Servicio de nomenclátor utilizando el motor de búsqueda Solr. Caso práctico en la provincia de Lugo.

Diego Alberto Arias Prado[1], Manuel Pérez Gómez[1], Francisco Velayos Pardo[2], Rafael Crecente Maseda[1]

[1] Laboratorio do Territorio, G.I.-1934-TB, Universidad de Santiago de Compostela
[2] Diputación Provincial de Lugo

diegoalberto.arias@usc.es
manuel.perez.gomez@usc.es
f.velayos@deputacionlugo.org
rafael.crecente@usc.es

En esta comunicación se presenta un servicio de nomenclátor de Lugo, que geo-localiza las entidades singulares de población reconocidas por el Instituto Nacional de Estadística. Dicho servicio se implementa sobre la IDE de la provincia de Lugo y utiliza los datos de la Encuesta de Infraestructura y Equipamientos Locales. La alta dispersión poblacional de Lugo configura un modelo de asentamientos rurales único y rico en topónimos. Es necesario disponer de un servicio de este tipo para identificar y geolocalizar dichos lugares en el territorio.

Desde el punto de vista técnico, la principal novedad introducida es que el servicio utiliza el motor de búsqueda Solr para realizar las búsquedas, en vez de optar por la solución tradicional, que se apoya en bases de datos relacionales.

Palabras clave: Nomenclátor, Encuesta de Infraestructura y Equipamientos Locales, Solr, motor de búsqueda.

1 Introducción

La Encuesta de Infraestructura y Equipamientos Locales (EIEL) es una herramienta creada por el Ministerio de Administraciones Públicas (MAP) – actualmente Ministerio de Hacienda y Administraciones Públicas - cuyo objetivo es crear un inventario de ámbito nacional, con información sistematizada y precisa de las infraestructuras y equipamientos de competencia municipal pertenecientes a todos los municipios cuya población sea inferior a 50.000 habitantes.

La elaboración y actualización de la EIEL es responsabilidad de las entidades locales - Diputaciones provinciales, cabildos, etc. -, correspondiendo al Ministerio el seguimiento de las tareas y el colaborar económicamente con éstas (real decreto 835/2003 [1] y orden ministerial APU/293/2006 [2]). La primera fase de la EIEL data del año 1985, siendo la actualización quinquenal hasta el año 2005 y anual desde el año 2008.

Actualmente la información recopilada por la EIEL se divide en 6 grandes bloques: información general (datos demográficos, planeamiento urbanístico y núcleos abandonados), equipamientos municipales, abastecimiento de aguas, saneamientos de aguas, red viaria y, por último, energía, comunicaciones y recogida de residuos

sólidos urbanos. La EIEL es un instrumento fundamental para que la administración central disponga de información que permita detectar deficiencias en infraestructuras y equipamientos y, así, poder asignar recursos económicos a las distintas entidades de forma equitativa.

Inicialmente la información recogida por la EIEL era de naturaleza alfanumérica. Está previsto que a partir de la edición del año 2011, la EIEL recoja también información geo-referenciada. A pesar de la falta de obligatoriedad de recopilar información geo-referenciada, desde hace años son numerosas las encuestas provinciales que recogen información geo-referenciada que, además, se pone a disposición del público en general mediante servicios como *Web Map Service* (WMS) o *Web Feature Service* (WFS), estándares del Open Geospatial Consortium (OGC).

En el caso de la provincia de Lugo, la Diputación provincial viene actualizando la EIEL desde el año 2000 en colaboración del Laboratorio do Territorio (LaboraTe) [3], grupo de investigación de la Universidad de Santiago de Compostela. El LaboraTe ha introducido de forma continua mejoras en los procedimientos de recogida, procesamiento y publicación de la información. En el año 2008 se puso a disposición del público en general toda la información alfanumérica de la EIEL, mediante una página web [4] desde la cual puede ser descargada. En la edición del año 2009 se añadió una Infraestructura de Datos Espaciales (IDE) mediante la cual se puso a disposición del público en general la información geo-referenciada, vía servicios WMS y WFS, junto con un visor de mapas. En la edición del año 2010 se añadió un catálogo de metadatos en formato ISO 19115, perfil Núcleo Español de Metadatos [5].

En paralelo, en el año 2007 entraba en vigor la directiva INSPIRE (20047/2/CE) [6] en los estados miembros de la Unión Europea (UE). Los objetivos de esta directiva son garantizar la compatibilidad e interoperabilidad de las Infraestructuras de Datos Espaciales (IDE) existentes en los distintos estados miembros de la UE, obligando, para ello, a la adopción de un conjunto de reglas [7] que afectan a áreas como especificaciones de datos, compartición de éstos y metadatos o servicios suministrados por red y su monitorización. En España, la transposición de esta directiva vino de la mano de la Ley sobre las Infraestructuras y los Servicios de Información Geográfica en España (LISIGE) [8], del año 2010.

Con motivo de la edición de la EIEL del año 2011, el LaboraTe ha añadido un nuevo servicio de nomenclátor, a mayores de los ya ofrecidos. Resulta de suma importancia el disponer para la provincia de Lugo de una herramienta que identifique y geo-localice cada una de las más de 9.900 entidades singulares de población - reconocidas por el Instituto Nacional de Estadística - en las cuales se asienta la población. Galicia, que cuenta con el 5,8% y el 5,9%, respectivamente, de la superficie y población de España y presenta la singularidad de que en ella se asientan el 49% de las entidades singulares de toda España. En el caso de la provincia de Lugo - 1,95% y 0,75%, respectivamente, de la superficie y población de España -, siendo un porcentaje de entidades elevado del 15%, merced a más de 9.900 entidades singulares. Además, de dichas entidades singulares, el 87% tienen carácter diseminado. Una de la consecuencias de esta elevada dispersión de la población es que el número de topónimos sea muy elevado.

La principal novedad de este servicio es de naturaleza técnica y radica en que en vez utilizar como fuente de datos una base de datos espacial o similares (*shapefiles*, *geodatabases*, etc.), se utiliza el motor de búsqueda Apache Solr. La principal ventaja que aporta este motor de búsqueda es que devuelve los resultados de la búsqueda ordenados en función de la similitud a los términos de búsqueda introducidos por el usuario. Otra ventaja es que permite la puesta en marcha de un servicio de nomenclátor sin necesidad de utilizar ningún sistema de bases de datos.

El resto del presente artículo está estructurado de la siguiente forma: en la sección 2 exponemos las propuestas de normalización de servicios de nomenclátor actualmente existentes y describimos brevemente dos soluciones tecnológicas similares desde un punto de vista funcional; en la sección 3 describimos el sistema, su arquitectura y la fuente de datos utilizado; y, finalmente, en la sección 4 presentamos nuestras conclusiones y líneas de trabajo futuras.

2 Trabajo relacionado.

En lo referente a servicio de nomenclátor, o gazetteer, dentro de INSPIRE existe una especificación, de datos para nombres geográficos: 'Data Specification on Geographical Names – Guidelines' [9]. El formato de los datos

geográficos sigue en principio, lo dictado por el estándard ISO 19112: el apéndice D de la especificación consiste en una propuesta de modificaciones sobre este estándar con el objeto de adaptarse a los requisitos de INSPIRE [10].

En el ámbito de España, el Consejo Superior Geográfico - dependiente del Ministerio de Fomento - ha creado el Modelo de Nomenclátor de España (MNE) [11] en el que se propone la utilización de un perfil del servicio WFS denominado 'Gazetteer Service - Application Profile of the Web Feature Service' [12], creado por el OGC. Cabe destacar que este perfil no es, a día de hoy, un estándar del OGC, si bien este documento refleja la posición oficial de este consorcio en lo referente a esta tecnología.

El MNE recomienda implementar los siguientes criterios de búsqueda:

- Búsqueda por nombre de la entidad, siendo posible especificar si el texto introducido es parte o coincide exactamente con el nombre.
- Búsqueda por localización espacial bien introduciendo las coordenadas por teclado, marcando una ventana de entorno en un mapa en pantalla o seleccionando un entorno por una entidad gráfica como por ejemplo la provincia.
- Búsqueda por el tipo de la entidad, seleccionándola de una lista controlada.

En lo referente a las diferencias que cabe encontrar entre un servicio WFS y un servicio WFS-G, este perfil añade al primero, entre otras, las siguientes funcionalidades:

- Un servicio WFS-G debe identificarse como tal, y no como un servicio WFS.
- Acceso a los metadatos de los servicio de nomenclátor proporcionados.
- Recuperación de nombres geográficos que tienen relaciones padre-hijo.
- Los nombres geográficos son devueltos en objetos de tipo *SI_LocationInstance*.
- Los servicios de nomenclátor son descritos mediante objetos de tipo *SI_Gazetteer*.

Sólo tenemos conocimiento de una solución de código abierto que implemente la propuesta WFS-G. Se trata de degree [13], que proporciona servicios OGC WMS, WFS, Catalogue Service for Web, Web Coverage Service y Web Processing Service. En la versión 2.1 de este software el servicio WFS podía ser configurado para que funcionase como un servicio WFS-G [14]. El Instituto de Cartografía de Andalucía implementa su servicio de nomenclátor WFS-G utilizando degree, versión 2.2 [15].

Por otra parte, López Otero et al. [16] describen el análisis, diseño e implementación de un servicio de nomenclátor que implementa el perfil WFS-G y la norma del MNE.

3 Descripción y arquitectura del sistema.

3.1 Descripción del sistema.

El sistema creado busca, por ahora, solamente entidades singulares de población que hayan sido reconocidas por el INE.

El cliente puede utilizar el servicio nomenclátor de dos formas: o bien hace una búsqueda a través de un formulario mostrado en página web o bien utiliza la API – *Application Programming Interface* – facilitada.

En el primer caso, se ofrece al usuario realizar la búsqueda de una entidad singular de población en toda la provincia o bien limitarla a un municipio determinado. Los resultados de la búsqueda se muestran en la página web. En las figuras 1 y 2 se muestran, respectivamente, capturas de pantalla con el formulario y los resultados de una búsqueda.

Figura 1. Formulario de búsqueda en página web.

En el segundo caso, la API ofrece al usuario mayor número de opciones de búsqueda: por nombre, restringiendo a un municipio, entidades dentro de una caja contenedora, etc. Los resultados son devueltos en formatos como eXtensible Markup Language (XML) o Javascript Object Notation (JSON), a elección del usuario; los tipos MIME son, respectivamente, *application/xml* y *application/json*.

3.2 Arquitectura del sistema.

Se ha creado una aplicación web utilizando el motor de búsqueda Apache Solr [17] (de ahora en adelante, Solr). Solr se ejecuta en un contenedor de servlets [18]. En nuestro caso hemos utilizado las versiones 7.0.29 y 3.6.0 de Apache Tomcat y Solr, respectivamente.

A grandes rasgos, el funcionamiento del sistema es el siguiente: el contenedor de servlets recibe, vía el protocolo HTTP, los parámetros de búsqueda introducidos por el usuario, los cuales, a su vez, son enviados a la instancia de Solr, que hace la búsqueda. Una vez completada ésta, el contenedor de servlets envía los resultados de ésta al cliente que hizo la petición.

Figura 2. Resultados de la búsqueda mostrados en página web.

En la figura 3 se muestra con mayor nivel de detalle de qué partes consta el sistema y cómo interaccionan entre si. Tanto la página web de la EIEL como la API son aplicaciones web independientes, al igual que el motor de búsqueda; tenemos, por lo tanto, tres aplicaciones web independientes. Toda comunicación entre las tres aplicaciones web se hace utilizando el protocolo HTTP.

Cuando el cliente hace una búsqueda a través de la página web de la EIEL, esta aplicación web hará un análisis de los parámetros de búsqueda dados por el usuario y los utilizará para formar una consulta dirigida al motor de búsqueda Solr; esta consulta de acuerdo a la sintaxis especificada por Solr para hacer consultas. Por otra parte, cuando el motor de búsqueda envía los resultados a la página web - en formato JSON -, ésta le da formato HTML.

Cuando el cliente hace un búsqueda a través de la API, el procedimiento seguido es similar, con la diferencia de que el análisis de parámetros es distinto y que la transformación aplicada para crear la salida también es distinta.

Cabe destacar que no es posible acceder de forma directa vía Internet al motor de búsqueda; todas las peticiones de búsqueda deben pasar o bien por la página web de la EIEL o bien por la API.

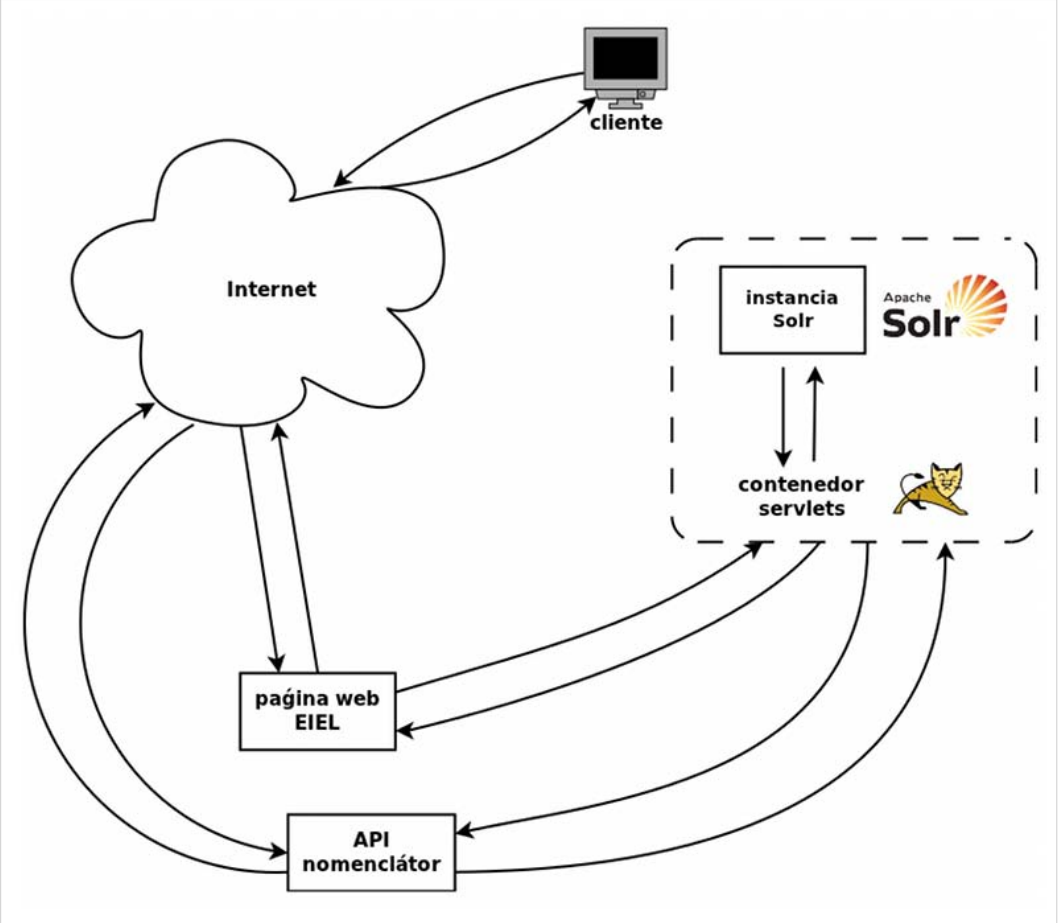

Figura 3. Arquitectura del sistema.

3.3 Fuente de datos.

Solr, al igual que la mayoría de los motores de búsqueda, crea un índice de todos los documentos que potencialmente pueden ser objeto de búsqueda. A este proceso se le denomina indexación.

La creación de un índice es un proceso previo a las búsquedas, ya que éstas se hacen sobre el índice. Sería posible hacer búsquedas directamente sobre los documentos, pero a poco que el volumen de datos fuese elevado, el proceso sería varios órdenes de magnitud más lento; sin índices, las búsquedas tendrían que hacerse secuencialmente, documento a documento. El tiempo ahorrado en las búsquedas lo es a expensas del espacio de almacenamiento - disco magnético, SSD o cualquier otra tecnología - consumido por el índice.

Solr tiene la capacidad de indexar documentos en varios formatos, entre los que están texto plano, HTML, PDF o XML. Nuestra elección para implementar el servicio de nomenclátor ha sido el formato XML.

Para ello, hemos creado un único fichero en formato XML que contiene todas las entidades singulares de la provincia de Lugo, con la excepción del municipio de Lugo - capital de la provincia -, ya que al superar el umbral de 50.000 habitantes no es objeto de la EIEL.

En la figura 4 mostramos un fragmento de una versión simplificada de este fichero XML.

```
<?xml version="1.0" encoding="UTF-8" ?>
<add overwrite="true">
<doc>
<field name="id">20442</field>
<field name="ineCode">270010101</field>
<field name="spanishName">Abadin</field>
<field name="galicianName">Abadín</field>
<field name="envelope">EPSG:25829;POLYGON((623314.279908025 4802146.92807464,623314.279745559
4802737.09325913,624049.428928126 4802737.09305599,624049.42909156 4802146.9278715,623314.279908025
4802146.92807464))</field>
<field name="centroid">EPSG:25829;POINT(623678.046408781 4802438.68112146)</field>
</doc>
<doc>
<field name="id">76791</field>
<field name="ineCode">270010102</field>
<field name="spanishName">Carballás</field>
<field name="galicianName">Os Carballás</field>
<field name="envelope">EPSG:25829;POLYGON((623198.656467093 4801612.76790331,623198.656263873
4802351.67171369,623544.608907637 4802351.67161834,623544.609111427 4801612.76780796,623198.656467093
4801612.76790331))</field>
<field name="centroid">EPSG:25829;POINT(623401.213335326 4801902.19145337)</field>
</doc>
...
```

Figura 4. Fichero XML indexado.

Este fichero está formado por elementos de tipo 'document' (etiqueta 'doc'), los cuales, a su vez, están formados por campos (etiqueta 'field'). Cada elemento de tipo document se corresponde con exactamente una entidad singular. Para cada una de las entidades singulares hay disponible la siguiente información: código INE, nombre en castellano, nombre en gallego, polígono delimitador y centroide de éste; tanto el polígono delimitador como el centroide están en formato *Extended Well-Known Text*, que permite especificar el identificador del sistema de referencia utilizado. Es posible hacer búsquedas consultando cualquiera de estos campos.

Hay un campo adicional, 'id', que consiste en un identificador numérico único para cada entidad - no puede haber, por lo tanto, dos entidades singulares cuyo campo 'id' sea igual - y que, al contrario que el código INE, carece de significado. Si bien un campo de este tipo, que juega dentro de un índice Solr un rol semejante al de claves primarias en los sistemas de bases de datos relacionales, no es estrictamente necesario, siempre es conveniente que exista. Por otra parte, no es posible hacer búsquedas consultando el campo 'id'.

Las razones que nos han llevado a elegir el formato XML frente a otros, como ficheros en texto plano o en formato *Comma Separated Values* son las siguientes:

- XML tiene capacidad para representar estructuras de datos complejas.
- XML tiene un buen soporte para trabajar con información en múltiples idiomas.
- Un fichero en formato XML puede ser validado de acuerdo a un conjunto de reglas pre-establecido, existiendo para ello tecnologías como *Document Type Definition* o *XML Schemas*. Esta funcionalidad es de gran utilidad, ya que no es factible revisar de forma manual ficheros que contengan grandes cantidades de datos.

3.4 Otras funcionalidades aportadas por Solr.

El motor de búsqueda Solr aporta las siguientes funcionalidades:

- Búsquedas complejas gracias al uso de filtros, la posibilidad de hacer consultas contra varios campos o anidar búsquedas.
- Clasificación de los resultados en función de la similitud a los términos de búsqueda introducidos por el usuario. Cabe destacar que es posible configurar el cálculo de la clasificación de forma que las coincidencias encontradas en ciertos campos tengan más peso que las encontradas en otros.
- *Stop words*: palabras frecuentes y formadas por pocos caracteres, como preposiciones o artículos, no son tenidas en cuenta a la hora de crear el índice. El responsable de administrar el motor de búsqueda facilita el listado de éstas introduciéndolas en un fichero de configuración.
- Escalabilidad, gracias a la replicación eficiente de buscadores.
- Posibilidad de hacer búsquedas distribuidas.
- Capacidad limitada de hacer búsquedas espaciales [19], sin necesidad de utilizar un sistema de base de datos espacial. Pese a lo limitado de este tipo de búsquedas, lo ofrecido por Solr es suficiente para implementar un servicio de nomenclátor.
- Paginación de resultados. En aquellos casos en los que una búsqueda tenga como resultado numerosas entidades, puede ser útil que el motor de búsqueda devuelva solamente un rango de éstos, por ejemplo, de 1 a 10, de 11 a 20, etc (*search pagination pattern*). Gracias a esta funcionalidad evitamos sobrecargar a los servidores de trabajo.

4 Conclusiones y líneas de trabajo futuras.

En esta comunicación hemos mostrado qué ventajas tiene utilizar el motor de búsqueda Solr en la implementación de un servicio de nomenclátor comparado con utilizar bases de datos relacionales con capacidades espaciales como PostGIS u Oracle Spatial, y tecnologías similares desde el punto de vista funcional. Hemos expuesto el caso práctico de un servicio nomenclátor de entidades singulares para la provincia de Lugo.

Solr tiene, en nuestra opinión, unas capacidades de búsqueda superiores a las ofrecidas por los sistemas de bases de datos relacionales. Por otra parte, su escalabilidad y posibilidad de hacer búsquedas distribuidas hacen que esta tecnología sea apta para ser utilizada en entornos con muy altas cargas de trabajo. Pese a que las funcionalidades de consultas espaciales que ofrece son mucho más limitadas que las de las bases de datos relacionales antes mencionadas, éstas son suficientes para implementar un servicio de nomenclátor.

Nuestras líneas de trabajo futuras son dos. Por una parte, ampliar el servicio de nomenclátor para que pueda buscar, además de entidades singulares de población, servicios de utilidad pública relacionados con los asentamientos rurales (abastecimientos y saneamientos de aguas, alumbrados públicos, etc). La búsqueda puede de ser utilidad para otras administraciones como la Agencia Tributaria o la Dirección General de Catastro. Por otra parte, el servicio implementado no sigue el estándar WFS-G, ni tampoco las recomendaciones del MNE, deficiencias que habría que subsanar.

5 Referencias bibliográficas.

[1] real decreto 835/2003: http://goo.gl/lm1Ze

[2] orden ministerial APU/293/2006, de 31 de enero: http://goo.gl/18jlk

[3] página web LaboraTe: http://laborate.usc.es/

[4] página web EIEL Lugo: http://www.idealugo.es/

[5] Sánchez Maganto, A., Rodríguez Pascual, A. F., Abad Power, P., López Romero, E.; 2005; “El Núcleo Español de Metadatos, perfil mínimo de metadatos recomendados para España”; III Jornadas Técnicas de la IDE de España (JIDEE 2005)

[6] texto directiva INSPIRE: http://goo.gl/NS4Yt

[7] INSPIRE implementing rules: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/47

[8] texto LISIGE (Boletín Oficial del Estado): http://goo.gl/sH2Ja

[9] 'INSPIRE Data Specification on Geographical Names – Guidelines': http://goo.gl/9k9uF

[10] Dutch National Stimulation Program on SDI (RGI-116), Wiki on geo-standards, sección 6.4.19, 'Gazetteers': http://goo.gl/Tetwi

[11] Modelo de Nonemclátor de España, versión 1.2: http://goo.gl/aGFE5

[12] 'Gazetteer Service - Application Profile of the Web Feature Service Best Practice': http://goo.gl/X1LCy

[13] página web de degree: http://www.deegree.org/

[14] manual de degree (versión 2.1), sección 6.6, 'Setting up WFS-Gazetteer (WFS-g)': http://goo.gl/VgJ4N

[16] Garrido Borrego, M. T., Torrecillas Lozano, C, Tarterá Ansay, Ll.; 2009; “Interoperabilidad del Servicio de Nomenclátor del Instituto de Cartografía de Andalucía. ” ; III Jornadas de SIG Libre

[16] López Otero, M.J., Luaces, M. R., Paramá , J.R.; 2007; “Implementación de un Servicio de Nomenclátor según la norma MNE y el estándar WFS-G ”; IV Jornadas Técnicas de la IDE de España (JIDEE 2007)

[17] página web Apache Solr: http://lucene.apache.org/solr/

[18] servlets Java, entrada en la Wikipedia: http://en.wikipedia.org/wiki/Java_Servlet

[19] Solr wiki, Spatial Search: http://wiki.apache.org/solr/SpatialSearch/