160" height="500">
  <Format>image/png</Format>
                <OnlineResource                xmlns:xlink="http://www.w3.org/1999/xlink"                xlink:type="simple"
    xlink:href="http://ovc.catastro.meh.es/Cartografia/WMS/simbolos.png" />
  </LegendURL>
  </Style>
```


Como se puede observar en el fichero XML, hay una serie de datos fácilmente accesibles que proporcionan información sobre el servicio. 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 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.

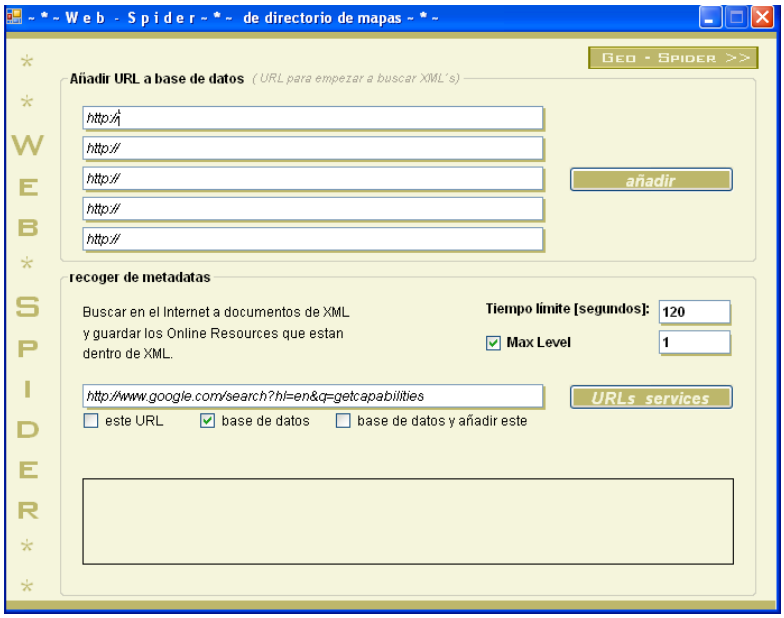


Figura 2: Interface del Web Crawler

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.

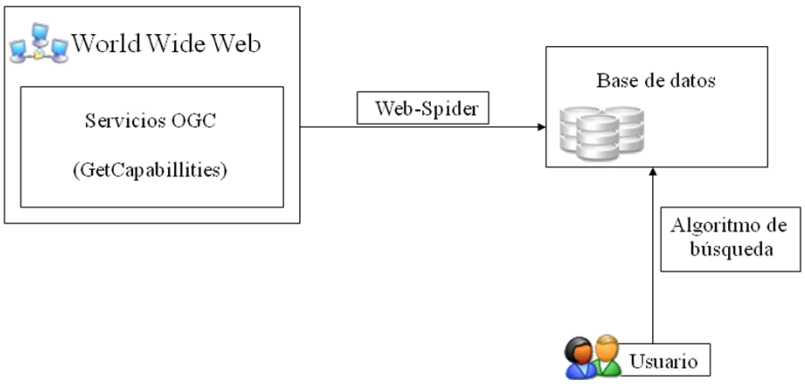


Figura 3: Esquema del sistema de Harvesting

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. Este proyecto piloto está sirviendo para definir una estrategia de despliegue del webcrawler totalmente diferente a la inicialmente prevista, con el fin de garantizar un paso a producción correcto.

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 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:

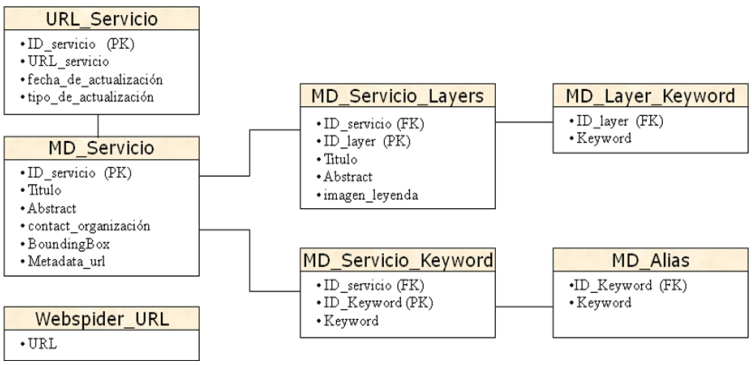


Figura 4: Modelo de datos

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.

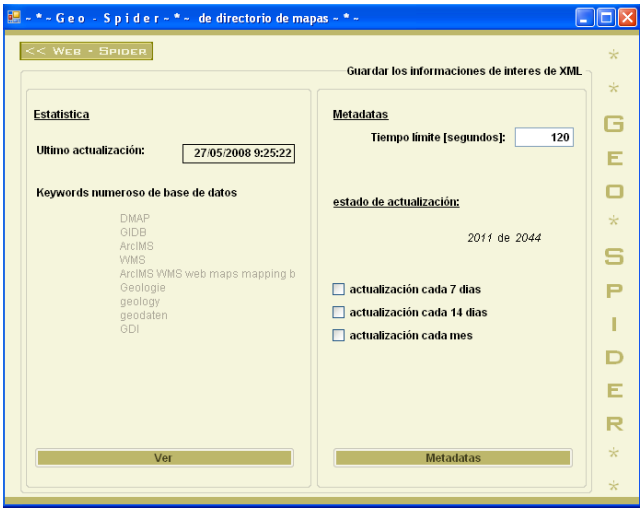


Figura 5: Interface del motor de indexación

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.

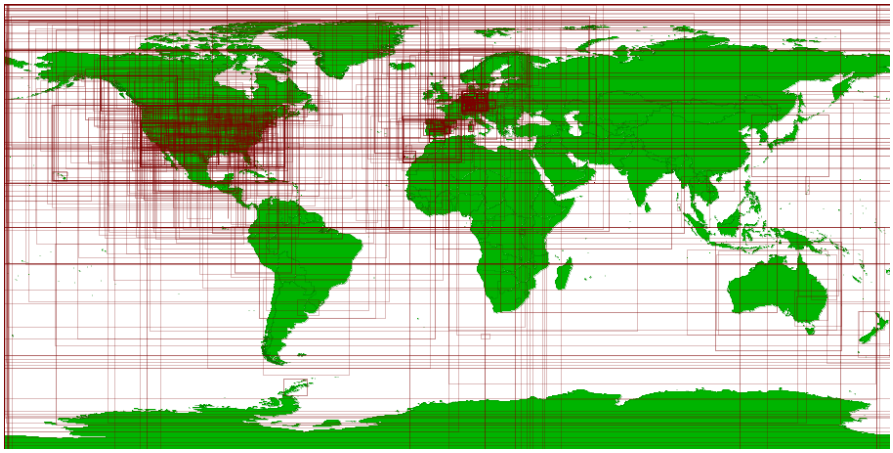


Figura 6: Representación gráfica de los límites de los servicios encontrados

5 Calidad de los metadatos

Una vez realizado el proyecto piloto, se ha realizado un estudio de la calidad de los datos recogidos, con el fin de conocer el estado del arte en lo relativo a los metadatos de servicios estándares, y tenerlo en cuenta a la hora de indexar las variables recogidas, generar alias y afinar el modelo de datos.

Respuesta de los servicios

Para una correcta recopilación de los metadatos, es necesario una correcta recuperación de los XML devueltos por los servicios en el método Getcapabilites. Durante el proyecto piloto, el 93% devolvieron una respuesta correcta. Del 7% restante, un 1% no permite utilizar el XML debido a una mala construcción del mismo, en un 1% de los casos el servidor no respondió, y en un 5% de las respuestas el valor del marco del servicio (boundingbox) fue incorrecto. Este valor es utilizado por el motor de indexación para poder localizar el servicio en el territorio, y puede impedir un correcto acceso al servicio dependiendo de la aplicación cliente.

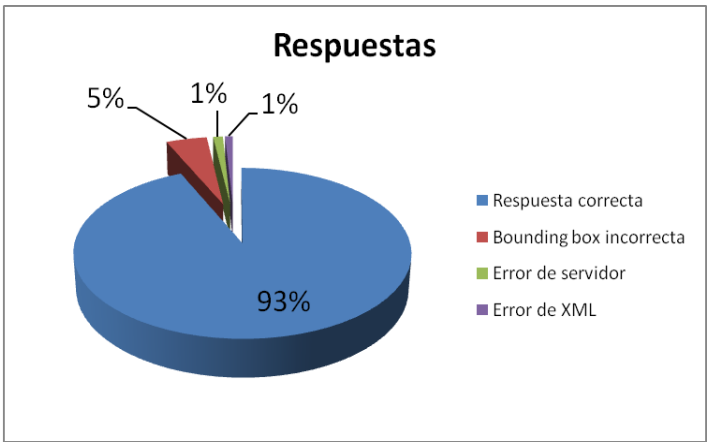


Figura 7: Tipos de respuestas recibidas en los servicios

Metadatos relativos a los servicios

Los tres campos estudiados son el título, el resumen (abstract) y el organismo de contacto. El análisis no ha tenido en cuenta el valor de los campos, tan sólo la presencia de los mismos. El dato más presente el título, se encuentra en un 94% de los servicios.

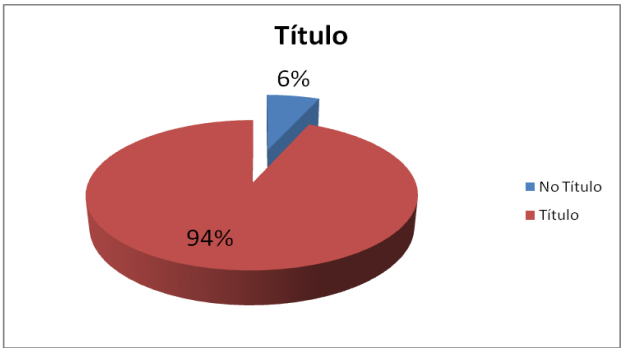


Figura 8: Presencia del título en los servicios

Respecto al resumen del servicio, que complementa al título, añadiendo más información sobre el mismo, aparece en un 85% de los servicios.

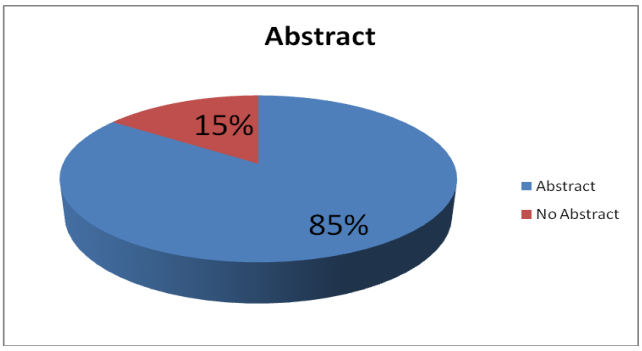


Figura 9: Presencia del resumen en los servicios

Los datos de contacto, que recogen el organismo responsable de los mismos, están presentes en el 86% de los servicios. Este dato es importante cuando se desea profundizar sobre el servicio, realizar alguna consulta sobre el mismo, comunicar incidencias, o simplemente contactar con el responsable.

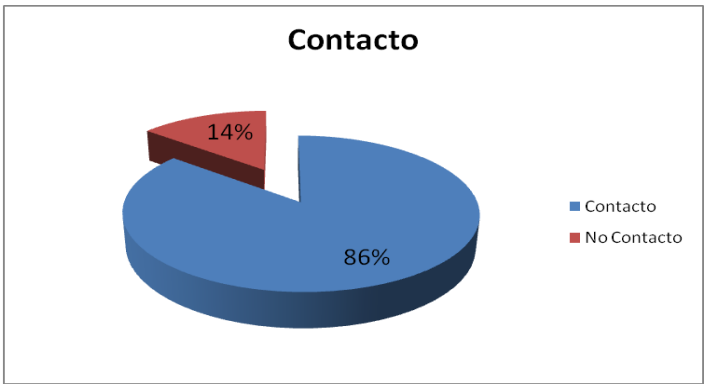


Figura 10: Presencia de la organización de contacto en los servicios

Como conclusión en cuanto a los campos descritos, un 73% de los servicios recopilados poseen todos los campos, mientras que el resto, o bien no tienen datos o sólo tienen parte de ellos.

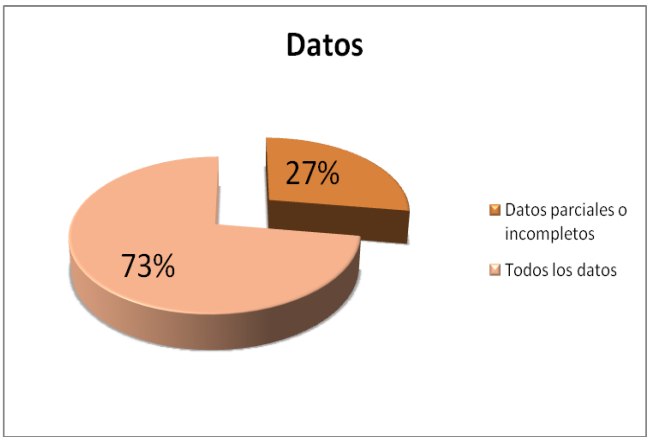


Figura 11: Porcentaje de servicios con los datos completos o incompletos

Metadatos relativos a las capas o layers

Los servicios de mapas ofrecen un acceso concreto a capas o niveles de información o layers. Cada una de estas capas posee una serie de metadatos que ayudan a identificar las mismas. Los campos estudiados han sido el título, el abastract y la imagen de leyenda. El título de la capa, al ser un campo obligatorio para la explotación del servicio, aparece en todos las capas estudiadas. En el caso de no existir el título de la capa, no se puede considerar en la práctica que la capa ni siquiera exista. Respecto al resumen, éste aparece en un número similar al encontrado en los servicios.

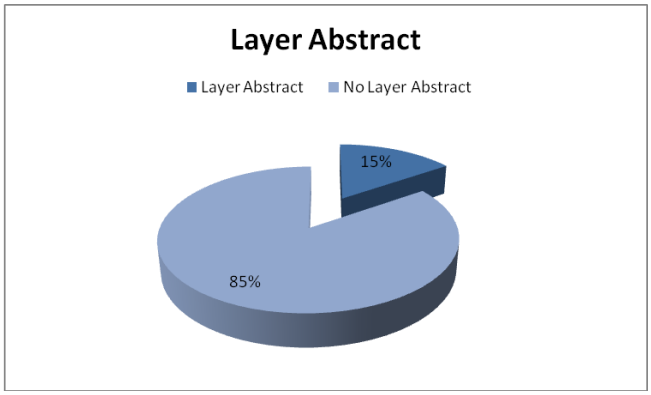
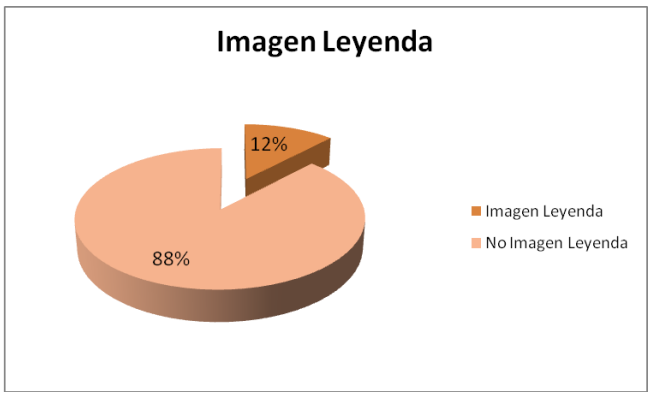


Figura 12: Aparición del resumen en los datos de las capas o layers.

El protocolo WMS permite añadir una referencia a una imagen que visualiza la simbología de la leyenda. Es una información que puede llegar a ser fundamental para interpretar los mapas servidos por el servicio. Además, permite enriquecer el interface de las aplicaciones cliente que se alimentan de un WMS. En el proyecto realizado, se han encontrado imágenes de leyenda en un 12% de los casos, mientras que en un 88% de los mismos no existía ninguna URL.



6 Conclusiones

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 como el presente.

Los metadatos intrínsecos al servicio pueden 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. Este hecho se ha cuantificado en el estudio de calidad, y refleja que más de un 25% de los servicios no contienen una información básica.

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.

Como en todas comunicaciones públicas del presente proyecto, 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.