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

AnetaJ. Florczyk, F.Javier López-Pellicer, Javier Nogueras-Iso, F.Javier Zarazaga-Soria
Universidad de Zaragoza, España

{florczyk,lopez,jnog,javy}@correo.es

Resumen

Este trabajo presenta el análisis de visibilidad de los publicadores de los recursos relacionados con IDEs Ibéricas. La análisis demuestra que mayoría de los sitios Web encontrados a partir de los servicios Web (OGC) están relacionados con las IDEs, pero casi un tercio de estos dominios, no esta configurado para adecuadamente. Por otro lado, se puede observar que 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, Metadato geográficos, Proveedores, Recursos Web.

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 portales, que pueden ser entendidos como; “el punto de acceso a una IDE” [2].

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 (i.e. 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 portales 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, no es de extrañar que se pueda llegar a usar un motor de búsqueda genérico (o sus variantes especializados) para descubrir 'nuevos horizontes' en la Web global [2].

Este hecho tiene repercusiones para los creadores de los portales. 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 violación no tiene ninguna repercusión directa gracias a la

[1] http://www.w3.org/

permisividad de las herramientas de soporte (e.g. browsers o motores de búsqueda). Pero existen buenas prácticas de publicación proporcionadas por la comunidad de buscadores (por ejemplo, Google) que permiten mejorar la visibilidad en un motor de búsqueda. Una de ellas es el uso de metadato. Además, la aparición y creciente oferta de aplicaciones Location Based Services, añade una dimensión más en la descripción del contenido.

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. El método se basa en creación automática (siguiendo principales recomendaciones de la comunidad Web) y evaluación de un metadato geográfico, que describe las paginas de estos sitios Web.

2 Análisis de la visibilidad

La visibilidad potencial de una pagina Web, siguiendo las principales recomendaciones de la comunidad Web respecto a los metadatos[2] [3] [4] [5]. Algunos de estos elementos del metadato no están usados por los buscadores (e.g. Google no usa Keywords[6]), pero existen otros elementos en la estructura de una pagina Web que usan los buscadores[7]. Por lo tanto, se evalúa 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

---

[2] http://geourl.org/add.html

[3] http://geourl.org/add.html

[4] http://www.metatags.org/all_metatags

[5] http://dublincore.org/documents/2008/08/04/dc-html/

[6] http://googlewebmastercentral.blogspot.com.es/2009/09/google-does-not-use-keywords-meta-tag.html

[7] http://support.google.com/webmasters/bin/answer.py?hl=en&answer=35624&topic=2371375&ctx=topic

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 del corpus

El listado de dominios ha sido extraído (sin repeticiones) de los servicios Web OGC (OWS), recogido por un robot especializado descrito en López-Pellicer et al (2011) durante un año a partir del Julio 2011. A propósito de este trabajo, el robot ha sido configurado para 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.

Los sitios Web (accedidos desde los dominios) se han analizado manualmente en Julio 2012. Durante la análisis, se ha extraído la siguiente información: el estatus ('OK', 'ERROR'), el tipo de página ('geoportal', 'portal', 'visor' and 'otro'), cobertura espacial (a base de contenido publicado en la pagina) y lenguaje del texto. Las páginas del tipo 'otro' (42.9% del total) están rechazados de la posterior evaluación. La mayoría de estas paginas son las paginas devueltas por el sistema por defecto (67.1%). El resto son las páginas repetidas (21.2%), paginas a las que no se podía acceder (10.6%) y algunas de compañías dedicadas a software. El 47.98% de las páginas encontradas son de tipo 'geoportal' (55.9%), 'portal' (27.4%) o 'visor' (16.8%).

En general, el análisis manual muestra que el 53.5% de las páginas encontradas son relevantes para las IDEs Ibéricos (incluyendo 12.9% de las paginas tipo 'otro' que han sido generadas automáticamente por un servidor).

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 [3]. El modelo del metadato generado sigue un modelo recomendado por especificación de la CSW. Esta herramienta extrae los metadatos encapsulados en las paginas 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[8], por lo tanto el uso de anotaciones semánticas dentro del cuerpo de una página, no se tiene en cuenta. Mas detalles sobre esta herramienta se puede encontrar en [3].

2.3 Analisis y discusion

Primero del todo se puede observar que mayoría 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 28.79% de total de las paginas (es decir, el 67% de los 85 paginas del tipo 'otro'). Eso significa que, casi un tercio de estos dominios, no esta configurado para por ejemplo redireccionar a un sitio Web relacionado.

El análisis de los resultados de generación del metadato muestra que el 88% de las paginas, para cuales se han generado los metadatos contienen el conjunto básico (es decir, por lo menos titulo o descripción). Por otro lado, se puede ver que la falta del metadato geográfico es bastante frecuente. Solo 3.16% del total de las paginas procesadas (es decir, tres paginas) proporcionan un metadato geográfico. Aunque la herramienta es capaz de generar un metadato de cobertura a partir del contenido publicado por la pagina, hay que tener en cuenta que solo genera la información, igual que la estimada manualmente, en 21.1% de los casos. Si se asume que la información geográfica esta aceptable si por lo menos se ha identificado adecuadamente el país (el nivel nacional es aceptable), la herramienta genera los resultados aceptables en 71% de los casos (incluyendo los 21.1% de los 'iguales'). Por lo tanto, es importante ver que

[8] http://www.w3.org/TR/html4/

el uso de metadatos es vital para eliminar la incertidumbre de una estimación basada en heurísticas.

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 del todo se puede observar que mayoría de los sitios Web encontrados a partir de los servicios OWS están relacionados con las IDEs, pero casi un tercio de estos dominios, no esta configurado para adecuadamente. Por otro lado, se puede observar que mayoría de las paginas 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, simples metadatos permiten caracterizar el recurso para su indexación adecuada. Es cierto que los buscadores de hoy en día son muy sofisticados y son capaces de mitigar la falta del metadato, pero un metadato, generado por un proceso automático que usa información contextual, esta asociado a un cierto grado de incertidumbre. Un simple metadato creado por un proveedor es siempre 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 identificación de sus duplicados son unas de las líneas de investigación abiertas. En este trabajo los duplicados, por ejemplo, se eliminan gracias a al 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 (p.ej. portal, geoportal, página de un servidor, página de una entidad como empresa o centro de investigación) a partir de un dominio se debe

investigar. En este trabajo, el análisis manual ha proporcionado esta información.

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

[3] 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.

[4] López-Pellicer, F.J., (2011), Semantic Linkage of the Geospatial Web. Ph.D thesis, Universidad de Zaragoza.

[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., Yang, C., 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] 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.