Una aproximación a la búsqueda distribuida de topónimos

C. Laborda[1], R. Recio[2], J.M. Agudo[2], A.F. Rodriguez[3], A. Florczyk[2], F.J. López-Pellicer[2].

[1]GeoSpatiumLab
C/ Carlos Marx, 6, 50.015 Zaragoza
claborda@geoslab.com

[2]Dpto. de Informática e Ing. de Sistemas
Universidad de Zaragoza
C/ María de Luna 1, 50.018 Zaragoza
{rociorm, joselilo, florczyk, fjlopez}@unizar.es

[3]Instituto Geográfico Nacional
C/ General Ibáñez de Ibero, 3, 28.003 Madrid
afrodriguez@fomento.es

Resumen

Un servicio fundamental con el que debe contar una Infraestructura de Datos Espaciales (IDE) es el servicio de nomenclátor, que permite realizar consultas sobre repositorios de topónimos. El siguiente objetivo a lograr es como compartir esta información, es decir, construir una aplicación en una IDE que permita acceder a los repositorios de topónimos de otras IDEs. A este tipo de aplicación se le denomina Nomenclátor Distribuido. En el presente artículo se presentan dos posibles soluciones a este objetivo, junto con los problemas que conllevan, para finalizar mostrando la arquitectura y puesta en marcha de una aplicación de nomenclátor distribuido capaz de comunicarse con los diferentes tipos de servicios de nomenclátor que pueden ofrecerse en cada una de las IDEs.

Palabras clave: Servicio, Nomenclátor, Gazetteer, Infraestructura de Datos Espaciales, Topónimo, Distribuido

1 Introducción

Uno de los servicios básicos que debería incluir una IDE es el servicio de nomenclátor (gazetteer). Según la Real Academia Española de la lengua un nomenclátor es un “catálogo de nombres, ya de pueblos, ya de sujetos, ya de voces técnicas de una ciencia o facultad”. En el caso de las Infraestructuras de Datos Espaciales hablaríamos de un nomenclátor como un catálogo de entes del mundo real que contiene alguna información sobre su posición (ISO19112)[2]. Estos topónimos pueden referirse tanto a lugares (países, municipios, etc.) como a accidentes geográficos (ríos, montañas, etc.) o cualquier otra entidad georreferenciable (árboles, puertos, etc.). Asociada a los nombres de los topónimos podemos encontrar por ejemplo la siguiente información: localización espacial, dimensiones del lugar (extensión o población de un municipio, altura de una montaña, etc.), unidad administrativa a la que pertenece, etc. Por lo tanto cuando hablamos del servicio de nomenclátor de una IDE nos estamos refiriendo al servicio que permite buscar nombres de topónimos y acceder a la información asociada a los mismos.

Una vez definido lo que es un servicio de nomenclátor y su utilidad en una IDE surge el siguiente problema: cómo compartir los datos accedidos por el nomenclátor de una IDE entre diferentes IDEs. Sería razonable pensar en una aplicación contenida en la Infraestructura de Datos Espaciales Española (IDEE) que permitiera realizar búsquedas sobre los topónimos geográficos contenidos en su propio servicio de nomenclátor así como en los servicios de nomenclátor propios de otras IDEs nacionales o IDEs a nivel autonómico o incluso IDEs de otros países. Este planteamiento permitiría crear por ejemplo una IDE a nivel europeo en la que se pueda acceder a todos los topónimos incluidos en todos los países que componen la Unión Europea.

Para solventar este problema generalmente se plantean estrategias de dos tipos: Técnicas de Harvesting y Técnicas de distribución de datos. Las Técnicas de Harvesting consisten en recopilar en un único repositorio todos los topónimos incluidos en cada uno de los repositorios de las diferentes IDEs. Esta solución tiene la ventaja de tener una respuesta más rápida ante las consultas que se le hagan ya que únicamente se está accediendo a un único repositorio. Pero tiene el gran problema de la réplica de la información. Mediante esta técnica los datos se encuentran tanto en los repositorios de cada una de las IDEs como en el repositorio de la IDE central. Esto complica de gran manera el mantenimiento de la consistencia de la información, ya que en todo momento los datos contenidos en ambos repositorios deben ser los mismos, es decir cualquier cambio en un dato en uno de los dos repositorios debe ser trasladado inmediatamente al otro. Por otra parte mediante esta técnica podríamos encontrarnos con repositorios de tamaño insostenible, aparte de que por diversas circunstancias, como pueden ser los derechos de propiedad de los datos, puede resultar imposible conseguir la información de todas las IDEs para recopilarla en un único repositorio externo a las mismas. Parece por ello una mejor solución mantener los datos separados en distintos repositorios.

Por otro lado, en las Técnicas de distribución de datos los datos se encuentran únicamente en el repositorio de cada IDE y la aplicación central que accede a todos ellos lo hace a través del servicio de nomenclátor que ofrece cada una de las IDEs. De esta manera aunque la devolución de resultados ante una búsqueda pueda resultar ligeramente más lenta, solventamos los serios y en ciertos casos irresolubles problemas que surgían con la solución anterior. La utilización de las técnicas de distribución de datos consiste por lo tanto en una aplicación central que accede a cada uno de los servicios de nomenclátor de diferentes IDEs. En el presente artículo se ofrece una solución al problema del nomenclátor distribuido empleando esta estrategia.

En las siguientes secciones se muestran, la arquitectura utilizada para la implementación del nomenclátor distribuido, así como la puesta en marcha de la misma en la IDEE.

2 Arquitectura del nomenclátor distribuido

El nomenclátor distribuido se ha estructurado siguiendo un modelo de arquitectura multicapa, donde cada uno de los componentes que integran el producto final puede agruparse en distintos niveles en función de cuál sea su relación con el acceso a los datos o su interacción con el usuario final. En concreto, se han distinguido tres niveles o capas arquitecturales claramente diferenciadas (véase Figura 1):

- un nivel de almacenamiento de datos (*Data Sources*) que, como su nombre indica, agrupa a las distintas fuentes de datos utilizadas por la aplicación, es decir, los diferentes servicios de nomenclátor a los que se accede;
- un nivel de servicios (*Application Services* y *Access Services*) encargado de la recuperación de los datos y su procesamiento;

- y por último, un nivel de aplicación (*Web Applications*), constituido por los componentes que interaccionan con el usuario final, bien recogiendo sus peticiones, bien facilitándole los resultados generados en el nivel anterior.

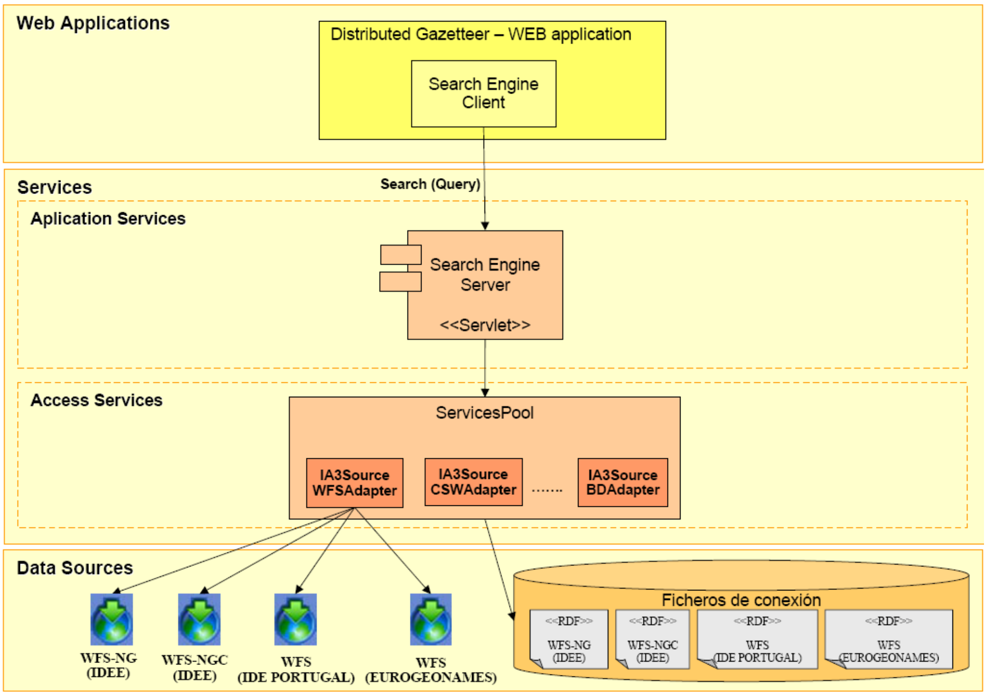

Figura 1. Arquitectura general del nomenclátor distribuido

En la Figura 1 se observa que el nivel más alto de la arquitectura, el nivel de aplicación, está formado por un componente fundamental, la aplicación *Distributed Gazetteer*, que es la aplicación final que un usuario puede ejecutar en su navegador de Internet. Dicho componente es el encargado de permitir al usuario introducir las restricciones de búsqueda a partir de las cuales se genera la consulta que será lanzada a los distintos servicios de nomenclátor para a continuación mostrar el listado con los resultados. Este componente también permite la conexión en línea con un cliente genérico de mapas, en el cual se podrá localizar el resultado seleccionado, así como permitir visualizar el mismo en detalle. Todos los componentes de este nivel han sido desarrollados haciendo uso de la tecnología *Google Web Toolkit* (GWT), un conjunto de herramientas de software libre, desarrolladas por Google, que facilitan la creación de aplicaciones Web con AJAX utilizando Java como lenguaje de programación.

En lo que respecta al nivel de servicios, se ha considerado oportuno distinguir entre aquellos componentes que realizan tareas de procesamiento de datos, a los que se ha denominado *Application Services*, y los que están más relacionados con la recuperación de la información, denominados *Access Services*. El primer nivel está constituido por el componente *Search Engine Server*, implementado mediante la tecnología de servlets de Java. El componente *Search Engine Server* es el encargado recibir las consultas del nivel aplicación (recibido desde el navegador del usuario) y pasarselas al componente del siguiente nivel *Services Pool* que actuará como un multiplexor entre la aplicación de búsqueda y cada uno de los adaptadores que contiene y que se detallarán más adelante. Por otra parte el componente *Search Engine Server* es también el encargado de construir las estructuras de datos necesarias para pasarle al nivel aplicación para que se le muestren al usuario dichos resultados.

El segundo subnivel de la capa de servicios está formado por un componente fundamental, el *Services Pool*. Este componente es el encargado de recuperar los datos existentes en el sistema de almacenamiento final mediante la creación de adaptadores necesarios para conectarse a cada uno servicios de nomenclátor a los que se haya configurado que deba conectarse la aplicación de nomenclátor distribuido. Como no existe una norma bajo la que se implementen todos los servicios de nomenclátor y que asegure que todo servicio de nomenclátor tiene una misma API (*Application Programing Interface*) de consulta y un mismo formato de respuesta, nos encontramos con que estos servicios pueden ser de diversos tipos. Se pueden encontrar servicios de nomenclátor implementados mediante un servicio estándar WFS (*Web Feature Server*)[3], un servicio estándar CSW (*Catalog Services for the Web*)[4] o cualquier otro servicio, tanto estándar como no estándar, para el que se haya implementado un adaptador propio. Para que este componente sepa qué tipo de conector/adaptador debe utilizar para conectarse con un determinado servicio de nomenclátor se utilizan una serie de ficheros de configuración expresados mediante RDF (*Resource Description Framework*)[5] (uno para cada servicio de nomenclátor) donde se describen las características del mismo: URL de consulta, formato de la consulta, formato de la respuesta, etc. A modo aclarativo, se podría decir que el conjunto formado entre el *ServicesPool* y cada uno de los conectores hacen un papel de middleware entre las capas superiores y los servicios finales ofreciendo una capa de abstracción sobre las distintas fuentes de datos (servicios de nomenclátor) que facilita a los componentes de niveles superiores el acceso a las mismas de forma transparente, independientemente de cuál sea su API o su soporte físico.

Por último, el nivel inferior de la arquitectura, encontramos las fuentes de los datos, es decir este nivel está constituido por los diferentes servicios de nomenclátor a los que tiene acceso el nomenclátor distribuido. Por otra parte en este nivel también se encuentra el repositorio con los diferentes archivos RDF que describen las conexiones a cada uno de esos servicios.

3 Puesta en marcha del nomenclátor distribuido en la IDEE

Siguiendo la arquitectura descrita en el apartado anterior se ha realizado un primer prototipo del nomenclátor distribuido de la IDEE (http://www.idee.es/IDEE-Gazetteer/index.html). Dicho prototipo permite realizar búsquedas distribuidas de topónimos en cuatro servicios de nomenclátor distintos:

- Nomenclátor NOMGEO: Este servicio publica los contenidos del Nomenclátor del Instituto Geográfico Nacional cuya estructura sigue las recomendaciones recogidas en el MNE (Modelo de Nomenclátor de España).
- Nomenclátor NGCE: Este servicio publica los contenidos del Nomenclátor Geográfico Conciso de España cuya estructura, al igual que en el nomenclátor NOMGEO, sigue las recomendaciones del MNE.
- Nomenclátor de la IDE de Portugal.
- Nomenclátor Eurogeonames.

La pantalla principal de la aplicación Web muestra el formulario de especificación de consultas de búsqueda (véase Figura 1). Dicho formulario permite establecer distintos criterios de restricción como son la extensión geográfica en la que se desea buscar topónimos, las unidades administrativas en las que se encuentren los lugares consultados, las palabras clave que los definan o el tipo de de topónimo buscado.

Al pulsar en el botón “Buscar” se lanzará una búsqueda contra cada uno de los servicios de nomenclátor soportados por la aplicación.

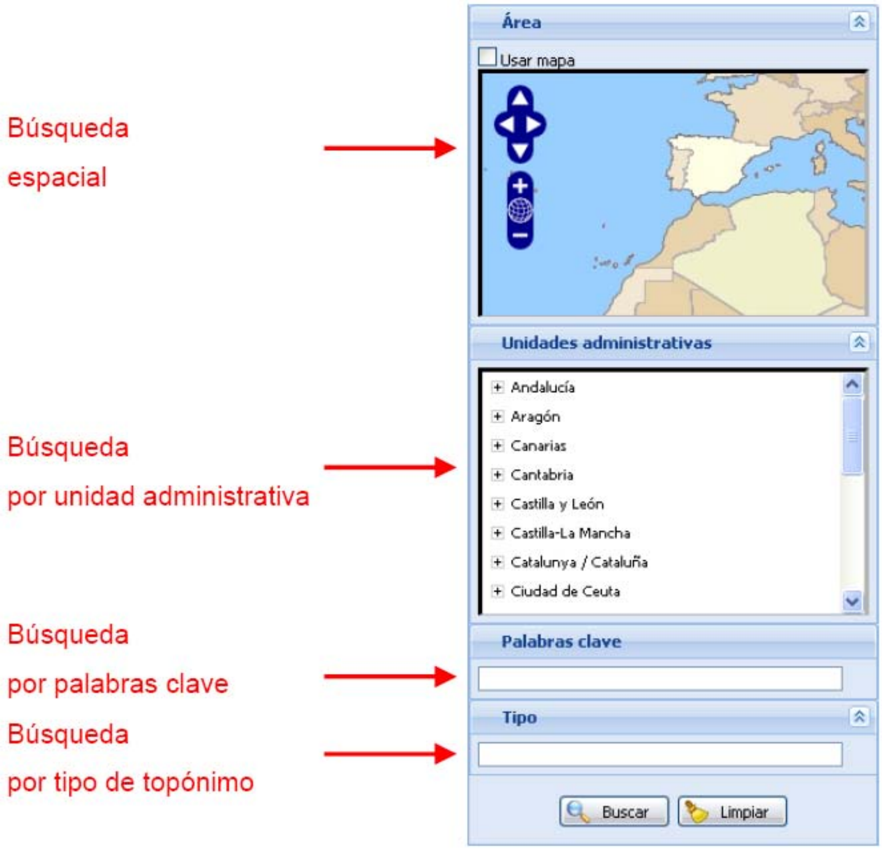

Figura 2. Criterios de búsqueda del nomenclátor distribuido

En la Figura 3 muestra la presentación de resultados. La aplicación genera una pestaña por cada uno de los servicios de nomenclátor consultados en la que mostrará los topónimos encontrados por cada uno de ellos. Cada pestaña contiene el número de resultados total encontrados por el servicio de nomenclátor, así como la lista de resultados paginada. Por cada resultado se muestra el nombre del topónimo (que da acceso a la visualización de los datos del mismo en detalle), el tipo del topónimo y un enlace para mostrar la localización del topónimo en un servicio de visualización de mapas.

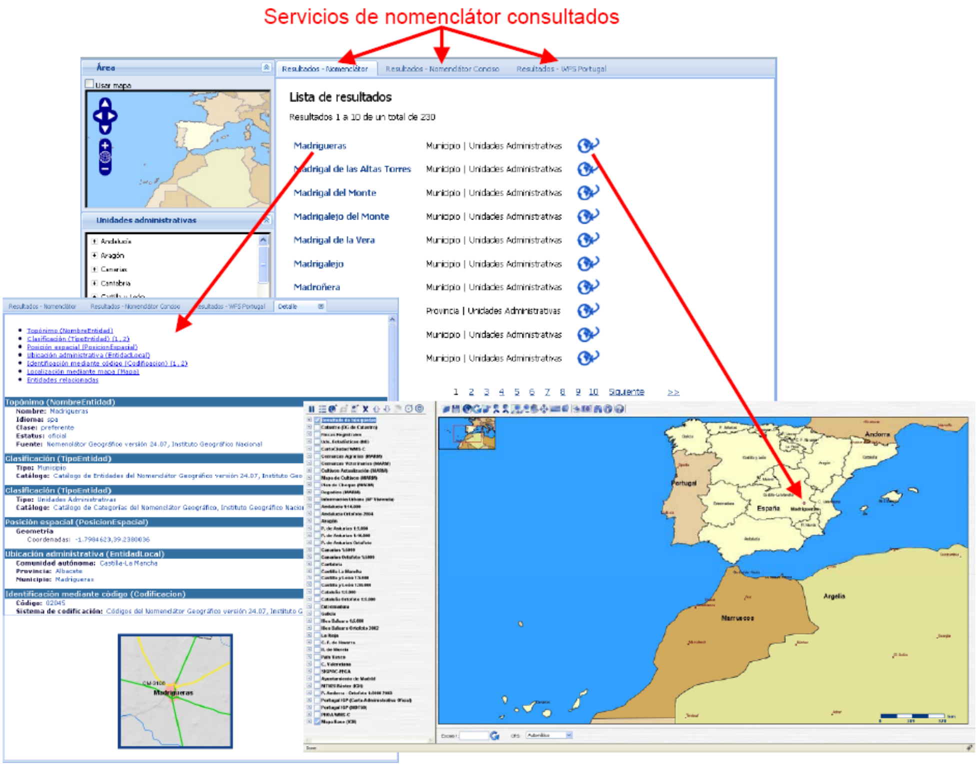

Figura 3. Presentación de resultados y conexión en línea con el servicio de visualización de mapas

4 Conclusiones y trabajo futuro

Este trabajo ha presentado el desarrollo y puesta en marcha de un primer prototipo de nomenclátor distribuido de la IDEE. Gracias a este prototipo es posible realizar búsquedas simultáneas de topónimos sobre diferentes servicios de nomenclátor desde la misma aplicación.

El principal problema encontrado en el desarrollo es que no existe un único estándar de consulta y devolución de resultados de los servicios de nomenclátor. Esto obliga a tener que implementar conectores distintos que sepan cómo consultar a cada uno de ellos, así como construir resultados en un determinado formato a partir de los diversos formatos que pueden ser devueltos.

En cuanto a las posibles mejoras del prototipo se pueden destacar las siguientes líneas de trabajo. En primer lugar, se pretende dar la posibilidad de que el usuario pueda seleccionar los diferentes servicios de nomenclátor a los que se desea consultar de manera distribuida, obteniendo la lista de servicios que se le ofrece al usuario mediante una consulta al catálogo de servicios de la IDEE. En segundo lugar se pretende poder añadir nuevos servicios de nomenclátor a los que poder consultar sin necesidad de tener que implementar *ad-hoc* nuevos conectores/adaptadores a los mismos. Es decir que estos conectores se autogeneren a partir de la información que dispongan los propios servicios sobre la forma de consultarlos y el formato de las respuestas mediante un proceso de descubrimiento de servicio.

Agradecimientos. Este trabajo es una iniciativa del Instituto Geográfico Nacional de España (IGN), junto con el Ministerio de Ciencia e Innovación (ref. TIN2007-65341) y fruto de la colaboración científico/técnica con el Grupo de Sistemas de Información Avanzados de la Universidad de Zaragoza (IAAA) y el apoyo tecnológico de GeoSpatiumLab S.L.

Referencias

[1] Nebert, D., ed.: Developing Spatial Data Infrastructures: The SDI Cookbook v.2.0. Global Spatial Data Infrastructure (2004) http://www.gsdi.org.

[2] Norma ISO19112: 2003 “Geografic Information – Spatial referencing by Geographic Identifiers”

[3] Panagiotis A. Vretanos: 2005 “Web Feature Service Implementation Specification” OGC 04-094

[4] Douglas Nebert, Arliss Whiteside, Panagiotis (Peter) Vretanos: 2007 “OpenGIS® Catalogue Services Specification” OGC 07-006r1

[5] Formato estándar RDF: http://www.w3.org/RDF/