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

Hacia las Infraestructuras de Datos Abiertos Espaciales

F.J. Zarazaga-Soria[1], R. García[2], F.J. Lopez-Pellicer[1], J. Nogueras-Iso[1], R. Béjar[1], R.Gil[2], J.M. Brunetti[2], J.M.Gimeno[2], P.R. Muro-Medrano[1]

[1] Universidad de Zaragoza
[2] Universitat de Lleida

javy@unizar.es, roberto.garcia@udl.cat, {fjlopez,jnog,rbejar}@unizar.es, {rgil,josepmbrunetti,jmgimeno}@diei.udl.cat, prmuro@unizar.es

Resumen

Los Datos Abiertos están impulsando la aparición de un nuevo tipo de sistema de sistemas que se puede denominar Infraestructura de Datos Abiertos que plantean problemas similares a los que nos enfrentamos cuando aparecieron las Infraestructuras de Datos Espaciales. Dado que gran parte de la información que se está publicando como Dato Abierto tiene una naturaleza espacial surge la necesidad de una solución híbrida compatible con ambas visiones: la Infraestructura de Datos Abiertos Espaciales. Esta comunicación presenta un proyecto de investigación que está trabajando en dicho concepto y en su sostenibilidad. Presentamos así mismo una serie de líneas de investigación que tratan de cubrir diversos aspectos de la cadena de valor del dato abierto con naturaleza espacial que incluyen el enriquecimiento de la representación de los aspectos geoespaciales y su utilización en la recuperación de información, el descubrimiento automático de recursos en la Web y su enlazado con la Web Semántica mediante su caracterización y etiquetado automático.

Palabras clave: Datos Abiertos, Web Semántica, Rastreadores de Información, Generación automática de interfaces, Web 2.0

1 Introducción

Los gobiernos públicos, a través de sus administraciones, generan, recogen, sufragan o poseen un ingente patrimonio de información que incluye información geográfica, medioambiental, social, económica, turística, estadística, meteorológica, datos de empresas, patentes y educación, datos procedentes de proyectos de investigación financiados con fondos públicos, materiales de archivo y libros digitalizados de las bibliotecas. La información del sector público es una materia prima importante para diversos productos y servicios con contenido digital, y también puede ser utilizada para mejorar la eficiencia de las administraciones. De entre todos los tipos de informaciones destaca la información geoespacial por su alto porcentaje de penetración en todo el ámbito de la administración pública, su elevado coste de elaboración y las posibilidades de reutilización.

La Unión Europea ha dado recientemente un impulso político y legislativo a su estrategia para facilitar el acceso abierto y en formato digital a los datos y documentos en poder de la administración, salvo las excepciones que la regulación marque, para su reutilización, puesta en valor y uso como herramienta de ayuda a la transparencia de las administraciones públicas europeas.

El propósito de esta comunicación es la caracterización de algunas de las líneas de investigación que estarían relacionadas con el desarrollo sostenible de *Infraestructuras de Datos Abiertos Espaciales* (IDAE) para facilitar la reutilización de información del sector público con algún rasgo espacial. Estas líneas mejorarían diversos aspectos de la cadena de valor asociada a la publicación de información vía el enriquecimiento de la representación de los aspectos geoespaciales, su utilización en la recuperación de información, el descubrimiento automático de recursos en la Web, y su enlazado con la Web Semántica mediante su caracterización y etiquetado automático.

2 Las Infraestructuras de Datos Abiertos Espaciales

La reciente comunicación de la CE titulada "*Datos abiertos. Un motor para la innovación, el crecimiento y la gobernanza transparente*" [1] explicita el actual rumbo de la política legislativa de la CE en relación con los datos del sector público. Esta estrategia es continuación de la Directiva 2003/98/CE

relativa a la Reutilización de las Información del Sector Público (RISP) [2] traspuesta en 2005.

El marco legal en formación así como las iniciativas particulares relacionadas con la publicación de Datos Abiertos están dando naturaleza a un nuevo tipo de sistema de sistemas (en el sentido recogido en [3]): la infraestructura de datos abiertos. Esta infraestructura englobaría la comunidad, las tecnologías, los estándares abiertos y las licencias libres que en conjunto apoyan y promueven la distribución, el uso y re-uso de los datos públicos utilizando sistemas de información heterogéneos, en constante evolución, gestionados por diferentes organizaciones con diferentes objetivos, y distribuidos por todo el mundo

La IDE podría ser el antecesor directo de dicha infraestructura. Son parecidas, pero se diferencian en el dominio y en los estándares de referencia. En una IDE el dominio lo delimita la naturaleza espacial de la información. El dominio de las infraestructuras de Datos Abiertos lo delimita la naturaleza pública de la información y su reutilización. Una IDE se basa en estándares desarrollados dentro del dominio espacial por organizaciones internacionales como OGC y el comité ISO/TC 211. Las infraestructuras de Datos Abiertos, por el contrario, toman como referencia estándares, recomendaciones y buenas prácticas desarrollados dentro del W3C. Un buen ejemplo es [4] que nos proporciona una guía para la publicación de información pública en el marco de los estándares de W3C.

Cuando en la información publicada bajo el paraguas de las buenas prácticas de Datos Abiertos predomina el carácter espacial y por algún motivo, por ejemplo legal, hay que dar soporte a estándares geográficos surge la necesidad de establecer una solución híbrida que podemos denominar como *Infraestructura de Datos Abiertos Espaciales* (IDAE).

2 Sostenibilidad de una IDAE: bajo coste, alta automatización y alta participación

Para que la estrategia de publicación descrita resulte sostenible es importante que la apertura del acceso a la información pública, en particular aquella con algún carácter espacial, no implique un coste añadido significativo a las administraciones públicas involucradas. La sostenibilidad económica pasa a ser un factor clave. La automatización del proceso de publica-

ción es una de las claves para lograr la sostenibilidad económica y cuando que su presupuesto esté bajo control. La automatización implica nuevas necesidades y por tanto nuevas oportunidades relacionadas con el desarrollo de las correspondientes infraestructuras de información.

Un factor de sostenibilidad adicional es el relacionado con la ciudadanía. Las administraciones públicas necesitan información sobre cómo impacta en la sociedad la apertura de datos para justificar el mantenimiento de la infraestructura como servicio público. Este impacto puede ser directo sobre el ciudadano o indirecto vía los consumidores masivos de datos que le proporcionan servicios. Por ello la participación ciudadana, entendida como los patrones de uso de los datos, los comentarios sobre su calidad, y otros aspectos que ahora relacionamos con la Web 2.0, han de incorporarse en el proceso de publicación automatizada.

3 El papel de la investigación

La investigación en Datos Abiertos debe hacer viable en el tiempo las iniciativas de publicación. Un ejemplo sería investigar qué características deberían tener las plataformas de publicación para ser herramientas efectivas que mejoran el acceso y el uso de los datos geoespaciales. Un programa de investigación para las IDAEs debe enfocarse en lograr avances concretos en diferentes campos. Por ejemplo, debe lograr avances tecnológicos en componentes y servicios basados en los conceptos de la Web Semántica para que puedan ponerse en un entorno de producción para reducir el coste de la automatización. Otro ejemplo es la investigación sobre las sinergias entre los rasgos específicos de la información pública con rasgos espaciales y los interfaces de usuario para mejorar la experiencia de usuario y la comprensión de la información.

4 Un plan de investigación en IDAEs

La presente comunicación se focaliza sólo en la automatización de algunos aspectos. El primero es el descubrimiento automático de nuevos recursos públicos disponibles en la red. También hay que crear una base de conocimientos espacial que ayude a la extracción de información enriqueciendo sus aspectos geoespaciales, su enlazado con la Web Semántica mediante etiquetado automático. Para el caso en que tenga que haber intervención humana, hacer más económico el trabajo facilitando la generación (se-

mi)automática de interfaces de usuarios y visualizaciones a partir de los propios datos abiertos. Finalmente, hay que mejorar la usabilidad y accesibilidad para hacer atractiva la participación ciudadana. A continuación se introducen, con un poco más detalle, las líneas de trabajo.

1. **Descubrimiento automático de recursos.** Actualmente se puede indexar la web para localizar información relacionada con un dominio [5]. Esta búsqueda puede estar enfocada a la información del sector público. Se requiere mejorar la capacidad del rastreo con el objetivo de aumentar los tipos de recursos que se accede. Esto requiere la especificación de nuevos tipos de recursos, así como la identificación de los patrones que caracterizan el acceso a los recursos, y el desarrollo de componentes tecnológicos que permitan su incorporación a rastreadores existentes.

2. **Base de conocimiento espacial para enriquecimiento**. Toda la información pública con propiedades espaciales ha sido descubierta por los rastreadores hay que alinearla con una base de conocimiento espacial común para poder ofrecer al consumidor del dato seguridad y calidad respecto a la parte espacial [6]. Esto requiere de una base de conocimiento espacial en constante evolución a través de la inclusión de nuevos sistemas de referencia espacial basada en identificadores geográficos y por la mejora de sus mecanismos de razonamiento. Pero también requiere mejoras en los algoritmos de alineamiento para tratar el ruido en los datos descubiertos y en los sistemas de referencia, así como mecanismos para enlazarla con la Web Semántica.

3. **Generación de interfaces de usuario.** Existen trabajos como [7] sobre la generación automática de interfaces de usuario para la exploración de datos abiertos, la mejora de la flexibilidad y la independencia de la estructura de los datos. Pero para que el costo de la implementación, incluso para grandes conjuntos de datos, fuera manejable habría que explorar los patrones de interacción y visualización para facilitar el uso de datos abiertos y su valorización, poniendo especial interés en su dimensión espacial.

4. **Participación ciudadana**. Los ciudadanos han manifestado su interés en compartir su conocimiento espacial sobre el territorio vía Web, por ejemplo en Wikipedia [8]. Una forma de contribuir a la sostenibilidad de los datos publicados abiertos, y mejorar su calidad, es la investigación

en métodos y técnicas que permitan aplicar ese conocimiento a los datos públicos de naturaleza espacial.

Conclusiones

El propósito de esta comunicación es la presentación de un proyecto de investigación que está trabajando en el desarrollo sostenible de infraestructuras de datos abiertos de carácter geoespacial. Las líneas de investigación presentadas tratan de cubrir diversos aspectos de la cadena de valor que incluyen el enriquecimiento de la representación de los aspectos geoespaciales y su utilización en la recuperación de información, el descubrimiento automático de recursos en la Web y su enlazado con la Web Semántica mediante su caracterización y etiquetado automático.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Gobierno de España a través del proyecto TIN2009-10971; del Instituto Geográfico Nacional (IGN) y de GeoSpatiumLab S.L.

Referencias

[1] Comisión Europea, “Propuesta de Directiva del Parlamento Europeo y del Consejo por la que se modifica la Directiva 2003/98/CE relativa a la reutilización de la información del sector público,” Comunicación de la Comisión al Parlamento Europeo, al Consejo al Comité Económico y social Europeo y al Comité de las Regiones., COM(2011) 877 final, Dec. 2011.

[2] Comisión Europea, “Directiva 2003/98/CE del Parlamento Europeo y del Consejo de 17 de noviembre de 2003 relativa a la reutilización de la información del sector público,” Diario Oficial de la Unión Europea, L 345/90 ES, Dec. 2003.

[3] M. W. Maier, “Architecting Principles for Systems-of-Systems,” *Systems Engineering*, pp. 267–284, 1998.

[4] D. Bennet and A. Harvey, “Publishing Open Government Data,” W3C, Sep. 2009.

[5] S. Chakrabarti, M. Van den Berg, and B. Dom, “Focused crawling: a new approach to topic-specific Web resource discovery,” *Computer Networks*, vol. 31, no. 11, pp. 1623–1640, 1999.

[6] N. Cardoso, M. J. Silva, and D. Santos, “Handling implicit geographic evidence for geographic IR,” presented at the CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, 2008.

[7] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon, “ManyEyes: a Site for Visualization at Internet Scale,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1121–1128, 2007.

[8] D. Hardy, J. Frew, and M. F. Goodchild, “Volunteered geographic information production as a spatial process,” *International Journal of Geographical Information Science*, vol. 26, no. 7, pp. 1191–1212, Jul. 2012.