Un Servicio de Resolución de Topónimos siguiendo el estándar OGC WPS*

Ana Cerdeira-Pena, Miguel R. Luaces, Óscar Pedreira, Diego Seco

Laboratorio de Bases de Datos
Universidade da Coruña
Campus de Elviña S/N
{acerdeira, luaces, opedreira, dseco}@udc.es

Resumen

Dentro del campo de los Sistemas de Información Geográfica (SIG) se está haciendo un esfuerzo muy importante por parte de diversos organismos internacionales para la definición de estándares para lograr la interoperabilidad de los sistemas desarrollados. Uno de los estándares más recientes propuesto por el Open Geospatial Consortium es el Web Processing Service (WPS), diseñado para estandarizar la forma en que los procesos SIG se ofrecen a través de Internet.

En este trabajo presentamos un WPS para realizar *Resolución de Topónimos*. Este servicio define dos operaciones espaciales. La primera de ellas, *getAll*, permite obtener todas las posibles descripciones geográficas con el nombre de lugar consultado ordenadas según un ranking de relevancia. La segunda operación, *getMostProbable*, filtra el resultado de la anterior devolviendo la descripción geográfica más importante. Además, ambas operaciones pueden ser parametrizadas de acuerdo con el nivel de detalle necesario en el resultado.

Palabras clave: Sistemas de Información Geográfica (SIG), Open Geospatial Consortium (OGC), Web Processing Service (WPS), ontología, servicio de nomenclátor.

* Este trabajo ha sido parcialmente financiado por el "Ministerio de Educación y Ciencia" (PGE y FEDER) ref. TIN2006-16071-C03-03, por la "Agencia Española de Cooperación Internacional (AECI)" ref. A/8065/07, por la "Xunta de Galicia" ref. 2006/4 y ref. 08SIN009CT, y por el "Ministerio de Educación y Ciencia" ref. AP2007-02484 (Programa FPU) para Ana Cerdeira.

1 Introducción

Los Sistemas de Información Geográfica (SIG) [1] constituyen un campo de investigación que ha recibido mucha atención en los últimos años. Las mejoras recientes en el hardware han hecho posible que la implementación de este tipo de sistemas sea abordable por muchas organizaciones. Además, se ha llevado a cabo un esfuerzo colaborativo por dos organismos internacionales (ISO [2] y el Open Geospatial Consortium [3]) para definir estándares y especificaciones para la interoperabilidad de los sistemas. Este esfuerzo ha hecho posible que muchas organizaciones públicas estén trabajando en la construcción de infraestructuras de datos espaciales [4] que les permitirán compartir su información geográfica.

El *Open Geospatial Consortium* (OGC) es una organización internacional sin ánimo de lucro destinada fundamentalmente a la realización de especificaciones no propietarias en el ámbito de los Sistemas de Información Geográfica (SIG) para facilitar la interoperabilidad de los sistemas [5]. Las especificaciones del OGC son documentos técnicos donde se detallan las interfaces y codificaciones que deben seguir los servicios y productos implementados para obtener servicios interoperables. Estas especificaciones tienen una amplia aceptación entre la comunidad de desarrolladores e investigadores en SIG y muchos de ellos "*han mostrado una extraordinaria cooperación en equipo para colaborar en las especificaciones OpenGIS*" [6]. El equipo del OGC está compuesto por profesionales de varios campos, a diferencia de otros comités cerrados en entornos corporativos, por lo que los estándares que realizan suelen ser de gran calidad y adecuados para numerosas áreas de aplicación.

Una de las especificaciones más recientes del OGC es la especificación *Web Processing Service* (WPS) [7]. La versión 1.0.0 de este estándar fue aprobada el 8 de junio de 2007. Por lo tanto, esta especificación es relativamente nueva si se compara con otras especificaciones más consolidadas, y ampliamente utilizadas, como son la especificación WMS (*Web Map Service*) o la especificación WFS (*Web Feature Service*). La especificación WPS describe un mecanismo por el cual los procesos geográficos pueden ser ejecutados en servidores remotos, empleando fundamentalmente XML [8] para la comunicación a través de la red. Esta especificación está diseñada para el desarrollo de sistemas totalmente independientes tanto de la plataforma como del lenguaje de programación empleados. En este artículo describimos brevemente las características más relevantes de esta especificación y presentamos un servicio de *Resolución de Topónimos* desarrollado siguiendo esta interfaz.

La *Resolución de Topónimos* es una línea de investigación en la que se trata el problema de relacionar nombres de lugar con la representación geográfica de los lugares referidos (por ejemplo, con sus coordenadas en latitud/longitud) [9]. Esta tarea se emplea frecuentemente en campos como la *Recuperación de Información Geográfica* (GIR), *question answering* o *generación de mapas*. El campo de investigación en GIR [10] apareció recientemente en la confluencia de los *Sistemas de Información Geográficos* [1] y de la *Recuperación de Información* [11]. El objetivo principal de este campo es definir estructuras de indexación y técnicas para almacenar y recuperar documentos de manera eficiente empleando tanto el texto como las referencias geográficas contenidas en el mismo. Por tanto, los documentos tienen que ser anotados con los topónimos mencionados en el texto. Esta tarea se ha automatizado recientemente obteniendo resultados comparables con un procesamiento manual [12]. Sin embargo, cuando los documentos tienen que ser indexados espacialmente no es suficiente con localizar los topónimos mencionados en ellos. En este caso, esos topónimos deben ser descritos siguiendo un modelo geográfico del mundo (por ejemplo, empleando las coordenadas en latitud/longitud). Para obtener estas georreferencias se puede emplear un *nomenclátor*.

Un nomenclátor es un catálogo de entidades del mundo real con información descriptiva de cada entidad. Entre esa información descriptiva se puede encontrar, además de información sobre su posición geográfica, otros nombres alternativos, poblaciones, etc. Por ejemplo, un nomenclátor puede catalogar los ríos del mundo, los municipios de España, o los dólmenes de Galicia. Sin embargo, los nomenclátores no son suficientes para automatizar completamente la tarea de la georreferenciación porque proporcionan los topónimos y las coordenadas asociadas con ellos sin ninguna medida de relevancia. Este problema está relacionado con la *ambigüedad referencial*. Por ejemplo, “Londres” es la capital del Reino Unido pero también es una ciudad en Ontario, Canadá. Ante la pregunta *dónde está Londres*, un nomenclátor devuelve ambas localizaciones sin dar ninguna indicación sobre cual de ellas es más apropiada.

Además, los nomenclátores no suelen proporcionar geometrías para los topónimos más complejas que un simple punto representativo (sus coordenadas). Sin embargo, algunas veces se necesita la geometría real del topónimo. En [13], los autores describen una estructura de indexación espacial donde los nodos de la estructura están conectados mediante relaciones de inclusión espacial. Por tanto, cada nodo no hoja almacena, además del topónimo, el *bounding box* de la geometría. Para aplicaciones de ese tipo, se necesita un servicio que devuelva no sólo la localización más probable, sino también su geometría completa para construir el

índice espacial.

En este artículo, presentamos un servicio para realizar *Resolución de Topónimos*. Este servicio proporciona una operación para obtener todas las descripciones geográficas posibles para un topónimo ordenadas siguiendo un ranking de importancia. Además, el servicio proporciona una operación para obtener la descripción geográfica más probable. Ambas operaciones pueden ser parametrizadas de acuerdo con el nivel de detalle necesario en el resultado (por ejemplo, si es suficiente con un simple punto representativo o se requiere la geometría completa). De acuerdo con la tendencia actual en el campo de los GIS, estas operaciones, o procesos espaciales, se ofrecen como un servicio que sigue la especificación WPS.

El resto de este artículo está organizado como sigue. En primer lugar, describimos algunos trabajos relacionados en la Sección 2. A continuación, en la Sección 3, se describen brevemente las características más importantes de la especificación WPS. La Sección 4 presenta la arquitectura general del servicio WPS desarrollado y describe sus componentes principales. Luego, la Sección 5 describe algunos detalles de la implementación del sistema. Finalmente, la Sección 6 presenta algunas conclusiones acerca del trabajo realizado y describe algunas líneas de trabajo futuro.

2 Trabajo relacionado

En el campo de la información geográfica están adquiriendo gran relevancia las llamadas Infraestructuras de Datos Espaciales (IDE) [4]. Una IDE es el conjunto de tecnologías, estándares y recursos humanos necesarios para adquirir, procesar, almacenar, distribuir y mejorar la utilización de la información geográfica. Como componentes principales de una IDE podemos mencionar los siguientes: un servicio de catálogo de metadatos, un servicio de publicación de mapas, y un servicio de nomenclátor.

El servicio de nomenclátor se define como aquel servicio que devuelve las descripciones completas de las entidades geográficas seleccionadas mediante la consulta de sus identificadores. El uso más común de un servicio de nomenclátor es almacenar un catálogo de entidades del mundo real, junto con los topónimos que los identifican, y permitir a un usuario localizar la ubicación de la entidad partiendo de su topónimo. Esta consulta debe soportar además la selección de atributos de las entidades, como pueden ser el nombre, el tipo de entidad o la localización

geográfica. La definición de estos servicios se encuentra en el perfil del estándar internacional del OGC *Gazetteer Service para WFS* (WFS-G) [14] y en el estándar nacional del Consejo Superior Geográfico para un *Modelo de Nomenclátor de España* [15]. Las principales diferencias de un WFS-G con respecto a un WFS son:

- El documento que describe los metadatos del servicio tiene una sección adicional que describe la estructura del nomenclátor.
- Los tipos de entidades geográficas de un WFS-G serán especializaciones del tipo predefinido *SI_LocationInstance*. De esta forma, todos los tipos de entidades geográficas del servicio tendrán un conjunto de atributos básicos comunes y un conjunto de atributos específicos del servicio en particular.

El Modelo de Nomenclátor de España (MNE) [15] define una estructura de datos cuya finalidad es el almacenamiento y gestión de los nombres geográficos o topónimos, con todas las propiedades y relaciones relevantes. Este modelo es completamente compatible con la especificación WFS-G y en él se establece el conjunto de atributos que se consideran fundamentales para caracterizar a un topónimo y otros opcionales que permiten enriquecer la descripción del mismo pero que no se consideran imprescindibles para la implementación del modelo. La inclusión del término *"de España"* refleja que la finalidad de esta iniciativa es llegar a consensuar un modelo común de nomenclátor en España que facilite el intercambio de datos, la interpretación de la información, la descentralización de la gestión, la actualización de un posible nomenclátor distribuido y la implementación de búsqueda en cascada en los nomenclátores integrados en la IDE de España.

Por otra parte, existen una gran cantidad de recursos gratuitos en Internet que proporcionan funcionalidades de nomenclátor, ontologías del espacio geográfico, etc. Los más conocidos son *Alexandria Digital Library* [16], *Getty Thesaurus of Geographic Names* [17] o *Geonames* [18]. Sin embargo, ninguno de estos recursos sigue las especificaciones antes mencionadas para lograr la interoperabilidad de sistemas.

Una desventaja muy importante de los nomenclátores es que no suelen proporcionar una descripción geográfica completa de las localizaciones obtenidas mediante una consulta. Normalmente el resultado incluye un simple punto representativo de la posición del lugar. Existen varios recursos cartográficos que se pueden emplear para completar la información proporcionada por los nomenclátores. Las cartografías de *Global Administrative Unit Layers* (GAUL) [19] y *Vector Map* (VMAP) [20] son muy interesantes ya que proporcionan una cartografía completa y actualizada de todo el mundo. Sin embargo, esta cartografía

no se suele ofrecer en los servicios de nomenclátor.

Los nomenclátores son un factor clave en la tarea de *Resolución de Topónimos*. El objetivo de esta tarea es obtener el *referent*e de cada nombre de lugar consultado. El trabajo de Leidner [9] en esta tarea se centra en el campo de investigación en Recuperación de Información Geográfica (GIR). Varios artículos describen la arquitectura de los sistemas GIR y la mayoría comparten la tarea de NERC+R (*Named Entity Recognition and Classification with Resolution*). El principal objetivo de esta tarea es localizar nombres de lugar en los textos de los documentos y relacionarlos con una correspondencia en un modelo del mundo. En los últimos años se han publicado varios artículos que tratan diferentes aspectos de este problema en el contexto de la GIR [21][22][23]. Web-a-where [21] emplea *contenedores espaciales* para identificar nombres de lugar en los documentos, MetaCarta (el sistema comercial descrito en [22]) emplea métodos basados en técnicas de Procesamiento del Lenguaje Natural (NLP) y STEWARD [23] emplea una aproximación híbrida. Una desventaja de los nomenclátores cuando se emplean para esta tarea es que, dado un nombre de lugar, el nomenclátor proporciona una lista de topónimos que no está ordenada por ningún criterio de relevancia. Por tanto, el usuario del nomenclátor debe encontrar un método para ordenar la lista de resultados.

3 OGC Web Processing Service

La especificación WPS [7] es una de las especificaciones del OGC más recientes y define un mecanismo por el cual un cliente puede enviar una tarea de procesado (espacial) a un servidor para que la realice. Recientemente, se han publicado varios artículos que revisan esta especificación y proponen varios ejemplos de su utilidad [24][25]. En esta sección resumimos las características más importantes de esta especificación.

Como ya se ha descrito, la especificación del WPS se centra principalmente en la definición de un protocolo de comunicación entre cliente y servidor. Para realizar esta comunicación, la especificación define un protocolo basado en XML que emplea el método POST de HTTP. Además, las peticiones también se pueden realizar empleando la codificación *Key-Value-Pairs* (KVP) sobre el método GET de HTTP. Además de este protocolo de comunicación, la especificación también define tres operaciones:

- *GetCapabilities*. Mediante esta petición se solicita al servidor una lista de

los servicios de procesado que tiene disponibles. La respuesta a esta petición que no requiere parámetros es siempre un documento XML que tiene dos partes principales. La primera de ellas, *Identificación del Servicio*, es común con la mayoría de las especificaciones del OGC que comparten esta petición y describe información básica acerca del proveedor del servicio, restricciones, etc. La segunda, *Procesos Ofertados*, consiste en la lista de servicios de procesado disponibles en el servidor.

- *DescribeProcess*. Después de analizar la respuesta de la operación *GetCapabilities*, el cliente dispone de una lista de los procesos ofertados por el servidor. La operación *DescribeProcess* se emplea para solicitar más información acerca de un servicio en concreto. La respuesta a esta petición es un documento XML con todas las características del servicio consultado como son el título, identificador, descripción breve, etc., además de toda la información necesaria para poder solicitar el servicio: descripción de los parámetros de entrada, resultados de salida del servicio, etc. Por tanto, los clientes capaces de analizar este documento tendrán conocimiento de qué entradas deben preparar para solicitar la ejecución del servicio y de cómo pueden obtener el resultado del proceso.
- *Execute*. Mediante esta petición el cliente envía los datos de entrada necesarios para la ejecución de un servicio en concreto, esperando por la respuesta del servidor. La respuesta a esta petición también es un documento XML que indica el estado del proceso, las entradas empleadas y la salida del proceso si ya está disponible. La salida puede ser un simple literal (por ejemplo, un resultado numérico o la URL donde está accesible un documento complejo) o una salida compleja (por ejemplo, una colección de *features* descritas en GML [26]).

Los procesos espaciales pueden ser muy complejos y largos (pueden tardar horas, días o incluso semanas). Por tanto, estos procesos deben ser realizados de forma asíncrona. La especificación define para este propósito la descripción del estado en el documento XML de respuesta a la petición *Execute*. El valor *ProcessAccepted* indica que el proceso se recibió correctamente. *ProcessStarted* indica que el servidor está realizando el proceso. P*rocessSucceded* indica que el proceso ha finalizado y, por tanto, el resultado está preparado. Finalmente, *ProcessFailed* indica que surgió algún problema durante la ejecución del proceso.

Una de las características más atractivas de esta especificación es que, debido a su generalidad, puede ser aplicada a un número ilimitado de casos. Cualquier proceso espacial puede ser ofertado a través de Internet siguiendo esta especificación. Sin embargo, hay ciertas cuestiones que se deben considerar para decidir cuando

conviene definir un proceso como WPS y cuando no. En primer lugar, procesos complejos que puedan tardar mucho tiempo en realizarse son los mejores candidatos para ser implementados como un WPS. Sin embargo, si la complejidad del proceso es baja y la mayor parte del tiempo de procesado se consume en la gestión de una gran cantidad de datos almacenados localmente, el proceso puede ser ejecutado de manera más efectiva en local. En segundo lugar, la especificación WPS tiene las ventajas comunes de los servicios web tradicionales de propósito general. Una de las más importantes es que el servicio está centralizado. Por tanto, un WPS es apropiado para el desarrollo de nuevos procesos que están sometidos a cambios continuos. Los desarrolladores de servicios WPS pueden liberar nuevas versiones actualizando únicamente la versión del proceso instalada en el servidor. Finalmente, la especificación WPS facilita la creación de servicios avanzados mediante la *orquestación* de varios servicios.

Recientemente han aparecido varios *frameworks* e implementaciones de la especificación WPS con el objetivo de facilitar su uso. Sin embargo, muchos de ellos son implementaciones de la versión 0.4.0 del estándar. En la implementación de nuestro servicio empleamos el *framework* de *52 North para WPS* [27]. Este *framework* proporciona una arquitectura extensible en la que se pueden integrar de forma sencilla nuevos procesos y codificaciones de datos. Además, su implementación está basada en la versión 1.0.0 de la especificación.

4 Arquitectura del Sistema

La figura 1 muestra nuestra propuesta de arquitectura de un WPS para realizar *Resolución de Topónimos*. En esta arquitectura se pueden diferenciar dos capas independientes: la *capa WPS* y la *capa de Resolución de Topónimos*.

La capa superior de la arquitectura se corresponde con la *capa WPS*. Esta capa está basada en la implementación de *52 North* [27]. En [28], los autores presentan la arquitectura de este framework y un ejemplo de aplicación para realizar generalización cartográfica. Esta arquitectura es muy sencilla. Define un manejador de peticiones (*Request Processor*) que gestiona el protocolo de comunicación con los clientes. Este manejador de peticiones implementa la especificación OGC WPS y encapsula todos los detalles relacionados con el protocolo de comunicación. Para lograr un grado de extensibilidad muy alto, la implementación de *52 North* está organizada en *repositorios* que proporcionan un acceso dinámico a la funcionalidad embebida del WPS. Para cada proceso espacial ofertado por el servidor se defina un algoritmo en el *repositorio de algoritmos*. Por ejemplo, el repositorio de

algoritmos en [28] está compuesto por varios algoritmos de generalización cartográfica. En nuestra implementación, adoptamos la noción de *repositorio* y diseñamos un componente intermedio para adaptar el algoritmo espacial a la interfaz específica del repositorio (ver el patrón de diseño *Adapter* en [29]). Por tanto, nuestra implementación particular de los algoritmos no depende para nada de los detalles de implementación del framework de *52 North.*

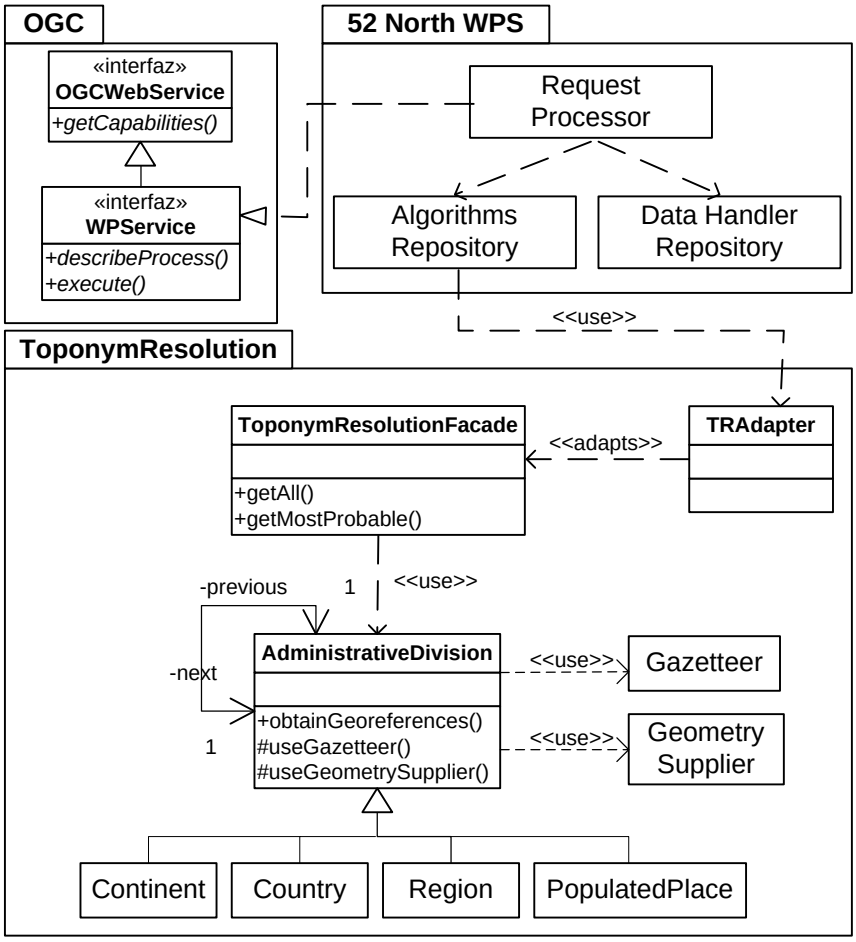

Figura 1. Arquitectura del Sistema

La parte inferior de la figura muestra la arquitectura del componente de *Resolución de Topónimos*. La clase *TRAdapter* representa el *adaptador* entre este componente

y los algoritmos del repositorio de *52 North*. El adaptador emplea la *fachada* [29] *ToponymResolutionFacade* que proporciona una interfaz de acceso simplificada a toda la funcionalidad ofrecida por el servicio. Esta fachada define dos operaciones públicas: *getAll* y *getMostProbable*. La primera de ellas devuelve todas las descripciones geográficas posibles en respuesta a un nombre de lugar. Existen dos diferencias fundamentales entre esta operación y la funcionalidad ofrecida por un nomenclátor. En primer lugar, nuestra implementación se puede configurar para obtener la geometría real, el *bounding box* o un simple punto representativo. En segundo lugar, proporcionamos las descripciones ordenadas según un ranking de relevancia. La segunda operación definida en la interfaz filtra el resultado de la anterior devolviendo sólo la descripción geográfica más probable.

La implementación de ambas operaciones emplea una jerarquía de divisiones administrativas (*AdministrativeDivision*) que define cuatro niveles administrativos (continente, país, región y lugar poblado). En el diseño de la jerarquía se emplearon varios patrones de diseño para lograr una arquitectura robusta y altamente extensible. En primer lugar, la jerarquía sigue el patrón *Chain of Responsibility* [29]. Por tanto, la clase que representa cada nivel administrativo se encarga de una parte del proceso completo y delega el resto en el siguiente nivel. En concreto, hay dos cadenas de responsabilidad configuradas en el sistema. La primera de ellas se emplea para descender la jerarquía obteniendo los nombres lugar que se corresponden con el consultado. Una vez que se encuentra un topónimo, se emplea la segunda cadena para ascender por la jerarquía construyendo una ruta completa que describa completamente el topónimo en la jerarquía del mundo. Por ejemplo, si el nombre de lugar consultado es *A Coruña*, la ruta completa estará compuesta por las descripciones geográficas de *Europa*, *España*, *Galicia* y *A Coruña*. Además, los algoritmos para obtener las referencias geográficas fueron diseñados siguiendo el patrón *TemplateMethod* [29]. Por tanto, la superclase (*AdministrativeDivision*) define el algoritmo general y los pasos concretos se definen en cada subclase. Estos pasos definen cómo emplear el *nomenclátor* y el *proveedor de geometrías* en cada nivel. En la Sección 5 presentamos más detalles sobre los algoritmos concretos implementados en el sistema para recuperar y elaborar el ranking de topónimos.

5 Implementación

Como hemos indicado en la sección anterior, la capa de *Resolución de Topónimos* emplea un nomenclátor (*Gazetteer*) y un proveedor de geometrías (*Geometry*

Supplier) para obtener las descripciones geográficas. En nuestra implementación de prueba empleamos *Geonames* [18] que proporciona una base de datos geográfica disponible bajo licencia con atribuciones *Creative Commons*. Esta base de datos contiene más de dos millones de lugares poblados de todo el mundo con sus coordenadas latitud/longitud en WGS84 (*World Geodetic System 1984*). Todos los lugares poblados están organizados en categorías de tal manera que es posible clasificarlos en los diferentes niveles de divisiones administrativas definidos por la arquitectura (continentes, países, regiones y lugares poblados).

Sin embargo, Geonames (y los Gazetteers en general) no proporcionan geometrías más allá de un simple punto representativo y, como ya explicamos en la motivación del trabajo, para ciertos dominios se necesita la geometría real del lugar (por ejemplo, el borde de los países) o al menos su *bounding box*. Por ese motivo, definimos un servicio proveedor de geometrías para obtener las geometrías de esos topónimos. Como base de este servicio empleamos la cartografía de Vector Map (VMap) [20]. VMap es una actualización y versión mejorada de la cartografía proporcionada por la *National Imagery and Mapping Agency's Digital Chart of the World*. Esta cartografía nos proporciona geometrías para las divisiones administrativas de primer y segundo nivel en un formato propietario. Sin embargo, existen varias herramientas de software libre que permiten crear archivos *shapefile* desde ese formato, en este trabajo empleamos FWTools [30]. A partir de estos archivos *shapefile* hemos creado una base de datos espacial sobre PostGIS [31] y hemos realizado varias correcciones y mejoras sobre la cartografía.

Aunque la implementación de prueba emplea Geonames y VMap, el sistema ha sido diseñado de manera que esos componentes sean fácilmente intercambiables. Todos los accesos a esos componentes se realizan mediante interfaces genéricas que pueden ser fácilmente implementadas por otros componentes.

Las operaciones espaciales definidas por la jerarquía de la arquitectura combinan ambos servicios para georreferenciar nombres de lugar. Cada nivel contiene una conexión con el nomenclátor y otra con el proveedor de geometrías para recuperar los datos necesarios para la ejecución del algoritmo. En otras palabras, las subclases en esta jerarquía redefinen los métodos abstractos de la superclase para implementar consultas reales sobre ambos servicios.

Además, el algoritmo para obtener las georreferencias se lleva a cabo en dos pasos cada uno de ellos empleando una de las cadenas de responsabilidad definidas en la estructura jerárquica. En el primer paso, cada nivel obtiene del nomenclátor todas las localizaciones con el nombre de lugar consultado. Después de esto, en el

segundo paso, el sistema construye la ruta completa de la descripción geográfica desde abajo hacia arriba. Por ejemplo, si el nombre de lugar consultado es *Londres*, en el primer paso el sistema obtiene al menos dos localidades con ese nombre. Después de eso, en el segundo paso el sistema devuelve dos rutas con las descripciones geográficas de *Europa*, *Reino Unido*, *Inglaterra*, *Londres* y de *América del Norte*, *Canadá*, *Ontario*, *Londres*. Finalmente, para elaborar un ranking de relevancia de los resultados (o para devolver el resultado más relevante) el algoritmo computa una medida de relevancia para cada resultado. Esta medida combina la longitud de la ruta que describe el topónimo en un modelo del mundo, la población del lugar, un factor de pesado que pondera si el lugar es capital, ciudad principal, etc. La mayor parte de estos datos provienen del nomenclátor.

La figura 2 presenta algunos ejemplos que emplean el *plugin* genérico de WPS para JUMP [32] desarrollado por *52 North*. JUMP proporciona un interfaz de usuario gráfico para visualizar y manipular conjuntos de datos espaciales. La arquitectura de esta herramienta es muy extensible y define un mecanismo de extensión basado en *plugins*. *52 North* desarrolló un *plugin* para JUMP que implementa un cliente WPS genérico. En este trabajo empleamos este *framework* para comprobar de forma visual el funcionamiento del WPS de *Resolución de Topónimos*.

En la figura se puede ver el resultado de varias peticiones de la operación *getMostProbable*. Todas estas peticiones fueron ejecutadas con el parámetro *full_path* establecido a *false*. Este parámetro se debe establecer a *true* para que el WPS devuelva la descripción geográfica de todos los nodos de la ruta que representa completamente el topónimo en el mundo (continente, país, etc.). Además, todas las peticiones, excepto la de capa nombrada *SICHUAN(BBox)*, fueron ejecutadas con el parámetro *bounding_box* establecido a *false*. Si este parámetro se establece a *true* el WPS devuelve el *bounding box* en lugar de la geometría real. Los nombres de lugar consultados que se muestran en la figura son, de abajo a arriba, *China* (un país), *Sichuan* (una provincia de China), el *bounding box* de esa provincia, *Qinghai* (una provincia de China), *Xining* (la capital de Qinghai) y Shanghai (otra ciudad en China).

6 Conclusiones y Trabajo Futuro

En este artículo se presenta un sistema para realizar *Resolución de Topónimos*. La interfaz de este sistema define dos operaciones espaciales *getAll* y

getMostProbable. La primera de ellas devuelve todas las descripciones geográficas con el nombre de lugar consultado ordenadas según un ranking de relevancia. La segunda filtra ese resultado devolviendo sólo la descripción geográfica más relevante con el nombre de lugar consultado. Además, ambas operaciones pueden ser parametrizadas. El parámetro *bbox* se emplea para obtener el *bounding box* de las geometrías en lugar de las geometrías reales. El parámetro *full_path* se emplea para obtener la ruta completa que representa el nombre de lugar consultado en el mundo en lugar de únicamente el nodo hoja de esa ruta. Además, siguiendo la tendencia actual en el campo de los GIS, hemos desarrollado un *Web Processing Service* (WPS) para ofrecer ambas operaciones como procesos que pueden ser ejecutados a través de Internet.

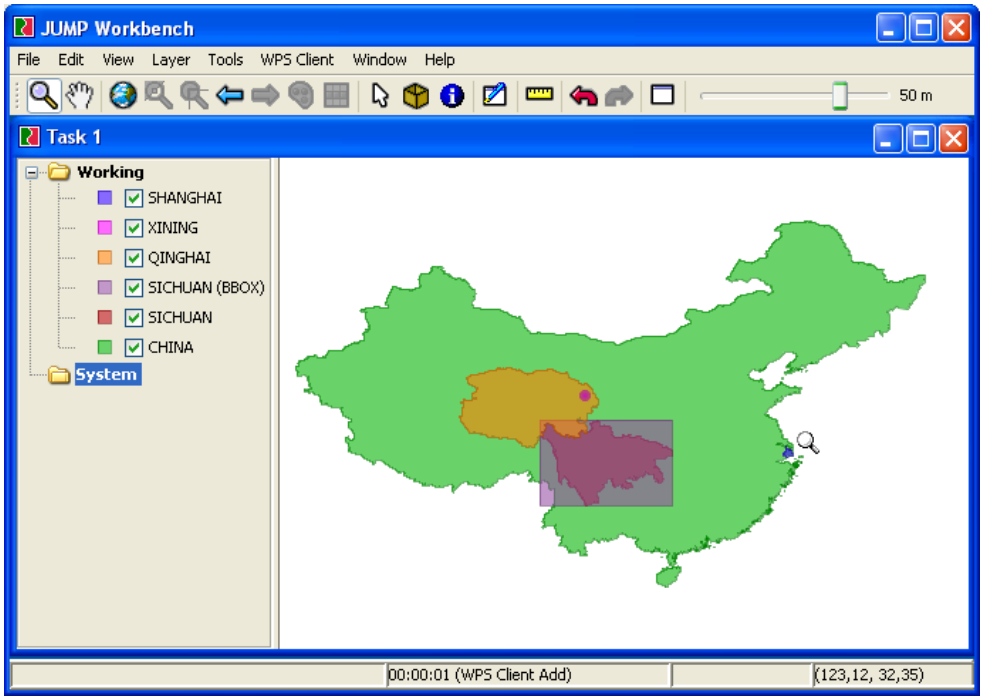

Figura 2. Ejemplos de resultados empleando el plugin de WPS para JUMP

Contemplamos la posibilidad de realizar futras mejoras sobre el servicio WPS desarrollado. Muchas veces, las características intrínsecas de los topónimos (tales como la población o el nivel administrativo) no son suficientes para decidir el

resultado más relevante en un cierto *contexto*. Por ejemplo, si el nombre de lugar *Santiago* aparece en el texto de un documento con otros nombres de lugar como *Atacama* o *Magallanes* lo más probable es que el documento se refiera a regiones de *Chile*. Sin embargo, si *Santiago* aparece con *Madrid* o *Barcelona*, lo más probable es que se refiera a lugares de *España*. Actualmente estamos trabajando en una operación que se pueda invocar con más de un nombre de lugar. El resultado de esta operación debe ser la descripción geográfica más probable para cada uno de esos nombres de lugar. Esta operación puede ser de gran utilidad en el campo de investigación en *Recuperación de Información Geográfica* (GIR). Por tanto, otra línea de trabajo futuro involucra la integración de este WPS en la arquitectura de sistemas GIR. Además, hay varios cambios en los algoritmos que son necesarios para mejorar el rendimiento del sistema. Finalmente, tenemos planteado explorar el uso de otros servicios de nomenclátor y cartografías para determinar como estas fuentes de datos afectan al rendimiento del sistema.

Referencias

[1] Worboys, M. F. (2004) GIS: A Computing Perspectiva. CRC. ISBN: 0415283752.

[2] ISO/IEC (2002) Geographic Information – Reference Model. International Standard 19101.

[3] Open GIS Consortium, Inc. (2003) OpenGIS referente Model. OpenGIS Project Document 03-040, OpenGIS Consortium, Inc.

[4] Global Spatial Data Infrastrucutre Association (2007) WWW document, http://www.gsdi.org.

[5] Open Geospatial Consortium, Inc. (2006) Vision and Mission. WWW document, http://opengeospatial.org/about/?page=vision.

[6] GIS Competitors Cooperate on OpenGIS Specs (1997) Information Today, Vol. 14(2):15.

[7] Open Geospatial Consortium, Inc. (2007) Web Processing Service (WPS) Specification. WWW document, http://portal.opengeospatial.org/files/?artifact_id=24151.

[8] World Wide Web Consortium (2006) Extensible Markup Language (XML). WWW document, http://www.w3.org/XML/.

[9] Leidner, J. L. (2004): Toponym Resolution in Text: “Which Sheffield is it?”. In Proc. of the 27th Annual Internacional ACM SIGIR Conference on Research and Development in Information Retrieval.

[10] Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld,

M., Weibel, R. (2002) Spatial Information Retrieval and Geographical Ontologies: an Overview of the SPIRIT Project. In Proc. of the 25th Annual Internacional ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 387-388.

[11] Baeza-Yates, R., Ribeiro-Neto, B. (1999) Modern Information Retrieval. Addison Wesley.

[12] Zheng, G., Su, J. (2002) Named entity tagging using an HMM-based Cheng tagger. In Proc. of the 40th Annual Meeting of the Association for Computacional Linguistics, pp. 209-219.

[13] Luaces, M. R., Paramá, J. R., Pedreira, O., Seco, D. (2008) An ontology-based index to retrieve documents with geographic information. In Proc. of the 20th Internacional Conference on Scientific and Statistical Database Management (SSDBM08).

[14] Open Geospatial Consortium, Inc. (2006) Gazetteer Profile of WFS (WFS-G) Specification. WWW document, http://www.opengeospatial.org/standards/requests/36.

[15] Modelo de Nomenclátor de España v1.9 (2005) WWW document, http://www.idee.es/resources/recomendacionesCSG/Propuesta_MNE_v1.0.pdf.

[16] Alexandria Digital Library (2007) WWW document, http://www.alexandria.ucsb.edu/.

[17] Getty Thesaurus of Geographic Names (2007) WWW document, http://www.getty.edu/research/conducting_research/vocabularies/tgn/index.html.

[18] Geonames Gazetteer (2007). WWW document, http://www.geonames.org.

[19] FAO (2008) Global Administrative Unit Layers (GAUL). WWW document, http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691.

[20] National Imagery and Mapping Agency (NIMA): Vector Map Level 0 (2007) WWW document, http://www.mapavility.com.

[21] Amitay, E., Har'El, N., Sivan, R., Soffer, A. (2204) Web-a-where: geotagging web content. In Proc. of 27th Annual International ACM SIGIR, pp. 273-280.

[22] Rauch, E., Bukatin, M., Baker, K. (2003) A confidence-based framework for disambiguating geographic terms. In Proc. of the HLT-NAACL 2003 workshop on Analysis of geographic referentes, pp. 50-54.

[23] Lieberman, M. D., Samet, H., Sankaranarayanan, J., Sperling, J. (2007) STEWARD: Architecture of a Spatio-Textual Search Engine. In Proc. of the 15th ACM International Symp. on Advances in Geographic Infomation Systems (ACMGIS07), pp. 186-193.

[24] Michaelis, C.D., Ames, D.P. (2008) Evaluation and implementation of the OGC Web Processing Service for use in client-side GIS. Geoinformatica.

[25] Cepický, J. (2008) OGC Web Processing Service and it's usage. In Proc. of the 15th International Symposium GIS Ostrava.
[26] Open GIS Consortium, Inc.(2007) OpenGIS Geographic Markup Language (GML)
Encoding Standard. OpenGIS Standard, Open GIS Consortium, Inc.
[27] 52 North (2007) Geoprocessing. Retrieved December 2007 from http://52north.org/.
[28] Foerster, T., Stoter, J. (2006) Establishing an OGC Web Processing Service for generalization process. In Proc. of the Workshop of the ICA Commission on Map Generalisation and Multiple Representation.
[29] Gamma, E., Helm, R., Johnson, R., Vlissides, J. (1996) Design Patterns: Elements of Reusable Object-oriented Software. Addison-Wesley.
[30] FWTools (2007) Open Source GIS Binary Kit for Windows and Linux. Retrieved September 2007 from http://fwtools.maptools.org.
[31] Refractions Research (2007) PostGIS. Retrieved June 2007 from http://postgis.refractions.net.
[32] The JUMP Project (2008) JUMP Unified Mapping Platform. Retrieved January 2008 from http://www.jump-project.org/.