Un Sistema de Gestión Documental y Workflow con Indexación Temática y Geográfica de los Documentos*

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

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

Resumen

Dentro del campo de los Sistemas de Información Geográfica (SIG) se está realizando un trabajo muy importante por parte de muchas organizaciones para la construcción de Infraestructuras de Datos Espaciales (IDEs) que les permitan compartir su información geográfica. En estas IDEs, y en los SIG en general, no sólo se gestiona información geográfica sino que también se deben almacenar y recuperar muchos tipos de documentos con texto (licencias de obra, expedientes, etc.). Para proporcionar un acceso eficiente a este tipo de documentos es necesario contar con estructuras de indexación textual sobre dichos documentos. Además, dentro del texto de los documentos aparecen muchas veces referencias geográficas por lo que la estructura de indexación debe permitir también realizar consultas que tengan en cuenta esas referencias geográficas y sus características especiales debidas a su naturaleza espacial.

En este trabajo presentamos un proceso de *workflow* que permite la construcción de un repositorio de documentos al que se puede acceder de manera eficiente realizando consultas acerca tanto del texto de los documentos como de las referencias geográficas citadas en dichos textos. Además, se describe brevemente la estructura de indexación que permite resolver dichas consultas y que combina un índice textual, un índice espacial y una ontología del espacio.

Palabras clave: GIR, *workflow*, estructura de indexación, ontología.

* 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, y por la "Xunta de Galicia" ref. 2006/4 y ref. 08SIN009CT.

1 Introducción

Los Sistemas de Información Geográfica (GIS) [1] constituyen un campo de investigación muy activo hoy en día. Fruto de los resultados de investigación en este campo muchas organizaciones están trabajando en la construcción de *Infraestructuras de Datos Espaciales* (IDES) [2] para compartir su información geográfica. Sin embargo, en las IDES no sólo se almacena y gestiona información geográfica sino también información documental. Existen una gran cantidad de documentos (licencias de obra, expedientes, etc.) que deben ser gestionados en las IDES y a los que se debe acceder de manera frecuente. Por tanto, es necesario emplear estructuras de indexación que permitan un acceso eficiente a esos documentos.

Por otra parte, la *Recuperación de Información* (IR) [3] es un campo de investigación ya clásico que debido al crecimiento de Internet y de la World Wide Web se ha convertido en uno de los campos de investigación más importantes hoy en día. Como resultado de la investigación en este campo se han desarrollado una gran cantidad de bibliotecas digitales, sistemas de gestión documental, estructuras de indexación y técnicas que permiten un acceso eficiente a las colecciones de documentos. Muchos de los documentos almacenados en estas bibliotecas digitales y bases de datos documentales contienen referencias geográficas. Sin embargo, en el campo de la IR no se están teniendo en cuenta las características especiales de esas referencias y que son debidas a la naturaleza espacial de las mismas.

Estos dos campos de investigación avanzaron de manera independiente durante muchos años. Como resultado existen una gran cantidad de estructuras de indexación textual propuestas desde el campo de la IR y una gran cantidad de estructuras de indexación espacial propuestas desde el campo de los GIS. Sin embargo, ninguna de estas estructuras es apropiada cuando se trata de realizar un acceso eficiente a los documentos teniendo en cuenta tanto su texto como las referencias geográficas contenidas en el mismo. Para solventar este problema surgió un nuevo campo de investigación en la confluencia de los dos anteriores, la *Recuperación de Información Geográfica* (GIR) [4]. 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*.

En [5,6] presentamos una arquitectura de sistema y una estructura de indexación que permite acceder a los documentos de una colección empleando tanto el ámbito textual como el geográfico de los mismos. Sin embargo, la arquitectura presentada

es demasiado rígida para poder ser empleada en organismos e instituciones donde la colección de documentos crece constantemente. Aunque la estructura de indexación tiene operaciones de actualización, la arquitectura está pensada para construir un índice sobre la colección de documentos y posteriormente explotarlo realizando consultas sobre el mismo. Para organismos como ayuntamientos o diputaciones donde se generan a diario nuevas licencias de obra o expedientes que deben ser indexados es necesario implementar un proceso de de *workflow* que defina todas las tareas que se tienen que realizar sobre los nuevos documentos para que pasen a formar parte del sistema de gestión documental. Además, este proceso de *workflow* debe contemplar otras tareas previas a la indexación como pueden ser el escaneado de documentos, el reconocimiento automático de su texto, etc. que no están contempladas en la arquitectura ya que ésta supone que se parte de una colección de documentos de la que ya se dispone de su texto.

El problema de definir procesos de *workflow* para alimentar las bibliotecas digitales con documentos, incluso cuando la colección de documentos puede ser enorme o crecer continuamente, está bien estudiado en el campo de la IR. En [7] presentamos un sistema de *workflow* para la alimentación masiva de bibliotecas digitales y describimos un caso de estudio de éxito para la creación de la Biblioteca Virtual Galega (BVG) [8]. Sin embargo, este proceso de *workflow* no es suficiente cuando se quieren explotar las referencias geográficas (y las características especiales derivadas de su naturaleza espacial) contenidas en el texto de los documentos.

Por tanto, el objetivo de este trabajo es definir un conjunto de estrategias de *workflow* y una arquitectura general de sistema que permita la creación de un sistema de gestión documental que proporcione un acceso eficiente a los documentos ante consultas textuales, espaciales e híbridas (por ejemplo, licencias *de obra de edificios civiles en la provincia de A Coruña*). El proceso de *workflow* define todas las fases por las que tiene que pasar un documento para formar parte de la colección de documentos indexada. Además, las estrategias propuestas mejoran el rendimiento general de este proceso y aseguran que todas las tareas necesarias se realizan correctamente y facilita el trabajo de las personas dedicadas a su realización.

El resto del artículo está organizado de la siguiente manera. En primer lugar, presentamos algún trabajo relacionado en la Sección 2. Luego, la Sección 3 presenta la arquitectura del sistema y el proceso de *workflow* definido. En la Sección 4, describimos brevemente la estructura de indexación empleada y los tipos de consulta que nos permite responder. Finalmente, en la Sección 5

presentamos algunas conclusiones y futuras líneas de trabajo.

2 Trabajo relacionado

Los *índices invertidos* se consideran la técnica de indexación de texto clásica. Un índice invertido asocia a cada palabra en el texto (organizado como un vocabulario) la lista de punteros a las posiciones donde la palabra aparece en los documentos. El conjunto de todas las listas se llama *ocurrencias* [3]. El principal inconveniente de esta técnica es que ignora por completo las referencias geográficas. Los nombres de lugar son considerados simplemente como palabras. Si el usuario realiza una consulta del tipo *hoteles en España*, el nombre de lugar *España* es considerado una palabra y sólo se recuperan aquellos documentos que contengan esa palabra. Sin embargo, un documento que contenga nombres de ciudades de *España* pero no la palabra exacta *España* no se recupera porque no se ajusta a la consulta textual.

En cuanto a la indexación de información geográfica, se han propuesto una gran variedad de estructuras de indexación espacial a lo largo de los años. En [9] se puede encontrar un buen resumen de esas estructuras. El objetivo principal de las estructuras de indexación espacial es mejorar el tiempo de acceso a las colecciones de objetos con datos geográficos. Una de las estructuras de indexación espacial más populares y un ejemplo paradigmático es el R-Tree [10]. Un inconveniente de estas estructuras es que no tienen en cuenta la jerarquía del espacio. Los nodos internos en la estructura carecen de significado en el mundo real, sólo tienen significado para la estructura de indexación. Por ejemplo, supongamos que queremos construir un índice para una colección de países, provincias y ciudades. Estos objetos están estructurados en una relación topológica de contenido, esto es, una ciudad está contenida en una provincia que a su vez lo está en un país. Si nosotros construimos un R-Tree con estos objetos geográficos la jerarquía de contenidos no se mantendrá. Los nodos internos del R-Tree no representan provincias o países y, por tanto, el índice no mantiene la jerarquía del espacio. No se puede asociar información al nodo de una provincia y que las ciudades que contiene hereden esa información porque no existe ninguna relación entre una provincia y sus ciudades en la estructura del R-Tree.

Se han realizado algunos trabajos para tratar de combinar ambos tipos de índices dentro del campo de investigación en GIR. Los artículos sobre el proyecto SPIRIT (*Spatially-Aware Information Retrieval on the Internet*) [4] son un buen punto de comienzo en el campo ya que presentan aproximaciones introductorias. Más avanzados son los sistemas Web-a-where [11], MetaCarta [12] (es un sistema

comercial) y STEWARD [13]. En cuanto a nuestro trabajo en el campo, en [5,6] presentamos una arquitectura de sistema y una estructura de indexación para indexar los documentos tanto de manera temática como geográfica. La estructura de indexación descrita combina un índice textual, un índice espacial y una ontología del espacio geográfico. Dicha estructura permite responder consultas puramente textuales, puramente espaciales y consultas híbridas. Puede encontrarse más información acerca de la misma en la Sección 4.

En cuanto al trabajo en sistemas de gestión documental y procesos de *workflow*, en [7] se presenta *DigiFlow* que es una herramienta para la construcción de repositorios de documentos. Esta herramienta proporciona un entorno integrado donde se pueden ejecutar todas las tareas necesarias para la construcción del sistema de gestión documental, desde la tarea de almacenamiento de metadatos, escaneado de documentos, etc. hasta la consulta del repositorio. Como ya se ha descrito en la sección anterior el objetivo de este trabajo es extender el proceso de *workflow* definido por *DigiFlow* para incluir las tareas necesarias que permitan tener en cuenta la naturaleza especial de las referencias geográficas citadas en el texto de los documentos.

3 Arquitectura del Sistema

Un proceso de *workflow* consiste en la *automatización de los procedimientos donde los documentos, información o tareas se pasan entre los participantes en el proceso siguiendo un conjunto prefijado de reglas para lograr, o contribuir, a un objetivo de negocio global* [14]. Los procesos de *workflow* pueden ser clasificados en varios tipos dependiendo de su naturaleza y de las características del proceso [15,16]. Los procesos de *workflow* colaborativo automatizan los procesos de negocio donde un grupo de gente participa para lograr un objetivo global. Este tipo de procesos de negocio involucra una cadena de actividades donde los documentos, que contienen la información, se procesan y transforman hasta lograr el objetivo global. En este trabajo nos basamos en la arquitectura de sistema recomendada para este modelo de proceso de *workflow* ya que el problema de la creación de un repositorio de documentos se ajusta a la perfección a dicho modelo.

En general, podemos diferenciar tres tipos de perfiles de usuario involucrados en la creación del repositorio:

- *Administrador*. Los administradores son los encargados del proceso completo de creación del repositorio. Son los responsables de asignar

tareas a los diferentes trabajadores y controlar el estado de cada documento.

- *Usuarios avanzados*. Los usuarios avanzados se encargan de realizar tareas críticas como el almacenamiento de metadatos o la revisión de los textos extraídos en el proceso de OCR.
- *Usuarios estándar*. Los usuarios estándar son los trabajadores que se encargan de tareas como el escaneado o la corrección OCR. Este rol lo desempeñan usuarios con algún conocimiento sobre la naturaleza de los documentos pero sin ninguna responsabilidad en el mantenimiento del sistema.

La figura 1 muestra la arquitectura general del sistema. Para definirla se siguieron las recomendaciones del *Workflow Reference Model* [15] que es un modelo ampliamente aceptado para el diseño y desarrollo de sistemas de gestión de *workflow*. El objetivo principal de este modelo es proporcionar un marco de trabajo general para la gran variedad de técnicas de implementación y entornos operacionales que caracterizan a esta tecnología. Por tanto, la arquitectura propuesta pude ser empleada en gran cantidad de entornos y situaciones.

Como se puede ver en la figura 1, el módulo de identificación y autorización se encarga de la autenticación de los trabajadores que quieran acceder al sistema. Cada usuario tiene un rol en el sistema dependiendo de las tareas que tiene que realizar. En función de este rol, el módulo de autenticación le proporcionará al usuario las características y herramientas necesarias para la realización de su trabajo. La arquitectura del sistema se compone de un módulo por cada actividad que se debe llevar a cabo para la creación del repositorio de documentos.

- *Almacenamiento de metadatos*. Este subsistema se encarga de la introducción y almacenamiento de los metadatos de cada documento (título, autor, año, fuente, etc.) según el formato deseado, como Dublin Core [17] o MARC [18]. Esta tarea la llevan a cabo los usuarios avanzados del sistema y, por tanto, sólo ellos tienen acceso a este módulo.
- *Escaneado*. El sistema proporciona acceso al hardware y software necesario para escanear los documentos y es el encargado de gestionar la especificación de los parámetros de escaneado para cada documento (por ejemplo, opciones como escanear dos páginas al mismo tiempo, orientación horizontal de las páginas, resolución, número de colores, etc.).
- *OCR*. Proporciona acceso al software de OCR que permite reconocer el texto de los documentos y lo almacena de manera automática.

- *Corrección*. Este módulo le proporciona al revisor tanto la imagen del documento escaneado como el texto extraído automáticamente para poder llevar a cabo las correcciones que sean necesarias.

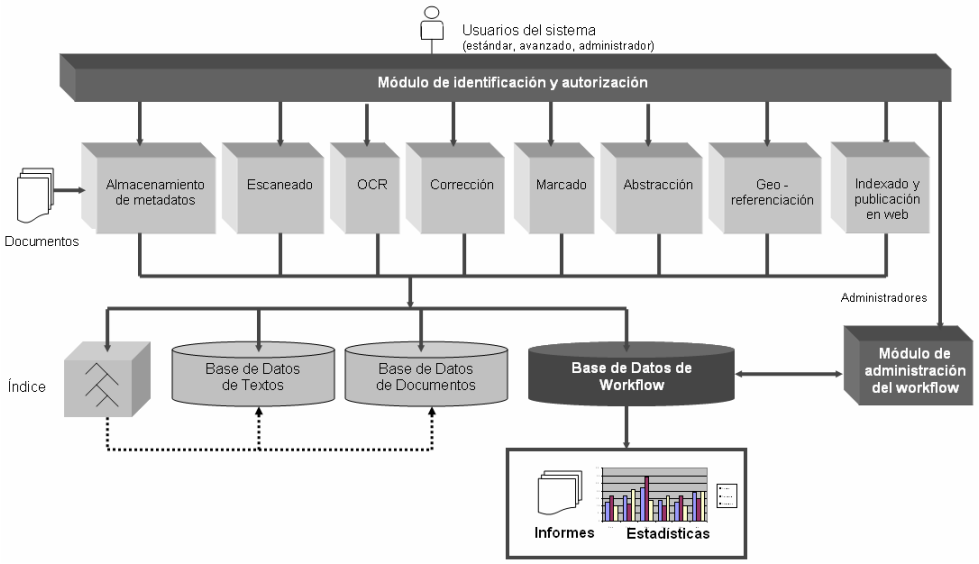


Figura 1. Arquitectura del sistema

- *Marcado*. Proporciona las herramientas necesarias para marcar el texto con metadatos como título, autor, página, etc.
- *Abstracción*. El sistema debe ser capaz de trabajar con textos marcados con diferentes conjuntos de metadatos (por ejemplo, el autor puede no estar disponible en todos los documentos a indexar). Para resolver este problema, hemos definido una abstracción de documentos que representa un *documento* como un conjunto de *campos* cada uno de ellos obtenido del texto marcado en la etapa anterior.
- *Geo-referenciación*. El módulo de geo-referenciación de documentos se encarga de anotar cada documento con las referencias geográficas que contiene y de su traducción a un modelo geográfico (por ejemplo, coordenadas en latitud/longitud, tipo de referencia geográfica, etc.). Existen varias técnicas que permiten automatizar esta tarea. Sin embargo están todavía en fase experimental y no obtienen tan buenos resultados como una anotación manual. Por tanto, el sistema debe permitir tanto una anotación manual de los documentos como utilizar dichas técnicas para

realizar una anotación asistida.

- *Indexado y publicación en web*. Una vez que se acepta el documento, este módulo se encarga de indexar su contenido, empleando técnicas de recuperación de información geográfica, y de su publicación en web.
- *Módulo de administración del workflow*. Este subsistema se encarga de la gestión del flujo de trabajo entre todas las actividades involucradas en el proceso de construcción del repositorio de documentos. Además, también proporciona herramientas de generación de informes para propósitos de monitorización del sistema.

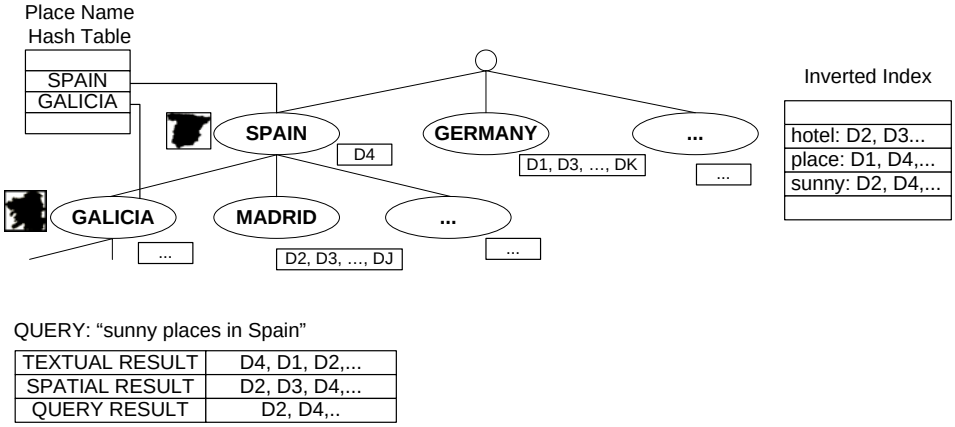
La arquitectura del sistema asume el empleo de diferentes repositorios y bases de datos. Las bases de datos de textos y documentos almacenan los textos extraídos de los documentos y los propios documentos. Sobre estas bases de datos se construye un índice para proporcionar una búsqueda de información eficiente. Este índice tiene en cuenta tanto el texto de los documentos como las referencias geográficas que se citan en dicho texto. En la siguiente sección se describe brevemente el índice y los tipos de consulta que permite responder. Finalmente, la base de datos de *workflow* almacena información sobre la cadena de digitalización, con las listas de tareas, el estado de cada documento, etc.

4 Indexación y Consultas Soportadas

En [5,6] presentamos una estructura de indexación que tiene en consideración no sólo el texto contenido en los documentos sino también las referencias geográficas incluidas en ese texto y la ontología del espacio geográfico. En esta sección vamos a describir brevemente esa estructura de indexación y los tipos de consulta que permite responder. La figura 2 muestra la estructura de indexación anotada con un ejemplo de consulta (*sunny places in Spain*). La base de esta estructura es una ontología espacial. Esta ontología modela tanto el vocabulario como la estructura espacial de las localizaciones geográficas para procesos de recuperación de información. La estructura de una ontología es fija por lo que la estructura de indexación debe ser construida *ad-hoc* para el dominio en el cual se va a emplear.

El componente principal de la estructura de indexación es un árbol compuesto por nodos que representan nombres de lugar. Estos nodos están interconectados por medio de relaciones de contenido espacial (por ejemplo, Galicia está contenida en España). En cada nodo almacenamos: (i) la palabra clave (un nombre de lugar), (ii) las referencias geográficas asociadas con el nombre de lugar, (iii) el *minimum*

bounding rectangle de la geometría que representa ese lugar, (iv) una lista con los identificadores de los documentos que incluyen referencias geográficas a ese lugar y (v) una lista de nodos hijos que están geográficamente contenidos en ese nodo. Además, para mejorar la eficiencia de la parte espacial de las consultas se emplea un R-Tree en cada nodo para el acceso a la lista de nodos hijo.



QUERY: “sunny places in Spain”

TEXTUAL RESULT	D4, D1, D2,...
SPATIAL RESULT	D2, D3, D4,...
QUERY RESULT	D2, D4,..

Figura 2. Estructura de Indexación

En el índice se emplean dos estructuras auxiliares. En primer lugar, una *tabla hash* almacena para cada nombre de lugar su posición en la estructura de indexación. Esto proporciona un acceso directo a un nodo concreto por medio de una palabra clave que se obtiene mediante un servicio de nomenclátor si la palabra procesada es un nombre de lugar. La segunda estructura auxiliar es un índice invertido tradicional con todas las palabras de los documentos que se emplea para resolver la parte textual de las consultas.

Mantener separados el índice textual del índice espacial tiene muchas ventajas. En primer lugar, todas las consultas textuales pueden ser procesadas de manera eficiente por el índice invertido y todas las consultas espaciales pueden ser procesadas de manera eficiente por el índice espacial. Además, el sistema soporta consultas que combinen aspectos textuales con espaciales. Del mismo modo, se pueden manejar de manera independiente las actualizaciones en cada uno de los índices, esto hace que se puedan añadir o eliminar datos de forma sencilla. Finalmente, se pueden aplicar optimizaciones específicas a cada estructura de indexación de manera individual. Por el contrario, los principales inconvenientes de esta estructura son: (i) el árbol que soporta la estructura es posiblemente

desbalanceado, lo cual penaliza la eficiencia del sistema y (ii) las ontologías tienen una estructura fija y, por tanto, nuestra estructura es estática y debe ser construida *ad-hoc*. En [5] se presentan unos experimentos realizados para demostrar que el desbalanceo del árbol no supone un problema grave.

Finalmente, la característica más importante de una estructura de indexación es el tipo de las consultas que se pueden resolver con él. Los siguientes tipos de consultas son relevantes en un sistema de recuperación de información geográfica:

- *Consultas puramente textuales*. Estas son consultas del tipo "*recuperar todos los documentos donde aparezcan las palabras* hotel *y* mar".
- *Consultas puramente espaciales*. Un ejemplo de este tipo de consultas es "*recuperar todos los documentos que se refieran a la siguiente área geográfica*". El área geográfica en la consulta puede ser un punto, una ventana de consulta, o incluso un objeto complejo como un polígono.
- *Consultas textuales con nombres de lugar*. En este tipo de consultas, algunas palabras son nombres de lugar. Por ejemplo, "*recuperar todos los documentos con la palabra* hotel *referidos a* España".
- *Consultas textuales sobre un área geográfica*. En este caso se proporciona un área geográfica de interés junto con el conjunto de palabras. Un ejemplo es "*recuperar todos los documentos con la palabra* hotel *que se refieren a la siguiente área geográfica*". Al igual que en las *consultas puramente espaciales* el área geográfica de la consulta puede ser un punto, una ventana de consulta o un objeto complejo.

La estructura de indexación puede resolver consultas puramente textuales recuperando del índice invertido la lista de los documentos asociados con cada palabra y luego realizando la intersección de las listas. Las consultas puramente espaciales se pueden resolver empleando el índice espacial descendiendo en la estructura teniendo en cuenta sólo aquellos nodos cuyos *bounding box* intersecan con el área geográfica de la consulta. Esta operación devuelve un conjunto de documentos candidatos que tiene que ser refinado con la referencia geográfica actual para decidir si el documento es parte del resultado o no.

Por tanto, las consultas puramente textuales se pueden resolver en nuestro sistema porque un índice textual forma parte de la estructura de indexación y las consultas puramente espaciales se pueden resolver porque la estructura de indexación es construida como un índice espacial. Sin embargo, la estructura de indexación que proponemos puede ser usada para resolver el tercer y el cuarto tipo de consultas que no pueden ser solucionados de manera sencilla empleando un índice textual y

un índice espacial. Para el caso de la consulta con nombres de lugar, nuestro sistema puede descubrir que *España* es una referencia geográfica consultando a un servicio de nomenclátor y posteriormente emplear la *tabla hash* de nombres de lugar de la estructura para recuperar el nodo del índice que representa *España*. De este modo se puede ahorrar algún tiempo de acceso suprimiendo parte del recorrido en el árbol. La figura 2 está anotada con un ejemplo de este tipo de consultas (*sunny places in Spain*). Cuando el usuario realiza una consulta con el texto *sunny places* y el nombre de lugar *Spain*, se emplea el índice textual para recuperar la lista de documentos que contienen esas palabras (*textual result*) y el índice espacial para obtener la lista de documentos que se refieran al área geográfica (*spatial result*). Esas dos listas se pueden ver en la parte inferior de la figura. Luego, el resultado de la consulta (*query result*) se obtiene como la intersección de las listas de resultado textual y espacial.

Con respecto al cuarto tipo de consultas, el índice invertido se emplea para recuperar la lista de documentos que contienen las palabras y la estructura de indexación se emplea para obtener la lista de documentos que hacen referencia al área geográfica. Por tanto, la intersección de ambas listas es el resultado de la consulta.

Otra mejora sobre los índices textuales y espaciales es que nuestra estructura de indexación puede realizar fácilmente expansión de los términos de consulta (*query expansion*) sobre referencias geográficas porque está construida sobre una ontología del espacio geográfico. Consideremos como se resuelve la siguiente consulta "*recuperar todos los documentos que se refieran a España*". El servicio de evaluación de consultas descubrirá que España es una referencia geográfica. La tabla hash de nombres de lugar se empleará para localizar rápidamente el nodo interno que representa el objeto geográfico *España*. Entonces todos los documentos asociados con este nodo forman parte del resultado de la consulta. Sin embargo, todos los hijos de este nodo son objetos geográficos que están contenidos en España (por ejemplo, la ciudad de Madrid). De este modo, todos los documentos referenciados por el subárbol forman también parte del resultado de la consulta. La consecuencia es que la estructura de indexación ha sido empleada para expandir la consulta porque el resultado contiene no sólo aquellos documentos que incluyen el término *España*, sino también aquellos documentos que incluyen el nombre de un objeto geográfico contenido en España (por ejemplo, todas las ciudades y regiones de España).

5 Conclusiones y Trabajo Futuro

La creación de un repositorio de documentos no es un proceso sencillo. Necesita la coordinación de un equipo humano y de herramientas para llevar a cabo todas las actividades que forman parte del proceso. Este proceso se complica aún más cuando se quieren explotar las referencias geográficas contenidas en los textos de los documentos y su naturaleza espacial. Las actividades que forman parte del proceso son la digitalización de documentos, extracción del texto, corrección, anotación de los documentos, indexación, etc. Para que todo este proceso sea realizado correctamente y de forma eficiente, se necesita el empleo de herramientas de soporte que faciliten el trabajo de cada participante y que aseguren la calidad de los resultados obtenidos.

Las estrategias de *workflow* propuestas y la arquitectura del sistema soportan el control y la coordinación del equipo humano y de las tareas involucradas en la creación y actualización constante del repositorio. El uso de esta arquitectura automatiza la realización de actividades propensas a errores y optimiza el rendimiento del proceso de construcción del repositorio, mejorando la calidad de los resultados obtenidos.

Además, la estructura de indexación construida sobre la colección de documentos y textos combina un índice textual, un índice espacial y una ontología del espacio geográfico. Esta estructura de indexación permite resolver nuevos tipos de consultas que tienen en consideración no sólo el texto contenido en los documentos sino también las referencias geográficas contenidas en esos textos y las características especiales de esas referencias debidas a su naturaleza espacial.

Actualmente estamos finalizando el prototipo del sistema. Una vez finalizado pretendemos aplicarlo en un dominio real para comprobar su verdadero rendimiento y potencial. Otras líneas de trabajo futuro tienen en consideración mejoras del proceso de *workflow* en general y de la estructura de indexación en particular. En primer lugar, está planificado incluir otros tipos de relaciones espaciales en la estructura de indexación complementarias a las de inclusión (por ejemplo, adyacencia). Estas relaciones pueden ser fácilmente representadas en la ontología y la estructura de indexación puede ser extendida para soportarlas. Otra línea de trabajo futuro implica la introducción de técnicas de *Resolución de Topónimos* para mejorar la tarea de geo-referenciación de documentos. Finalmente, es necesario definir algoritmos para elaborar el ranking de los documentos recuperados por el sistema. Para esta tarea debemos definir una medida de

relevancia espacial y combinarla con la relevancia obtenida empleando el índice textual.

Referencias

[1] Worboys MF (1995) GIS: A Computing Perspective. Taylor & Francis. ISBN: 0-7484-0065-6.

[2] GSDI (2007) Global Spatial Data Infrastructure Association, http://www.gsdi.org.

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

[4] 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.

[5] 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).

[6] Luaces, M. R., Paramá, J. R., Pedreira, O., Seco, D. (2008) LBD LOCAL: Un Sistema para la Recuperación de Documentos con Referencias Geográficas. En Actas de las II Jornadas de SIG Libre, Girona, España.

[7] Places, A. S., Brisaboa, N. R., Paramá, J. R., Pedreira, O., Seco, D. (2007) Managing the workflow of massive feeding of digital libraries. Research in Computer Science, 32, pp. 352-362. Mexico.

[8] BVG (2007) Biblioteca Virtual Galega, http://bvg.udc.es/.

[9] Gaede V, Günther O (1998) Multidimensional access methods. ACM Comput. Surv., 30(2), pp. 170-231.

[10] Guttman A (1984) R-Trees: A dynamic index structure for spatial searching. In B. Yormark, editor, SIGMOD'84, Proceedings of Annual Meeting, Boston, Mas-sachusetts, June 18-21, 1984, pp. 47-57. ACM Press.

[11] 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.

[12] 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.

[13] 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.

[14] Hollingsworth, D. (1995) Workflow Management Coalition – the workflow reference model. Technical Report, Workflow Management Coalition.

[15] van der Aalst, W., van Hee, K. (2002) Workflow management: Models, methods, and systems.

[16] Fischer, L. (2003) Workflow handbook 2003. Future Strategies, Inc., USA.

[17] Hillman, D. (2005) Using Dublin Core. Technical Report, Dublin Core Metadata Initiative.

[18] Furrie, B. (2003) Understanding MARC – Bibliographic machina readeble cataloging. Technical Report, Library of Congreso – Network development and MARC standards office.