(Draft) Actas de las III Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE'2012), Madrid, 17-19 de octubre de 2012. 2012

Análisis de la visibilidad global de los publicadores de los recursos geográficos estandarizados

Aneta J. Florczyk, F.Javier López-Pellicer, Juan Valiño-García, Javier Nogueras-Iso, F.Javier Zarazaga-Soria

Universidad de Zaragoza, España

{florczyk,lopez,juanv,jnog,javy}@unizar.es

Resumen

Este trabajo presenta el análisis de visibilidad de los publicadores de los recursos relacionados con IDEs Ibéricas. El análisis demuestra que la mitad de los sitios Web encontrados a partir de los servicios Web (OGC) están relacionados con las IDEs, y casi un tercio de estos dominios no está configurado adecuadamente. Por otro lado, se puede observar que la gran mayoría de las páginas no contiene metadato geográfico. Por lo tanto, los publicadores de recursos de las IDEs podrían mejorar su visibilidad si se siguen las recomendaciones correspondientes a la plataforma tecnológica que usan, es decir, las recomendaciones y buenas prácticas de la comunidad de la Web.

Palabras clave: Visibilidad, Publicadores, Recurso Web, Web Geoespacial, Metadato geográfico.

1 Introducción

La Web es un medio libre, tecnológicamente maduro y de extensión global para la comunicación entre los participantes, principalmente los publicadores y consumidores de información y recursos digitales. En la comunidad de las IDEs pronto se han visto las ventajas de esta capacidad, y actualmente la Web esta siendo usada como la plataforma base de la implementación tecnológica para las IDEs. La publicación de forma distribuida de información y de recursos de información geográfica para su reuso es una de las características de una IDE.

En los entornos distribuidos, como la Web y la propia IDE, los participantes deben tener un soporte de búsqueda de información y los recursos. Los principios de búsqueda de información o búsqueda de recursos en una IDE, vienen del mundo de las librerías digitales [1], y se caracteriza por un mecanismo de búsqueda especializado que típicamente cubre la extensión (digital) de la IDE. Este mecanismo es un elemento clave en los geoportales [2], que pueden ser entendidos como “el punto de acceso a una IDE” [3].

El uso de la plataforma Web por las IDEs convierte automáticamente sus contenidos en los recursos de la Web [4]. En el entorno Web los buscadores genéricos son un mecanismo que se adapta a las

características de la Web como la libertad y dinamismo, que en términos prácticos se traduce en creación, evolución y desaparición de recursos sin ningún control. Aunque las IDEs propiamente dichas son las iniciativas tipo “top-down”, su proceso de desarrollo parece compartir las características del entorno dinámico como la Web, tanto dentro de una IDE, como en la comunidad de las IDEs. Por lo tanto no debe extrañar la aparición de las propuestas a soporte a la búsqueda de los recursos especializados de una IDE (esto es, datasets y servicios) que vienen de la comunidad de la Web [5,6]

La dinámica del desarrollo de las IDEs muchas veces se debe a apariciones de nuevos geoportales que pueden ser de interés para los potenciales usuarios de recursos de información geográfica. Teniendo en cuenta la cobertura limitada de una IDE y la gran cantidad de herramientas conocidas por los usuarios Web en sus tareas de búsqueda, los usuarios pueden llegar a usar un motor de búsqueda genérico (o sus variantes especializados) para descubrir 'nuevos horizontes' en la Web global [3]. Además, la aparición y creciente oferta de aplicaciones Location Based Services, añade una dimensión más en la descripción del contenido. A parte del valor económico de acceso a información ofrecida por las IDEs a los usuarios, hay que tener en cuenta el valor económico indirecto para los actores de las IDEs (por ejemplo publicadores de los servicios o datos) como la publicidad. El mercado de servicios ajustados a las necesidades del cliente, la oferta educativa, o incluso el potencial de atraer nuevos socios pueden ser ejemplos de ello, lo cual no se limita al sector empresarial.

Por lo tanto los creadores de los geoportales y otros actores que participan en el proceso de publicación de contenidos geoespaciales (por ejemplo los publicadores) deberían cuidar su visibilidad en la Web. La publicación de los contenidos en la Web se basa en las recomendaciones que cubren aspectos tecnológicos (es decir, las recomendaciones de W3C[1]) y su incumplimiento no tiene ninguna consecuencia directa gracias a la permisividad de las herramientas de soporte (e.j. navegadores o motores de búsqueda). Sin embargo existen buenas prácticas de publicación proporcionadas por la comunidad de buscadores (por ejemplo, Google [7]) que permiten mejorar la visibilidad en un motor de búsqueda. Una de ellas es el uso de metadato.

Este trabajo se dedica a la evaluación de la visibilidad potencial de los portales de las IDEs. Se asume que los sitios Web que publican recursos relevantes para una IDE deben estar relacionados con ella. Por lo tanto, como base para el análisis se toma un conjunto de sitios Web que publican recursos relevantes para una IDE. La evaluación de la visibilidad se basa en evaluación de un metadato geográfico que describe las páginas de estos sitios Web. El proceso de creación de este metadato es un proceso automático que sigue las principales recomendaciones de la comunidad Web.

2 Análisis de la visibilidad

1http://www.w3c.es (last accessed 26/11/2012)

La visibilidad potencial de una página Web se puede analizar siguiendo las principales recomendaciones de la comunidad Web respecto a los metadatos[2] [3] [4]. Algunos de estos elementos del metadato no son usados por los buscadores (e.j. Google no usa Keywords[5]), pero existen otros elementos en la estructura de una página Web que usan los buscadores[6]. Por lo tanto, se evalúan los metadatos asociados a las páginas Web publicados por los portales de las IDEs.

Primero, hay que crear de manera automática una muestra de los sitios Web de interés para su análisis. Se puede asumir que una página Web a la que se accede a través del dominio extraído de un recurso Web (por ejemplo un servicio) debe pertenecer a la vez al sitio Web que publica este recurso. Si partimos de los recursos relevantes para las IDEs, se puede asumir que los sitios Web identificados de esta manera deben ser vinculados a una IDE. En este trabajo, los tipos de recursos de interés son los servicios que siguen las especificaciones OGC, porque, primero, estos estándares son los estándares tecnológicos recomendados por las IDEs y, segundo, existen herramientas que permiten una recogida de los recursos Web de manera automática, por ejemplo un robot especializado [5].

2.1 Creación y análisis de corpus

El listado de dominios ha sido extraído (sin repeticiones) de los servicios Web OGC (OWS) recogido por un robot especializado descrito en [8] durante un año a partir de Julio 2011. A propósito de este trabajo, el robot ha sido configurado para la identificación de servicios relevantes para la comunidad de las IDEs ibéricas (es decir, de Portugal, Andorra y España). Como resultado, se han recogido 50.537 entradas, de las cuales hay 6.899 servicios únicos y 259 dominios únicos.

En total, 192 dominios únicos han sido seleccionados de manera aleatoria y los documentos Web (accedidos desde estos dominios) han sido analizados manualmente en Julio 2012. Durante ella análisis, se ha extraído la siguiente información: el estatus ('OK', 'ERROR'), el tipo de página ('geoportal', 'portal', 'visor', 'otro', 'error'), cobertura espacial (a base de contenido publicado en la página) y lenguaje del texto.

En general, no ha sido posible acceder a 3.7% de los sitios Web ('ERROR') debido a un error del navegador. Las páginas con el estatus 'OK' han sido analizadas con más detalle. Un 25.5% de los dominios analizados han proporcionado las páginas clasificadas como 'geoportal', un 11.5% como 'portal', y un 8.9% como 'visor'. El tipo 'otro' ha sido asignado a más de la mitad de los dominios analizados (50.5%), de los cuales 22.2% (del total) son las respuestas erróneas (por ejemplo, respuestas vacías, el listado de los ficheros, información del error del servicio o del servidor), un 10.9% (del total) son las páginas generadas por defecto por el servidor/servicio (por ejemplo, página de inicio del Tomcat), un 2.1% (del total) son páginas de compañías dedicadas a software, un 10.9%

2 http://geourl.org/add.html (last accessed 26/11/2012)

3 http://www.metatags.org/all_metatags (last accessed 26/11/2012)

4 http://dublincore.org/documents/2008/08/04/dc-html/ (last accessed 26/11/2012)

5 http://googlewebmastercentral.blogspot.com.es/2009/09/google-does-not-use-keywords-meta-tag.html (last accessed 26/11/2012)

6 http://support.google.com/webmasters/bin/answer.py?hl=en&answer=35624&topic=2371375&ctx=topic (last accessed 26/11/2012)

(del total) son las páginas repetidas, y el resto (3.7% del total) son las páginas generadas por servidor de manera automática pero su contenido sugiere que están vinculadas a una IDE.

En general, el análisis manual muestra que el 49.5% de los dominios analizados devuelve contenido relevante para las IDEs, teniendo en cuenta las páginas clasificadas como 'geoportal' (25.5%), 'portal' (11.5%), 'visor' (8.9%), y la parte relevante en este contexto de las clasificadas como 'otro' (3.7%).

2.2 Generacion del metadato geografico

En este trabajo se toman las recomendaciones de la comunidad Web como punto de partida para creación de una herramienta capaz de generar automáticamente un metadato geográfico mediante las heurísticas [9]. El modelo del metadato generado sigue un modelo mínimo recomendado por especificación de la CSW [10]. Esta herramienta extrae los metadatos encapsulados en las páginas Web, y también aplica algunas de las técnicas típicas para este tipo de herramientas. Pero su ventaja principal es la capacidad de estimar la extensión geográfica cuando esta información no es proporcionada por el metadato de la página. Su funcionalidad se limita a las páginas Web que siguen una especificación HTML 4[7], por lo tanto el uso de anotaciones semánticas dentro del cuerpo de una página, no se tiene en cuenta. Más detalles sobre esta herramienta se puede encontrar en [9].

2.3 Análisis y discusión

Primero de todo se puede observar que casi la mitad de los sitios Web encontrados a partir de los servicios OWS están relacionados con las IDEs. Por otro lado, se puede apreciar que hay un 33.2% del total de los dominios cuyo URL genera una respuesta errónea o respuesta 'por defecto' cual no se puede vincular con un sitio Web concreto. Eso sugiere que una gran parte de los dominios usados por los publicadores de los contenidos geoespaciales no está configura adecuadamente, por ejemplo, para redireccionar a un sitio Web.

El análisis de los resultados de generación del metadato muestra que el 88% de las páginas para cuales se han generado los metadatos contienen el conjunto básico (es decir, por lo menos título o descripción). Por otro lado, se puede ver que la ausencia del metadato geográfico es bastante frecuente. Solo el 3.16% del total de las páginas procesadas (es decir, tres páginas) proporcionan un metadato geográfico. Aunque la herramienta es capaz de generar un metadato de cobertura a partir del contenido publicado por la página, hay que tener en cuenta que solo genera la información, igual que la estimada manualmente, en el 20% de los casos. Si se asume que la información geográfica esta aceptable cuando al menos se ha identificado adecuadamente el país (el nivel nacional es aceptable), la herramienta genera los resultados aceptables en el 69.5% de los casos (incluyendo los 20% de los 'iguales'). Por lo tanto, es importante ver que el uso de metadatos es vital para eliminar la incertidumbre de una estimación basada en heurísticas.

Un ejemplo de buenas prácticas respecto a los metadatos de las páginas Web es IDERRIOJA[8], el geoportal del Gobierno de La Rioja. El siguiente listado presenta el contenido del elemento 'head' que corresponde al metadato proporcionado por la página de este portal.

---

7 http://www.w3.org/TR/REC-html40/ (last accessed 26/11/2012)

8 http://www.iderioja.larioja.org/ (last accessed 26/11/2012)

```
<head>
  <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
   <meta name="keywords" content="IDERIOJA, Gobierno de La Rioja, Cartografia,
Rioja, IDE, Infraestructura Datos Espaciales, Mapas, Visor, Geografia, Territorio,
Ortofoto, CAD, Metadatos, Saicar" />
   <meta name="description" content="Infraestructura de Datos Espaciales de La
Rioja." />
  <meta name="title" content="Infraestructura de Datos Espaciales - Gobierno de La
Rioja - IDERIOJA" />
<!-- (…) -->
  <meta name="geo.position" content="42.27189379;-2.28201889" />
  <meta name="geo.placename" content="La Rioja, Logroño" />
  <meta name="geo.region" content="ES-LO" />
  <!-- Metadatos del Dublin Core -->
  <link rel="meta" href="http://www.w3c.es/index.rdf" />
  <link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
  <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
  <meta name="DC.title" xml:lang="es" content="Infraestructura de Datos Espaciales
- Gobierno de La Rioja - IDERIOJA" />
   <meta name="DC.subject" content="IDERIOJA, Gobierno de La Rioja, Cartografia,
Rioja, IDE, Infraestructura Datos Espaciales, Mapas, Visor, Geografia, Territorio,
Ortofoto, CAD, Metadatos, Saicar" />
    <meta name="DC.description" xml:lang="es" content="Infraestructura de Datos
Espaciales de La Rioja." />
  <meta name="DC.publisher" content="Gobierno de La Rioja" />
  <meta name="DC.date" scheme="DCTERMS.W3CDTF" content="2006-03-10" />
  <meta name="DC.type" scheme="DCTERMS.DCMIType" content="Text" />
  <meta name="DC.creator" content="Agencia del Conocimiento y la Tecnologia" />
  <meta name="DC.format" scheme="DCTERMS.IMT" content="application/xhtml+xml" />
  <meta name="DC.identifier" scheme="DCTERMS.URI" content="http://www.w3c.es" />
  <meta name="DC.language" scheme="DCTERMS.RFC1766" content="es" />
  <meta name="DC.rights" content="http://www.w3.org/Consortium/Legal/" />
  <!-- Fin de los metadatos del DC -->
    <title>Infraestructura de Datos Espaciales - Gobierno de La Rioja -
IDERIOJA</title>
<!-- (…) -->
</head>
```

Basándose en este análisis, es interesante observar que aunque la comunidad de las IDEs se caracteriza por un fuerte reconocimiento de valor del metadato, esta característica no se refleja en los portales Web de la propia comunidad.

3 Conclusiones y trabajo futuro

Este trabajo presenta el análisis de visibilidad de los publicadores de los recursos relacionados con IDEs Ibéricas. Primero de todo se puede observar que casi la mitad de los sitios Web encontrados a partir de los servicios OWS están relacionados con las IDEs, y casi un tercio de estos dominios no está configurado adecuadamente. Por otro lado, se puede observar que la gran mayoría de las páginas tratadas por la herramienta de extracción de metadato no contiene metadato geográfico. Resumiendo, los publicadores de recursos de las IDEs podrían mejorar su visibilidad si se siguen las recomendaciones correspondientes a la plataforma tecnológica que usan, es decir, las

recomendaciones y buenas prácticas de la comunidad de la Web. Aunque el uso de los visores es muy común en el contexto de publicación de información geográfica, metadatos sencillos permiten caracterizar el recurso para su indexación adecuada. Es cierto que los buscadores de hoy en día son muy sofisticados y capaces de mitigar la ausencia del metadato, pero un metadato generado por un proceso automático que usa información contextual está asociado a un cierto grado de incertidumbre. Un metadato básico creado por el publicador es siempre la mejor opción.

En el futuro, el trabajo se centrará en el análisis de la dinámica de desarrollo de las IDEs. Uno de los aspectos que lo reflejan es el comportamiento de los recursos publicados por la comunidad de las IDEs. En este contexto, el seguimiento de la evolución de un recurso o incluso la identificación de sus duplicados son algunas de las líneas de investigación abiertas. En este trabajo los duplicados, por ejemplo, se eliminan gracias a un análisis manual, pero para el análisis de cobertura global se debe disponer de aproximaciones automáticas.

Por otro lado, la cuestión de caracterización automática de un sitio Web (e.j. portal, geoportal, página de un servidor, página de una entidad como empresa o centro de investigación) a partir de un dominio y el contenido publicado se debe investigar. En este trabajo, el análisis manual ha proporcionado esta información.

4 Agradecimientos

El trabajo de Aneta Jadwiga Florczyk ha sido cofinanciado por el Ministerio de Educación a través de la beca AP2007-03275. Este trabajo ha sido parcialmente financiado por el Gobierno de España a través del proyecto TIN2009-10971, el Instituto Geográfico Nacional (IGN), GeoSpatiumLab S.L., el Gobierno de Aragón y el Fondo Social Europeo.

5 Referencias bibliografiítas

[1] Béjar, R., Nogueras-Iso, J., Latre, M.A., Muro-Medrano, P. R. and F. J. Zarazaga-Soria (2009). “Digital Libraries as a Foundation of Spatial Data Infrastructures”, *Handbook of Research on Digital Libraries: Design, Development, and Impact*. IGI Global, pp. 382-389.

[2] Percivall, G. (2002). OpenGIS Service Architecture (Version 4.3). The OpenGIS Abstract Specification and ISO/DIS 19119. Geographic information — Services.

[3] Florczyk, A.J. (2012). Search Improvement within the Geospatial Web in the context of Spatial Data Infrastructures. Ph.D thesis, Universidad de Zaragoza.

[4] López-Pellicer, Béjar, R. and F.J. Zarazaga-Soria (2011). Providing semantic links to the Invisible Geospatial Web. Zaragoza: Universidad de Zaragoza, 2012. Notes in Geoinforrnatics Research/Cuadernos de Investigación en Geoinforrnática.

[5] López-Pellicer, F.J., Florczyk, A.J., Béjar, R., Muro-Medrano, P.R. and F.J. Zarazaga-Soria (2011). Discovering geographic web services in search engines, *Online Information Review*, 35(6):909-927.

[6] Li, W., Yang, C., and C. Yang (2010). 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 24 (8), 1127–1147.

[7] Google (2010). Google Search Engine Optimization Starter Guide, Online, URL http://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf (last accessed 26/11/2012).

[8] López-Pellicer, F. J., Florczyk, A. J., Béjar, R., Muro-Medrano, P. R. and F.J. Zarazaga-Soria (2011). Discovering geographic web services in search engines *Online Information Review, 35*, 909-927.

[9] Florczyk, A.J., López-Pellicer, F.J., Béjar, R., Nogueras-Iso, J. and F.J. Zarazaga-Soria (2011). Automatic Generation of Geospatial Metadata for Web Resources, *IJSDIR*, 7:152-172.

[10] Nebert, D., Whiteside, A. and P. Vretanos (eds.) (2007). OpenGIS Catalogue Services Specification. OpenGIS Implementation Specification. Version 2.0.2, Corrigendum 2 Release. OGC 07-006r1. Open Geospatial Consortium Inc.