UNIVERSITAT JAUME I
INSTITUTO DE NUEVAS TECNOLOGÍAS DE LA IMAGEN

Descripción, Publicación y Descubrimiento de Recursos Georreferenciados

Tesis Doctoral

Arturo Beltran Fonollosa

Dirigida por:

Dr. **Joaquín Huerta Guijarro**
Dra. **Laura Díaz Sánchez**

Castellón, junio de 2013

Descripción, Publicación y Descubrimiento de Recursos Georreferenciados

Arturo Beltran Fonollosa

Trabajo realizado bajo la dirección del Doctor D. Joaquín Huerta Guijarro y de la Doctora Dña. Laura Díaz Sánchez. Presentado en el Instituto de Nuevas Tecnologías de la Imagen para optar al grado de Doctor por la Universitat Jaume I.

Castellón, junio de 2013

Este trabajo ha sido parcialmente financiado mediante una beca predoctoral de la Universitat Jaume I (ref. PREDOC/2008/06) y por el proyecto “España Virtual” (ref. CENIT 2008-1030) a través del Instituto Geográfico Nacional (IGN).

A mi familia.

Resumen

El interés de los usuarios en la información con contexto geográfico hace que esta juegue un papel fundamental en la sociedad haciendo que la cantidad y variedad de recursos georreferenciados disponibles en la web aumente día tras día. Actualmente existen numerosos servicios Web especializados en compartir tipos concretos de recursos como imágenes, video o texto que nos permiten realizar búsquedas en base a una localización. Por otra parte, de manera más formal, en el contexto de los Sistemas de Información Geográfica (SIG) se han realizado grandes esfuerzos en generar grandes catálogos de metadatos. Sin embargo, aún resulta complicado encontrar contenidos georreferenciados relevantes de una forma integrada y sencilla.

Tomando como referencia el mundo Web, observamos que ya en su inicio comenzó a ser poblado con recursos de forma masiva, hecho que dificultaba encontrar contenidos relevantes. Tras la solución inicial de los directorios, la revolución llego con los buscadores, empresas como Yahoo! o Google que se dieron cuenta de las deficiencias del sistema y empezaron a recopilar ellos mismos información de cada recurso que encontraban disponible. Estos sistemas se dedican a recorrer sistemáticamente los recursos con el fin de obtener de ellos el máximo de información (metadatos) que su tecnología les permite. En base a estos metadatos se podrán indexar los recursos de una forma exacta y eficiente proporcionando resultados relevantes y exactos a las consultas formuladas por los usuarios. Tal ha sido el éxito de estos sistemas que hoy es inimaginable la búsqueda de información y la navegación por la red sin acceder a alguno de estos servicios.

Sin embargo, en el caso de los recursos georreferenciados el descubrimiento sigue dependiendo de que estos hayan sido publicados de forma no automatizada en servicios especializados como catálogos. Algunos usuarios en ciertas situaciones requieren un rápido acceso a información actualizada. Por ejemplo, en el campo de la gestión de emergencias una buena disponibilidad y un rápido

acceso a información actualizada son muy importantes dado que las primeras horas de respuesta a un desastre son críticas para salvar vidas y reducir daños. Lo que plantea nuevos retos de investigación en el descubrimiento de contenidos georreferenciados.

Partiendo de estos hechos, esta tesis se basa en la premisa de que son necesarios mecanismos más sencillos, flexibles y eficientes que den soporte a la descripción, publicación y descubrimiento de recursos georreferenciados, capaces de ofrecer al usuario una solución integrada que le permita recolectar, catalogar y buscar recursos georreferenciados de una forma natural y homogénea.

Por ello, en esta tesis se revisa todo el ciclo de vida de los recursos georreferenciados, se exponen diferentes conceptos teóricos y técnicos relevantes y se proponen diferentes soluciones. Para la descripción de los recursos se propone una metodología para la generación de metadatos que pretende automatizar su producción. Además, se propone un mecanismo común para anotar y georreferenciar recursos independientemente de su naturaleza. Y también se propone un nuevo paradigma cuyo fin es proporcionar descripciones homogéneas on-line. Por otra parte, se propone una metodología para la publicación de los recursos de forma interoperable en servicios de catálogo de forma integrada en el flujo de trabajo e incluyendo posibles anotaciones semánticas. Además, se propone otro método que permite publicar los recursos mediante la indexación de sus descripciones combinando índices textuales y espaciales. Por otra parte, para mejorar el descubrimiento de los recursos se automatiza su recopilación, se propone una interfaz de consulta común y homogénea, y se proponen aplicaciones cliente basadas en un globo virtual permitiendo la visualización e integración de datos heterogéneos. Todas estas contribuciones conceptuales han sido comprobadas y demostradas mediante sus respectivas implementaciones en casos de uso reales. Finalmente, juntando todas las piezas, se presenta una arquitectura y una primera implementación de un sistema automatizado de indexación y búsqueda de recursos georreferenciados. Mediante este sistema se consigue mejorar el descubrimiento y consecuentemente la accesibilidad a los recursos, nuestro objetivo final.

Palabras Clave: *Descubrimiento, Recursos Georreferenciados, Metadatos, Publicación, Indexación.*

Agradecimientos

Me gustaría que estas líneas sirvieran para expresar mi más profundo y sincero agradecimiento a todas aquellas personas que con su ayuda han colaborado en la realización de esta tesis.

En primer lugar, me gustaría dar las gracias a mis directores de tesis por el apoyo y la orientación que me han brindado durante estos años. Quiero agradecer a Laura Díaz su ayuda para completar la escritura de esta tesis y sus comentarios siempre constructivos y acertados. También me gustaría agradecer a Joaquín Huerta por su atención, por su ayuda para encontrar la forma de seguir adelante y por haber solucionando todos los problemas que han aparecido.

También me gustaría agradecer a Michael Gould el haberme dado la primera oportunidad y el haberme introducido en el mundo de la investigación y de los sistemas de información geográfica. Asimismo, quiero dar las gracias a mis compañeros del grupo de investigación y demás miembros del instituto por estar siempre dispuestos a ayudar y mantener un entorno de trabajo agradable y productivo.

Gracias al comité de expertos que evaluaron esta tesis: Miguel Ángel Manso y Francisco Javier López, cuyos valiosos comentarios han permitido mejorar la calidad de la misma.

Por otra parte, quiero dar las gracias a todos mis amigos por su apoyo y entendimiento a lo largo de estos últimos años y por enriquecer mi vida de muchas formas diferentes.

En último lugar, y no por ello menos importante, quiero agradecer y dedicar este trabajo a mi familia, especialmente a mis padres, por ser el pilar fundamental en todo lo que soy, en toda mi educación, tanto académica, como de la vida, por su incondicional apoyo mantenido perfectamente a través del tiempo. Y a Helena por su paciencia y atención, por estar siempre a mi lado y por ayudarme a ver la parte positiva de todas las cosas.

Índice de Contenidos

Listado de Figuras

1. Introducción

Este capítulo describe la motivación y los objetivos de esta tesis, así como los proyectos que han marcado su desarrollo. El capítulo también describe la metodología llevada a cabo y presenta las principales contribuciones realizadas en esta tesis, para finalizar explicando cómo se estructura el presente documento.

1.1 Motivación y Objetivos

Existen estudios que muestran que actualmente más del 80% de los datos existentes en todo el mundo es susceptible de contener una referencia espacial, ya sea de forma explícita o implícita [79] [136]. El interés y el papel clave que la información con contexto geográfico juega en nuestra sociedad crece día a día, así como el número y variedad de *recursos*[1] georreferenciados disponibles en la red [137] [128].

El contexto colaborativo que impulsa la Web 2.0 [173] [217] ha influido en la forma de compartir recursos en la sociedad. La principal innovación es el cambio en el papel que juegan los usuarios finales, donde ya no son simples consumidores de información sino que se

[1] Cuando hablamos de *recursos* nos referimos a cualquier cosa o entidad que pueda ser identificada, nombrada, enlazada o manejada de algún modo, en Internet o en cualquier sistema de información.

establece un sistema bidireccional donde todos pueden crear y compartir contenido. De esta forma, los usuarios se han vuelto rápidamente proveedores de contenido en esta nueva era multimedia y social identificada por el creciente flujo de servicios Web basados en compartir y el sentido de la comunidad [25]. Dentro de todo este contenido generado por los usuarios, la localización se ha mostrado como el contexto predominante para anotar cualquier tipo de recurso, lo que nos lleva a enormes cantidades de recursos georreferenciados en prácticamente cualquier dominio [70]. Un claro ejemplo de estos hechos es la reciente aparición de numerosos y populares servicios Web basados en la colaboración abierta y distribuida (*crowdsourcing*[2]) que contienen tipos específicos de recursos como imágenes (*Flickr*[3], *Panoramio*[4], *Instagram*[5]), video (*YouTube*[6]) o texto (*Twitter*[7], *Wikipedia*[8]) y que ofrecen búsquedas basadas en la localización.

Por otra parte, en el lado más formal y autoritativo de los recursos georreferenciados encontramos las Infraestructuras de Datos Espaciales (IDE). Estas describen y engloban un conjunto de políticas, acuerdos institucionales y económicos, datos espaciales, servicios geo-espaciales y tecnologías que facilitan la disponibilidad y el acceso a los servicios y a los datos geoespaciales [165] [148] [22] [56] [216] [89]. Durante los últimos años los organismos públicos han realizado grandes esfuerzos para el despliegue de nodos IDE a diferentes escalas con el fin de construir una IDE global [149] [185] que permita mejorar la disponibilidad y el descubrimiento de este tipo de recursos [113].

La geolocalización de cualquier tipo de recurso y su disponibilidad está adquiriendo un papel fundamental en un amplio rango de aplicaciones tanto a nivel popular como a nivel oficial. Esto plantea

---

[2] *Crowdsourcing*, del inglés *crowd* (multitud) y *outsourcing* (externalización) consiste en externalizar tareas que, tradicionalmente, realizaba un empleado o contratista, a un grupo numeroso de personas o una comunidad, a través de una convocatoria abierta. El término se ha hecho popular para referirse a la tendencia a impulsar la colaboración en masa posibilitada por las tecnologías Web 2.0.

[3] http://www.flickr.com

[4] http://www.panoramio.com

[5] http://instagram.com

[6] http://www.youtube.com

[7] https://twitter.com

[8] http://www.wikipedia.org

nuevos retos de investigación en el descubrimiento de contenidos multimedia georreferenciados [196] [160].

La cantidad de información, tanto oficial como creada por los usuarios, crece de forma exponencial lo que nos permite compartir contenidos a gran escala [17]. Sin embargo, esta enorme cantidad de datos existente en diferentes recursos distribuidos hace que su descubrimiento sea una tarea ardua [196]. Además, algunos usuarios en ciertas situaciones requieren un rápido acceso a información actualizada. Por ejemplo, en el campo de la gestión de emergencias una buena disponibilidad y un rápido acceso a información actualizada son muy importantes dado que las primeras horas de respuesta a un desastre son críticas para salvar vidas y reducir daños [63] [162] [238] [239].

Consecuentemente, es de gran importancia ser capaz de descubrir estos datos geoespaciales de acuerdo a su contenido y sus características, por ejemplo su cobertura espacial o su extensión temporal [191]. Además, los usuarios requieren cada vez más la posibilidad de compartir todo tipo de recursos georreferenciados a través de aplicaciones colaborativas geoespaciales como globos virtuales o servicios de visualización de mapas [25]. Es entonces necesario desarrollar mecanismos para permitir a los usuarios encontrar y acceder a estos recursos distribuidos de forma homogénea y eficiente.

Iniciativas como *Shared Environmental Information System*[9] (SEIS) [109], *Single Information Space in Europe for the Environment* (SISE) [109], INSPIRE[10] [113], *Global Earth Observation System of Systems*[11] (GEOSS) [45], o *Copernicus*[12], anteriormente conocido como *Global Monitoring for Environment and Security* (GMES) [101], pretenden dar soporte a los responsables de la toma de decisiones a todos los niveles (incluyendo los ciudadanos) proporcionando datos medioambientales. En este contexto emergen retos para abordar la recogida, análisis, y la distribución de los recursos para facilitar la generación y la comunicación de información sobre el medio ambiente.

[9] http://ec.europa.eu/environment/seis
[10] http://inspire.jrc.ec.europa.eu
[11] http://www.earthobservations.org/geoss.shtml
[12] http://copernicus.eu

Como se deduce del ejemplo de la gestión de emergencias, es destacable la importancia de disponer de una información actualizada, especialmente en situaciones críticas. En estas situaciones multitud de información dinámica es creada. Esta nueva información recientemente generada y actualizada resulta normalmente infrautilizada, dado que es usada en el momento pero rara vez publicada de una forma estructurada e interoperable para que otros puedan aprovecharla [239]. Por lo tanto, resulta necesario desarrollar mecanismos que automaticen la publicación y puesta a disposición de estos recursos tan valiosos en situaciones límite.

Es en este escenario heterogéneo donde las descripciones de los recursos cobran sentido y pasan a ser la pieza clave de cualquier Sistema de Información (SI) [169]. Los metadatos nos permiten describir los recursos en base a sus propiedades, características y contexto. Indexando y catalogando los recursos de acuerdo a sus características (tipo de datos, contenido, origen, calidad, fecha de creación, localización, etc.) y su contexto, permite y facilita su posterior descubrimiento [199].

Partiendo de estos hechos, esta tesis se basa en la premisa de que son necesarios mecanismos más sencillos, flexibles y eficientes que den soporte a la descripción, publicación y descubrimiento de recursos georreferenciados, capaces de ofrecer al usuario una solución integrada que le permita recolectar, catalogar y buscar recursos georreferenciados de una forma natural y homogénea.

Con esto se pretende que la información georreferenciada quede disponible y fácilmente descubrible de forma automatizada. Este objetivo, pese a ser de utilidad general, resulta especialmente interesante en entornos y situaciones críticas como la gestión de emergencias. Grandes desastres naturales ocurridos durante los últimos años, como terremotos, huracanes o incendios, han mostrado que pese a la gran cantidad de recursos espaciales existentes para manejarlos, estos no estaban suficientemente organizados para proporcionar un sistema de análisis y respuesta efectivo. Estos recursos están disponibles desde múltiples fuentes y almacenados en múltiples formatos y es necesario organizarlos para que los científicos y otros usuarios puedan acceder a ellos y explotarlos [92]. Resulta esencial para los científicos y las personas responsables de la toma

de decisiones una disponibilidad global de datos geoespaciales actualizados para extraer información útil y precisa.

1.2 Contexto

El trabajo de esta tesis se ha desarrollado en el marco de los siguientes proyectos:

1.2.1 gvSIG

gvSIG[13] es un proyecto de desarrollo de SIG en software libre impulsado por la Consejería de Infraestructuras y Transportes[14] (CIT) de la Generalitat Valenciana. En el proyecto implementa un programa informático para el manejo de Información Geográfica (IG) con precisión cartográfica que permite acceder a información vectorial y *raster* así como a servidores de mapas que cumplan las especificaciones del *Open Geospatial Consortium*[15] (OGC).

Parte del trabajo de esta tesis se enmarca en la funcionalidad que gvSIG ofrece para la generación semi-automática de metadatos y su posterior publicación en servidores de catálogo. El trabajo fue desarrollado en el contexto de una beca de investigación para llevar a cabo trabajos de investigación y desarrollo en el Departamento de Lenguajes i Sistemas Informáticos de la Universidad Jaume I, a cargo del contrato *"Servicios Informáticos de Implementación Sistema de Extracción de Metadatos en gvSIG y de Seguridad para Acceso a Geodatos a través de WFS en la IDE de la Conselleria de Infraestructuras y Transportes"* (ref. 07I352), financiado por la Conselleria de Infraestructuras y Transportes de la Generalitat Valenciana.

[13] http://www.gvsig.org
[14] http://www.cit.gva.es
[15] http://www.opengeospatial.org

1.2.2 España Virtual

La mayor parte del trabajo de esta tesis se ha desarrollado en el contexto del proyecto *España Virtual*[16] (EV). EV es un proyecto CENIT (ref. 2008-1030), subvencionado por el Centro para el Desarrollo Tecnológico Industrial[17] (CDTI) dentro del programa *Ingenio 2010*, orientado a crear un puente entre el mundo geográfico y las tecnologías de Internet.

Durante los últimos años, el mundo ha vivido una revolución en la manera en que los ciudadanos hacen uso de las tecnologías de información geográfica y 3D. Los satélites de observación de la Tierra, Internet, los dispositivos móviles y las tecnologías 3D y de código abierto han universalizado el acceso a esta información, rompiendo las fronteras entre el mundo físico y el virtual.

Dada esta revolución, el objetivo del proyecto EV es la definición de la arquitectura, protocolos y estándares de la Internet Geográfica, con un foco especial en la visualización 3D, mundos virtuales e interacción entre usuarios. España Virtual incluye aspectos semánticos y tecnologías para el procesamiento masivo y almacenamiento de datos geográficos.

El proyecto EV se estructuró en nueve paquetes de trabajo que se agrupan en dos áreas: Datos e Infraestructura Geográfica y Arquitectura y Tecnologías de Internet 3D. La primera de las áreas trata todos los temas relativos a la obtención y el tratamiento de los datos geográficos, incluyendo investigaciones para mejorar el procesamiento y el almacenamiento de los mismos. La segunda área trata de hacer estos datos accesibles a los usuarios finales a través de productos de visualización y servicios en Internet.

EV ha representado en gran parte la contribución de nuestro país a dicha revolución, aún en curso, creando la tecnología para unir dos mundos: el geográfico y el de servicios de Internet, desarrollando los máximos avances de la mano de los principales actores nacionales en cada campo.

[16] http://www.españavirtual.org
[17] http://www.cdti.es

El trabajo desarrollado en esta tesis encaja claramente en los objetivos del proyecto. De esta forma, el trabajo queda enmarcado en los paquetes de trabajo relacionados con metadatos, incorporación de nuevos tipos de datos, IDEs e indexación y búsqueda de recursos.

1.2.3 ENVIronmental Services Infrastructure with ONtologies

El proyecto *ENVIronmental Services Infrastructure with ONtologies*[18] (ENVISION) es un proyecto FP7 financiado por la Comisión Europea dentro del área de tecnologías de la información y la comunicación para servicios medioambientales y adaptación al cambio climático. El objetivo del proyecto es proporcionar una infraestructura de servicios ambientales con ontologías que pretende apoyar a los usuarios sin conocimientos TIC en el proceso de descubrimiento semántico y el encadenamiento y composición de servicios ambientales [150].

Los modelos ambientales son tradicionalmente aplicaciones independientes. Para permitir el acceso a estos modelos, el objetivo de iniciativas como GEOSS o GMES es proporcionar infraestructuras que suministren observaciones de la Tierra procedentes de sensores físicos o modelos ambientales para una amplia gama de comunidades de información. ENVISION sigue la idea de ofrecer los modelos como servicios, cuya realización es una de las tareas identificadas en el plan de trabajo GEOSS [87]. Se supone que permitirá a científicos y ciudadanos acceder, combinar y comparar modelos ambientales, con el objetivo final de aumentar la confianza en los resultados del modelo y apoyar el desarrollo colaborativo de modelos que representen mejor el medio ambiente.

Dentro del proyecto ENVISION, se definen tres demostradores, un piloto sobre deslizamientos de tierra basado en modelos de riesgo de deslizamientos de tierra, un piloto sobre derrames de petróleo basado en modelos de predicción de la deriva del petróleo, y un piloto sobre inundaciones basado en la monitorización de las inundaciones en tiempo real y en las previsiones del nivel de agua. Flujos de trabajo complejos, implementados en forma de composiciones de servicios, modelan los pilotos y hacen que los resultados estén disponibles en la

[18] http://www.envision-project.eu

Web. A través de una gestión integrada de los recursos, los diseñadores de los flujos de trabajo tienen la oportunidad de añadir, eliminar y anotar semánticamente dichos servicios (tanto servicios de datos como de procesamiento). Sin embargo, el diseñador tiene que ser consciente de que los servicios cumplan los requisitos del flujo de trabajo que está siendo modelado. En este sentido, las descripciones (metadatos) ayudan a descubrir estos servicios o identificar el contenido de los mismos.

En el contexto de este proyecto se ha trabajado en servicios de descripción de recursos, anotación semántica y publicación de los recursos junto con sus anotaciones en servidores de catálogo.

1.3 Metodología de la Investigación

La metodología adoptada para alcanzar los objetivos propuestos se puede resumir en los siguientes puntos:

- Estudio del estado del arte de la tecnología y arquitectura de sistemas distribuidos relativa a los datos cartográficos y datos multimedia en general.
- Estudio y análisis de los actuales flujos de trabajo (*workflows*[19]) completos para el descubrimiento de recursos, lo que proporcionará una visión global del proceso.
- Análisis y comparación de los beneficios y limitaciones que ofrecen las especificaciones, estándares y metodologías actuales de descripción, publicación y descubrimiento de recursos georreferenciados, particularmente en los relativos a los mundos cartográficos y de la Observación de la Tierra.
- Proponer un escenario de trabajo en el ámbito de la investigación (según necesidades reales del proyecto CENIT "España Virtual").
- Diseñar, proponer y/o extender mecanismos, *framework*[20] y/o arquitecturas que permitan la descripción, publicación y

[19] Un *workflow* incluye todos los aspectos operacionales de una actividad de trabajo: cómo se estructuran las tareas, cómo se realizan, cuál es su orden correlativo, cómo se sincronizan o cómo fluye la información que soporta las tareas.

[20] En el contexto de la informática, un *framework* es una estructura conceptual y tecnológica, en base a la cual otro proyecto de software puede ser más fácilmente organizado y desarrollado.

descubrimiento de recursos multimedia del mundo cartográfico de forma semiautomática o automática en un entorno abierto como la Web.

- Implementación de un prototipo del sistema y evaluación de la validez de nuestra aproximación mediante la aplicación de un escenario de uso.
- Participación y presentación de las nuevas ideas en diversos foros, como congresos nacionales e internacionales. Lo que ha permitido la depuración y posterior mejora de las ideas de cara a su exposición final en esta tesis doctoral.

1.4 Contribuciones

Las principales contribuciones de este trabajo son la investigación y propuesta de un *workflow* y una solución técnica para ofrecer a los usuarios un sistema integrado que le permita recolectar, describir, publicar y buscar recursos georreferenciados de una forma homogénea.

Para ello, a nivel general, se propone **un *workflow*** basado en la evolución de la WWW y adaptado a las necesidades de la IG cubriendo todo su ciclo de vida (ver Capítulo 5). En base a este *workflow*, se ha diseñado **una arquitectura** genérica para un sistema de indexación y búsqueda de recursos georreferenciados (ver Sección 5.1). Implementando esta arquitectura se ha desarrollado ***GeoCrawler*** como prototipo basado en este *workflow* y que integra las diferentes soluciones parciales que se presentan en esta tesis (ver Sección 5.2 y Sección 5.3).

El título de esta tesis está compuesto por las tres etapas principales del *workflow* propuesto, a continuación se detallan las contribuciones relativas a cada una de estas etapas. Para mayor claridad, estas contribuciones han sido resumidas de forma gráfica en la Figura 1, que también muestra las relaciones existentes entre ellas.

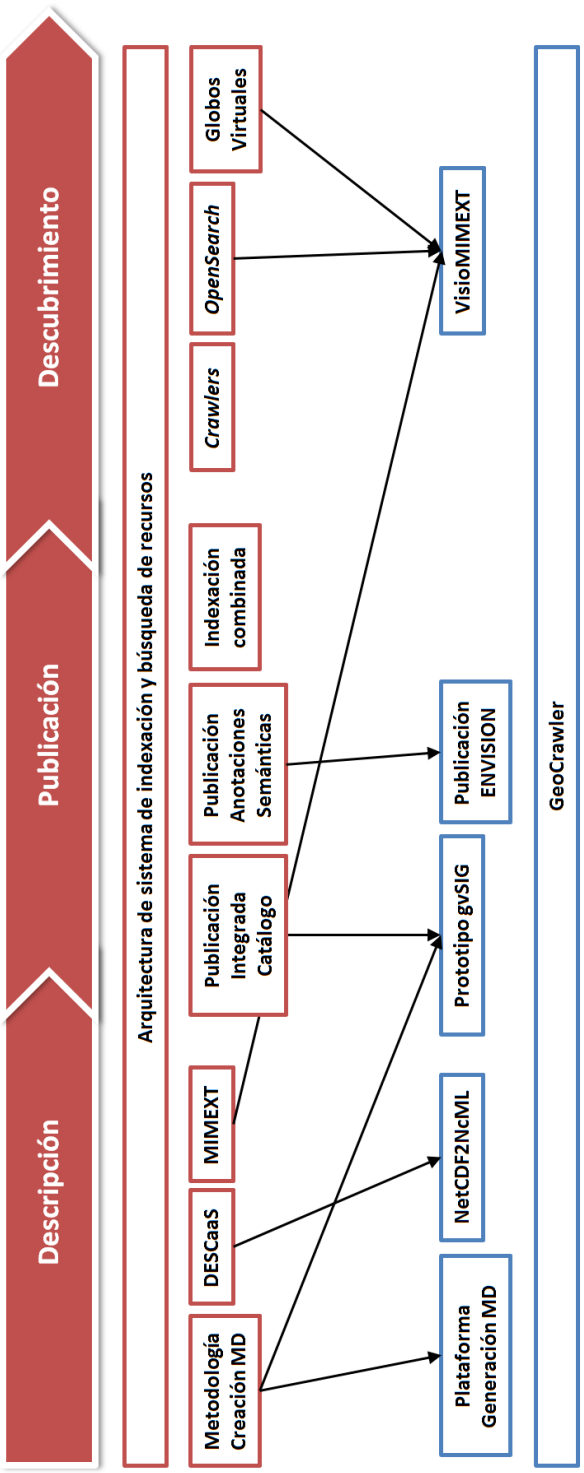

Figura 1: Contribuciones

Las principales contribuciones que se aportan para la descripción de recursos georreferenciados son:

- Una metodología para la generación de metadatos que pretende automatizar su producción para evitar el tedioso trabajo manual más dado a errores y facilitar la descripción de los recursos (ver Sección 2.2.1). Esta metodología se comprueba y se demuestra mediante:
  - La implementación de un prototipo de gestor de metadatos sobre la aplicación gvSIG (ver Sección 2.2.2).
  - El desarrollo de una plataforma común de generación de metadatos que permite obtener información y acceder de forma homogénea a recursos heterogéneos (ver Sección 2.2.3).
- Un mecanismo común para anotar y georreferenciar recursos independientemente de su naturaleza llamado MIMEXT (ver Sección 2.3).
- Un nuevo paradigma llamado *Description as a Service* (DESCaaS) cuyo objetivo es proporcionar descripciones de recursos homogéneas *on-line* (ver Sección 2.4).
  - Este paradigma ha sido implementado para un caso de uso real en el contexto del proyecto ENVISION, para proporcionar descripciones de recursos NetCDF (NetCDF2NcML).

Las principales contribuciones que se aportan para la publicación de recursos georreferenciados son:

- Una metodología para publicar los recursos generados en servicios de catálogo de forma integrada en el flujo de trabajo (ver Sección 3.2.1). Esta metodología se comprueba y se demuestra mediante:
  - La implementación de un prototipo de extensión de metadatos sobre gvSIG (ver Sección 3.2.2).
- Un mecanismo para publicar recursos anotados semánticamente en servicios de catálogo (ver Sección 3.2.3).
  - Este mecanismo ha sido implementado en el contexto del proyecto ENVISION.
- Un método para publicar los recursos mediante la indexación de sus descripciones combinando índices textuales y espaciales, de forma que los índices espaciales se integran sobre los índices textuales (ver Sección 3.3).
  - Este método ha sido comprobado gracias a su implementación en *GeoCrawler*.

Las principales contribuciones que se aportan para el descubrimiento de recursos georreferenciados son:

- Automatizar la recopilación de los recursos georreferenciados mediante aplicaciones de tipo *crawler* (ver Sección 4.2).
  - Esta contribución ha sido puesta en práctica en la implementación de *GeoCrawler*.
- Una interfaz de consulta común y homogénea basada en *OpenSearch*, aplicable a un amplio espectro de servicios de recursos georreferenciados en la red (ver Sección 4.3).
- Visualización e integración de datos heterogéneos para cualquier tipo de usuario gracias a herramientas basadas en globos virtuales (ver Sección 4.4). Tanto esta contribución como la anterior, junto con el mecanismo para la anotación de recursos MIMEXT, han sido comprobadas y demostradas mediante:
  - El desarrollo de la aplicación VisioMIMEXT basada en un globo virtual, OpenSearch y MIMEXT. Integrando mecanismos de búsqueda y visualización de forma natural sobre el globo virtual (ver Sección 4.5).

1.5 Organización de la Tesis

En esta tesis se propone un *workflow* (ver Figura 2) que, tras recopilar los recursos georreferenciados disponibles, pretende averiguar qué son y qué contienen estos recursos, por lo que intenta describirlos en función de sus propiedades. El siguiente paso es la publicación de los recursos utilizando la información obtenida con el fin de que puedan ser descubiertos de forma fácil y eficiente. A partir de este punto, los usuarios podrán descubrir los recursos y, finalmente, visualizarlos y explotarlos correctamente.

Figura 2: *Workflow* general (simplificado)

Este *workflow* (ver Figura 2) servirá como hilo conductor para estructurar esta tesis. Por lo tanto, los restantes capítulos de esta tesis se organizan de la siguiente manera:

El Capítulo 2 introduce el concepto de recurso georreferenciado y conceptos teóricos y técnicos relevantes para la *Descripción* de estos a través de los metadatos, pieza básica de nuestro sistema de indexación y búsqueda de recursos georreferenciados. En este capítulo también se exploran diferentes métodos que permiten crear metadatos, se estudia cómo anotar los recursos con nueva información y se investiga cómo facilitar la creación y distribución de metadatos.

El Capítulo 3 presenta conceptos teóricos y técnicos relevantes sobre la *Publicación* de recursos de forma interoperable. En este sentido, el capítulo analiza diferentes estrategias de catalogación e indexación de los recursos en base a sus descripciones, permitiendo y facilitando su posterior descubrimiento.

Una vez entendido cómo describir recursos georreferenciados y cómo ponerlos a disposición de los usuarios, el Capítulo 4 se centra en su *Descubrimiento*, nuestro objetivo final. Por ello, se revisan diferentes métodos actuales que permiten buscar y recuperar recursos georreferenciados. Además, cualquier SI necesitará previamente descubrir los recursos que incorporará, por ello el capítulo también estudia diferentes opciones para la recopilación de los recursos. Por otra parte, entendemos que la visualización de los recursos supone un factor fundamental para su descubrimiento por lo que también se considera en este capítulo.

En estos tres capítulos, para cada uno de los problemas que se aborda, se presenta la parte correspondiente del estudio del estado del arte y los conceptos relacionados, seguidamente, se presentan las contribuciones conceptuales e implementaciones realizadas para dicho problema. De esta forma se pretende dar una visión más completa e integrada de cada uno de los problemas concretos y es posible justificar y comprender mejor la solución propuesta a través de las diferentes opciones y de la evolución que se detalla en el estado del arte. Además, las contribuciones conceptuales quedan validadas por sus respectivas pruebas de concepto mediante implementaciones en casos de uso reales.

El Capítulo 5 detalla el *workflow* propuesto y explica cómo encajan todas las piezas que se han presentado en esta tesis. Este capítulo también presenta una primera aproximación para desarrollar un sistema de indexación y búsqueda de recursos georreferenciados, describiendo su arquitectura, explicando los detalles de implementación más relevantes y analizando el resultado obtenido.

Finalmente, el Capítulo 6 resume las principales aportaciones de esta tesis, presenta las conclusiones alcanzadas y se discuten las limitaciones y futuras direcciones de investigación que permitirán continuar con la investigación iniciada en esta tesis.

Adicionalmente, en el Anexo A se presentan las publicaciones más representativas derivadas del trabajo de investigación realizado en esta tesis, las cuales han sido un referente externo de evaluación.

2. Descripción

Combinando las definiciones que nos ofrecen la Wikipedia[21] y la Real Academia Española[22] (RAE), podemos afirmar que una descripción es la explicación, de forma detallada y ordenada, de cómo es cierta persona, lugar o cosa, a través de la definición de sus varias partes, características y circunstancias.

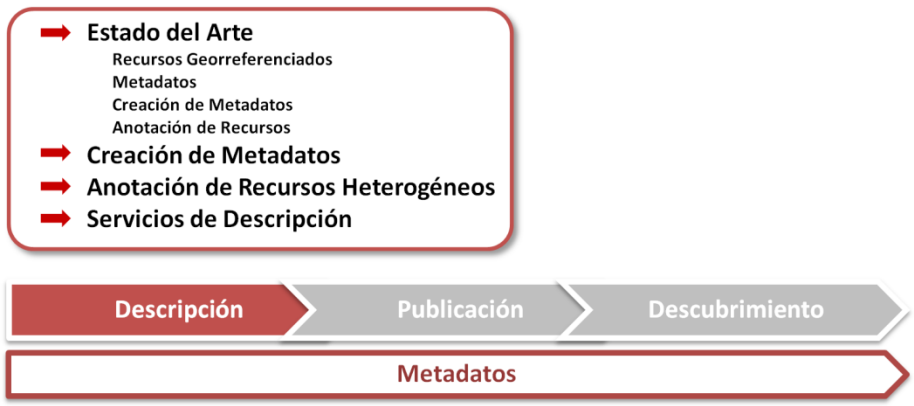

Figura 3: Visión general del Capítulo 2

El concepto de descripción es general, amplio y aplicable a cualquier cosa. Este capítulo presenta conceptos teóricos y técnicos relevantes para conseguir descripciones de recursos. Con el fin de acotar nuestro alcance, nos centraremos en la descripción de recursos

[21] http://es.wikipedia.org/wiki/Descripción
[22] http://lema.rae.es/drae/?val=descripcion

georreferenciados. La Figura 3 representa la estructura del capítulo y su posición dentro del *workflow* general.

En primer lugar, en este capítulo, se pretende responder a la pregunta: *¿Qué vamos a describir?* Por lo que definirá qué entendemos nosotros por recursos georreferenciados.

En segundo lugar, se pretende responder a la pregunta: *¿Cómo lo vamos a describir?* Los metadatos permiten describir recursos de forma estructurada y, en base a estas descripciones, es posible organizarlos, publicitarlos y facilitar el acceso a ellos. Los metadatos son conjuntos estructurados de datos que describen otros datos y cuyo propósito es mejorar el conocimiento sobre los recursos descritos. De esta forma, las descripciones de los recursos resultan ser la pieza clave de cualquier sistema de información [169] [62]. Por ello, en este capítulo se presenta el concepto de metadatos y se analizan los diferentes aspectos respecto a ellos, como sus objetivos, su funcionalidad o su relevancia en el contexto de la información geoespacial.

El hecho de describir los recursos mediante metadatos implica una nueva pregunta: *¿Cómo creamos los metadatos?* Los metadatos deben ser creados para responder a las necesidades actuales, especialmente el descubrimiento de los recursos. Muchos autores de renombre en este campo han identificado diferentes formas de crear metadatos: editándolos a mano, por extracción, por cálculo o por inferencia. Pero la mayoría de ellos están a favor de automatizar la producción de metadatos para evitar errores y para evitar a los creadores este arduo y monótono trabajo que suele derivar en una baja calidad y/o falta de metadatos. Por ello, se exploran los diferentes métodos y herramientas para la generación de metadatos. Posteriormente, en la Sección 2.2, como contribución se propone una metodología para la creación de metadatos y, en base a ella, se exploran dos casos de estudio reales que implementan parte de dicha metodología.

Además de crear los metadatos deberemos explorar *¿cómo representar y adjuntar los metadatos a los recursos?* En este capítulo también se estudian y analizan los diferentes formatos existentes para la anotación de recursos y su codificación. Hay que tener en cuenta que, dada la creciente demanda por georreferenciar cualquier tipo de

recurso, es necesario un mecanismo genérico que nos permita anotarlos a todos. Con este objetivo, en la Sección 2.3, la contribución que se presenta es MIMEXT, una solución que permite georreferenciar recursos heterogéneos sin depender de su tipo de datos o formato.

Finalmente, como respuesta a la pregunta *¿cómo facilitar la creación y distribución de metadatos?* en la Sección 2.4 se presenta un nuevo paradigma cuya funcionalidad es proporcionar descripciones homogéneas de recursos a través de la red, con el fin de facilitar su publicación y explotación y mejorar su accesibilidad y descubrimiento. La validez y beneficios de esta contribución son comprobados a través de un caso de uso específico dentro de un proyecto real.

2.1 Estado del Arte y Conceptos Previos

En esta sección se introducen conceptos previos importantes para el desarrollo de la presente tesis, como son los recursos georreferenciados y los metadatos. Además se realiza un amplio estudio del estado del arte tanto para la creación de los metadatos como para la anotación de los recursos.

2.1.1 Recursos Georreferenciados

Cuando hablamos de información geoespacial estamos hablando de datos intrínsecamente relacionados con una posición geográfica.

Algunos estudios muestran que la mayoría de los recursos o conjuntos de datos (más del 80%) existentes son susceptibles de estar relacionados con una posición geográfica [79] [136]. Por lo tanto, existen multitud de recursos de naturaleza espacial y formatos de datos diseñados especialmente para los datos geoespaciales que contienen. Sin embargo cualquier otro dato o recurso, aunque en un principio no sea considerado de naturaleza espacial, puede ser georreferenciado y puede considerarse como tal.

Georreferenciación[23] es un neologismo que se refiere al establecimiento de relaciones entre un objeto y su existencia en un espacio físico, por las que se definen el posicionamiento y la localización del objeto (representado mediante punto, vector, área, volumen) en un sistema de coordenadas y *datum* determinado [104].

En consecuencia, el concepto de recurso georreferenciado es un concepto más amplio que engloba a todos aquellos recursos que contienen IG y cualquier información que haya sido georreferenciada. Podemos definir los recursos georreferenciados como aquellos recursos de cualquier naturaleza que han definido su existencia en un espacio físico. Es decir, aquellos que han establecido su localización en términos de proyecciones geográficas o sistemas de coordenadas.

Actualmente, la georreferenciación ha ido más allá de las especialidades de geociencias y de los Sistemas de Información Geográfica (SIG), debido a la aparición en los últimos años de nuevas herramientas cuya facilidad de uso ha extendido y democratizado esta tarea fuera del ámbito técnico existente hasta ahora.

La masificación y evolución constante de la georreferenciación se ha visto impulsada por el uso *mashups*[24] en sitios Web 2.0, permitiendo la localización de contenidos digitales (vídeo, noticias, modelados 3D, etc.) en cartografía digital, dentro de lo que se conoce actualmente como neogeografía [211] [91] [90].

El uso de herramientas como *Google Earth*[25], *Flickr*[26], etc. ha implicado un salto cualitativo y sobretodo cuantitativo en cuanto a georreferenciación. Se ha extendiendo el uso de datos georreferenciados, tradicionalmente limitados a datos geográficos por parte de especialistas de las geociencias y SIG. Ahora la georreferenciación tiene un impacto sociológico puesto que se realiza

---

[23] Es muy común encontrar la palabra escrita como *georeferenciación*, con un única r, pero es un error. En español las palabras compuestas cuyo segundo formante comienza por r, de manera que el sonido vibrante múltiple quede en posición intervocálica, se escriben con doble erre, por lo tanto la forma correcta es georreferenciación.

[24] Un *mashup* es una aplicación web hibrida que usa y combina datos, presentaciones y funcionalidad procedentes de una o más fuentes para crear nuevos servicios.

[25] http://www.google.com/earth

[26] http://www.flickr.com/map

sobre todos los contenidos sociales presentes en el mundo. Esto está acelerando la aparición de una web geosemántica [39].

Todos estos recursos georreferenciados, como la IG, pueden ser descritos mediante el uso de metadatos e integrados en las IDEs. De hecho la propia georreferenciación de un recurso implica añadirle los metadatos necesarios para establecer su localización.

2.1.2 Metadatos

Concepto

Como reflejó M.A. Manso en su tesis doctoral [143], el concepto de metadatos no es nada nuevo, la definición más común del término *metadatos* es *"datos sobre los datos"*. Si buscamos los orígenes de la palabra *metadatos*, en primer lugar, encontraremos sus raíces en la palabra griega *"μετα"* (meta) cuyo significado es *"después de, más allá de"*. En segundo lugar *"datos"* resulta ser el plural de la palabra Latina *"datum"* cuyo significado es *"lo que se da"* o *"antecedente necesario para llegar al conocimiento exacto de algo o para deducir las consecuencias legítimas de un hecho"* según la RAE[27]. Literalmente su significado sería *"sobre datos"*, es decir, son datos que describen otros datos.

Definición

A pesar de sus raíces antiguas la palabra metadatos no apareció impresa hasta 1973 [108], aunque fue acuñada por Lack Myers en los 60 para describir conjuntos de datos y productos en el contexto de la gestión bibliográfica [143]. El término metadatos no tiene una definición única, desde entonces, en la literatura relacionada, han aparecido multitud de autores que proporcionan una interpretación el alcance del significado teórico y práctico del término. Además de la definición más difundida *"datos sobre datos"*, existen otras definiciones populares como *"información sobre datos"* [197], *"datos sobre información"* [200], *"información sobre información"* [228] u otras más complejas como *"descripciones estructuradas y opcionales que están disponibles de forma pública para ayudar a localizar objetos"* [34], "una

[27] http://www.rae.es

abstracción jerárquica y descriptiva sobre los datos que se describen" [233] o *"datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas"* [66]. Estas últimas pretenden evitar los problemas de las definiciones simples que resultan demasiado difusas y generales dificultando la tarea de acordar estándares.

En el contexto de los sistemas informáticos, el término metadatos se empezó a utilizar para evitar algunos prejuicios de nomenclatura por parte de los profesionales en el campo de la información, tras un *workshop* en 1995 [37]. Las primeras referencias sobre este término en el contexto de la IG datan de 1996 [8] [124]. A partir de entonces, a pesar de las diferentes definiciones que aparecen en la literatura [153] [71] [68] [226], los metadatos son usados para describir el contexto, la calidad, la condición o las características de los datos de forma que los usuario pueden descubrir y entender los recursos.

Resumiendo las aportaciones de todos estos autores, en [143] M.A Manso define el término metadatos como *"un conjunto estructurado de datos que describen otros datos, y cuya finalidad es mejorar el conocimiento de la información descrita ayudando a responder preguntas como 'qué', 'quién', 'dónde', 'cuándo', 'cuánto' y 'cómo' "*. En el contexto de la IG, también pueden ser descritos como *"esos productos autónomos que, enlazados a los datos, permiten mantener un inventario de éstos, su publicación y posterior referencia en los catálogos presentes en las IDEs y, finalmente, la reutilización de los datos"*.

Objetivos

La creación de metadatos geográficos persigue tres objetivos (y a su vez beneficios) principales [75]:

- Organizar y mantener la inversión en datos hecha por una organización: los metadatos buscan fomentar la reusabilidad de datos sin tener que recurrir al equipo humano que se encargó de su creación inicial.
- Publicitar la existencia de información geográfica a través de sistemas de catálogo: Los registros de metadatos se suelen publicar a través de sistemas de catálogos, en ocasiones también denominados como directorios o registros. Los catálogos

electrónicos no difieren demasiado de los catálogos tradicionales de una biblioteca excepto por el hecho de que ofrecen una interfaz estandarizada de servicios de búsqueda. Así pues, estos catálogos son la herramienta que pone en contacto a los consumidores con los productores de información. Mediante la publicación de recursos de información geográfica a través de un catálogo, las organizaciones pueden encontrar datos a usar, otras organizaciones con las que compartir datos y esfuerzos de mantenimiento, así como clientes para esos datos.

- Facilitar el acceso a los datos, su adquisición y una mejor utilización de los mismos logrando una interoperabilidad de la información cuando esta procede de fuentes diversas. Los metadatos ayudan al usuario u organización que los recibe en el procesamiento, interpretación y almacenamiento de los datos en repositorios internos.

Funcionalidad

Tras caracterizar conceptualmente los metadatos y repasar sus objetivos, vamos a explorar su propósito. En [143] M.A. Manso realizó una clasificación de los metadatos en base a su funcionalidad según multitud de autores [16] [55] [121] [69] [115] [165] [176] [119] [180] [155] [86] [60]. En base a las coincidencias encontradas en la literatura y en base a dicha clasificación podemos extraer que las principales funciones de los metadatos son: el descubrimiento, la evaluación, el acceso y el uso de los recursos. Además de permitir el desarrollo de sistemas interoperables a diferentes niveles.

El descubrimiento de recursos implica su búsqueda y su localización. En este sentido, los metadatos deben proporcionar suficiente información para responder a la pregunta: *¿Qué recursos contienen datos como los que estoy interesado?* Para ello, los metadatos deben responder preguntas como ‘qué’, ‘quién’, ‘dónde’, ‘cuándo’, ‘cuánto’ y ‘cómo’ para discernir el contenido, formato y alcance de un recurso. Estos metadatos permitirán a los usuarios descubrir los recursos y favorecen la gestión, almacenaje y reutilización de los datos.

La evaluación de los recursos implica su exploración y valoración. En este sentido, los metadatos deben proporcionar suficiente información para responder a la pregunta: *¿Contiene el recurso datos*

suficientes y válidos para la aplicación que requiero? Es decir, los metadatos deben proporcionar información que ayude a los usuarios a determinar si los datos van a ser útiles para cierta aplicación. Para ello, resulta necesario describir detalladamente la calidad de los datos. En el contexto de la IG podemos destacar cinco aspectos de calidad: completitud, exactitud temática, exactitud temporal, exactitud posicional y consistencia lógica [181]. Estos metadatos permitirán a los usuarios determinar si los datos se ajustan al uso previsto.

Tras haber localizado los recursos y evaluado si son adecuados, en algunos casos los usuarios necesitan acceder a ellos. El acceso a los recursos implica su recuperación y transferencia. En este sentido, los metadatos deben proporcionar suficiente información para responder a la pregunta: *¿Cuál es el proceso para obtener el recurso?* Es decir, los metadatos deben proporcionar al usuario suficiente información para recuperar los datos, esto puede ser tan simple como una proporcionar una URL que identifique la localización de un recurso digital o ser más complejo incluyendo información sobre asuntos de seguridad, persona de contacto, formato de distribución de los datos, restricciones de acceso o información sobre costes.

Una vez que disponemos de los datos, los usuarios necesitan saber cómo manejarlos y manipularlos. El uso de los recursos implica su explotación. En este sentido, los metadatos deben proporcionar suficiente información para responder a la pregunta: *¿Cómo usar el recurso?* Es decir, los metadatos deben proporcionar la información necesaria para utilizar o explotar los recursos recuperados, como su tamaño o la estructura lógica y física (formato) tanto de los datos como de los propios metadatos. Estos metadatos permitirán a los usuarios conocer cómo fusionar y combinar estos datos con los suyos propios, cómo aplicarlos correctamente y entender completamente sus propiedades y limitaciones.

Interoperabilidad

El término interoperabilidad tiene muchas connotaciones, incluyendo los objetivos de comunicación, intercambio de información, cooperación y la compartición de recursos entre los diferentes tipos de sistemas [143]. De hecho, la esencia de la interoperabilidad es asegurar las relaciones entre sistemas, siendo cada relación una forma de compartir, comunicar, intercambiar y cooperar [38]. Nuestra

atención se centra en todos los aspectos de las IDEs que deben facilitar la localización, evaluación, acceso y uso de IG de forma transparente al usuario, ya sean agentes humanos o aplicaciones informáticas [221] [88] [89].

Los metadatos proporcionan a los sistemas la capacidad de interactuar a diferentes niveles por lo que resultan ser la pieza clave para lograr la interoperabilidad. La extensa literatura sobre el tema muestra numerosos y diferentes niveles de interoperabilidad. Todos estos niveles han sido analizados y discutidos en el contexto de la IG y las IDEs, en [145] estos niveles se resumen en: técnico, sintáctico, semántico, pragmático, dinámico, conceptual y organizativo.

La interoperabilidad técnica permite la interconexión de los sistemas a través de protocolos de comunicación comunes permitiendo el intercambio de información en su nivel más básico (juego de caracteres, codificación, protocolos, tipos de servicios y formatos, sus versiones, etc.). Interoperabilidad sintáctica permite el intercambio de información entre los sistemas mediante el uso de un formato de datos, estructura, lógica, etc. común (estándares y especificaciones). La interoperabilidad semántica permite el intercambio de información utilizando un vocabulario compartido que evita imprecisiones y errores al interpretar el significado de los términos. Interoperabilidad pragmática permite que los sistemas interconectados se conozcan, siendo capaces de explotar sus interfaces, invocando métodos o procedimientos y manejando los datos necesarios para el intercambio con otros sistemas (*APIs*[28]). La interoperabilidad dinámica permite a los sistemas supervisar otros sistemas y responder a cambios detectados en la transferencia de información o retardo de tiempo, tomando ventaja de los mismos. La interoperabilidad conceptual permite conocer y reproducir las funciones de un sistema en base a su documentación. Por último, la interoperabilidad organizativa permite conocer los objetivos de negocio, modelos de procesos, regulaciones y políticas de acceso y uso de datos y servicios.

[28] Del inglés *Application Programming Interface* (API) es el conjunto de funciones y procedimientos que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción.

Cada elemento de metadatos promueve, en diferentes grados, el papel que estos juegan y facilita la interoperabilidad entre sistemas en un entorno distribuido como las IDEs.

Clasificación

En [143], de acuerdo con la literatura relacionada, los metadatos son categorizados en base a su relación con los datos (implícito, explicito) [157] [222] [120] [12] [61], su comportamiento temporal (estático, dinámico) [120], su propósito (estructural, de control, descriptivo, administrativo) [26] [120] [166], su naturaleza (objetivo, subjetivo) [68] o si puede derivar de otro (calculado, inferido, contextual) [16] [91]. Aunque podrían clasificarse en base a muchos criterios, entre ellos, su funcionalidad o el papel que juegan en la interoperabilidad.

Sin embargo, la forma más usual de clasificar los elementos de metadatos es respecto al rol que desempeñan dentro del paradigma *"descubrimiento, evaluación y acceso"* establecido en [165]:

- Los metadatos de descubrimiento son aquellos elementos que permiten describir mínimamente la naturaleza y contenido de un recurso. Estos elementos suelen responder a las preguntas *"Qué, Por qué, Cuándo, Quién, Dónde y Cómo"*. Los elementos típicos en esta categoría serían el título, la descripción del conjunto de datos o su extensión geográfica.
- Los metadatos de exploración proporcionan la información que permiten verificar que los datos se ajustan al propósito deseado, permiten evaluar sus propiedades, o permiten contactar con la entidad que facilitará más información.
- Los metadatos de explotación incluyen aquellas descripciones necesarias para acceder, transferir, cargar, interpretar y utilizar los datos en la aplicación final que los explote.

Por otra parte, también es posible clasificar los metadatos desde el punto de vista de la creación de metadatos [141]. En particular, se consideran metadatos *implícitos* aquellos fuertemente relacionados con los datos y su uso; metadatos *explícitos* aquellos relacionados con el tipo de datos y su almacenamiento; metadatos *calculados* aquellos que pueden ser obtenidos a partir de otros en base a algún tipo de cálculo o tratamiento; metadatos *inferidos* aquellos que pueden ser deducidos a partir de otros en base a reglas lógicas; y metadatos

contextuales aquellos que pueden ser obtenidos o impuestos por el contexto en el que datos y metadatos son creados y publicados.

Relevancia

Como hemos visto los metadatos juegan un papel fundamental en cualquier sistema de información dada su funcionalidad. En el contexto de la IG, las IDEs no resultan una excepción. Por ello, importancia de los metadatos en este contexto ha sido reconocida por las principales entidades a todos los niveles (nacional, europeo e internacional) promoviendo el desarrollo de una IDE global.

Europa, consciente del creciente valor de la información geográfica y de su falta de homogeneidad en los 27 países miembros, ha impulsado iniciativas como INSPIRE[29] (bajo la Directiva 2007/2/EC del Parlamento Europeo y del Consejo del 14 de marzo de 2007) con el fin de hacer que la información geográfica esté armonizada, sea de alta calidad y se encuentre fácilmente disponible para su uso a nivel local, regional, nacional o internacional. El mundo geográfico ha apoyado con fuerza la creación de IDEs con la vocación de compartir de forma estándar e interoperable información geográfica de interés general como servicio público. Esta directiva está haciendo que las IDEs regionales o nacionales no solo sean deseables, sino que sean legalmente requeridas.

El objetivo de la iniciativa INSPIRE consiste en la creación de una infraestructura de información geoespacial europea que proporcionará a usuarios finales (ya sea administración, empresas y organizaciones tanto a nivel europeo, nacional o local, así como a ciudadanos individuales) servicios integrados alimentados por la información geoespacial disponible. De este modo, los usuarios finales pueden acceder a los datos mediante los catálogos de servicios definidos en la infraestructura INSPIRE, la cual proporciona un acceso transparente a un amplio rango de fuentes de información distintas, tanto a nivel global como local, para cualquier tipo de usuario final.

La Figura 4 representa los principales componentes que forman parte de una IDE genérica. Se pueden observar las cuatro piezas principales: aplicaciones de usuario, servicios de geoprocesamiento, catálogos de metadatos y los repositorios de IG.

[29] http://inspire.jrc.ec.europa.eu

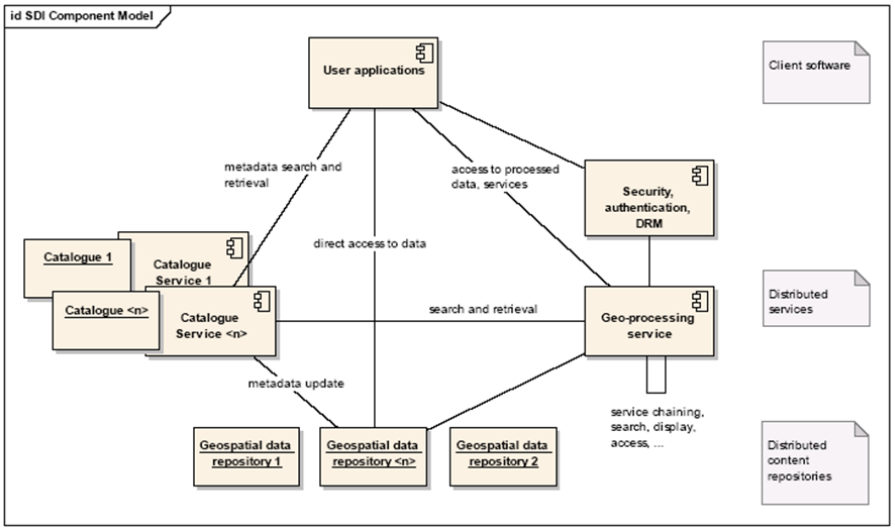

Figura 4: Arquitectura general de las IDEs

Analizando el contexto de las IDEs, podemos observar que los metadatos suponen una pieza clave, un elemento fundamental en base al cual funcionan el resto de componentes. Por ello, directivas como INSPIRE establecen la creación y mantenimiento de metadatos y servicios de descubrimiento relacionados [50] que, a menudo, son los primeros elementos de valor añadido visibles en una IDE.

A nivel Americano encontramos el *Federal Geographic Data Committee*[30] (FGDC) que promociona de forma coordinada la creación, uso, compartición y diseminación de IG en el país, a través del desarrollo de su IDE nacional: *National Spatial Data Infrastructure*[31] (NSDI). En la cual los metadatos también tienen un papel fundamental [74].

A nivel internacional la *Global Spatial Data Infrastructure Association*[32] (GSDI) promueve la cooperación y la colaboración internacional para apoyar el desarrollo de IDEs a nivel local, nacional e internacional que permitan a las naciones abordar temas sociales, económicos y ambientales de primordial importancia. De nuevo GSDI destaca el importante papel que desempeñan los metadatos [89].

[30] http://www.fgdc.gov
[31] http://www.fgdc.gov/nsdi/nsdi.html
[32] http://www.gsdi.org

Finalmente, a nivel Español el Consejo Superior Geográfico[33] (CSG), órgano dependiente del Ministerio de Fomento[34] es el encargado de promover y planificar el desarrollo de la IDE de España[35]. Que, al ser de un país de la Unión Europea, debe seguir las directrices marcadas por INSPIRE. Y donde los metadatos también juegan un papel destacado como muestra su portal dedicado a los metadatos[36].

Consideraciones

La mayoría de las veces no es posible diferenciar entre datos y metadatos. Por ejemplo, un poema es un grupo de datos, pero también puede ser un grupo de metadatos si está adjuntado a una canción que lo usa como letra.

Muchas veces, los datos son tanto *datos* como *metadatos*. Por ejemplo, el título de un documento es parte del texto como a la vez es un dato referente al texto, por lo tanto es tanto dato como metadato.

Debido a que los metadatos son datos en sí mismos, es posible crear metadatos sobre metadatos. Aunque, a primera vista, parece absurdo, los metadatos sobre metadatos pueden ser muy útiles. Por ejemplo, al fusionar dos recursos distintos y sus respectivos metadatos puede ser muy importante deducir cuál es el origen de cada grupo de metadatos, registrando ello en metadatos sobre los metadatos.

2.1.3 Creación de Metadatos

Los metadatos son normalmente creados por los proveedores de los datos, generados de forma manual y almacenados (separados del recurso) en catálogos, para más tarde ser encontrados con fines informativos. Sin embargo, problemas prácticos con su creación y mantenimiento están limitando su efectividad para tareas como el descubrimiento o la evaluación de los recursos que describen.

---

[33] http://www.idee.es/web/guest/consejo-superior-geografico
[34] http://www.fomento.gob.es
[35] http://www.idee.es
[36] http://metadatos.ign.es

Varios expertos están a favor de asignar la tarea de crear los metadatos a los propietarios de los recursos [141], dado que piensan que los dueños son los mejor capacitados para proporcionar información sobre sus datos [95] [126] [16]. En la práctica la creación de metadatos ha ocupado un papel secundario en las organizaciones, siendo creados habiendo transcurrido tiempo tras la producción de los recursos. Por ello, algunas organizaciones han considerado la creación de metadatos como un coste añadido [161]. Hecho que ha sido desmentido en otros estudios que aseguran que pese a que inicialmente la creación de metadatos pueda parecer costosa, a largo plazo las ventajas superan los inconvenientes, dado que el coste inicial de documentar los datos es muy inferior al potencial coste de la generación de datos duplicados y redundantes [40].

Por otra parte, algunos autores destacan como principales problemas de la creación de metadatos la complejidad de las reglas y estándares existentes en el contexto de la IG [23], y la baja automatización y sincronización entre la creación de los datos y los metadatos [34] [147]. Estos problemas, entre otros, ha motivado a la comunidad científica a revisar los existentes procedimientos y tareas para la creación de metadatos tanto de forma manual, automatizada o mixta [35] [53]. El principal reto es proponer nuevos procedimientos capaces de maximizar la automatización de la generación de metadatos para IG.

Creación manual vs. automática

Como se refleja en [141], el hecho de que la creación de metadatos no ocurra de forma simultánea a la creación de los datos supone en muchos casos la falta de información, lo que algunas veces hace que la creación de metadatos resulte ser una tarea imposible [126] [16] [36] [131]. Este hecho, junto con los extensos y complicados estándares de metadatos de IG (p.ej. ISO19115 define más de 400 elementos), resulta en que la creación manual de metadatos resulta ser una tarea monótona, desagradable y costosa [15] [220]. Además, la consecuencia natural de que la creación de metadatos se realice de forma manual es la presencia de errores [144].

Por otra parte, de acuerdo a la revisión de la literatura relacionada que se realiza en [143], algunos autores sugieren que las técnicas de creación automática de metadatos pueden producir resultados de una

calidad razonable sólo bajo ciertas circunstancias [132]. Sin embargo, otros afirman que los metadatos creados de forma automática tienden a ser más eficientes, consistentes y menos costosos que aquellos creados de manualmente [5]. Otros mantienen que el desarrollo y la adopción de herramientas para la catalogación automatizada de los recursos simplificaría el trabajo, pese a la dificultad de programarlos [65]. Finalmente, otros autores proponen una combinación de métodos automáticos y manuales para producir una documentación de calidad [93] [51].

El proyecto ARIADNE[37] explora las posibles vías para la creación de metadatos, contemplando la creación manual, automática y mixta tanto por parte del creador de los datos como por parte de especialistas en información [99]. La Figura 5 muestra estas posibles vías de creación de metadatos: (a) automática, (b) automática mejorada por el creador de los datos, (c) automática mejorada por el creador de los datos y por el especialista en información, (d) manualmente por el creador de los datos y mejorada por el especialista en información o (e) manualmente por el especialista en información.

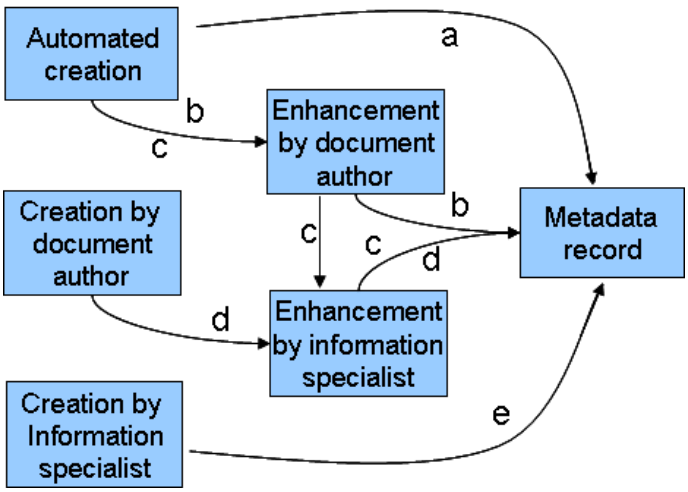

Figura 5: Posibles vías de creación de metadatos [99]

Aunque teóricamente solo sería necesario completar manualmente las descripciones subjetivas, como el resumen o el título, actualmente los metadatos normalmente son creados de forma manual, y sólo

[37] http://www.ariadne.ac.uk

algunos de ellos son extraídos automáticamente. Esto es debido a que la complejidad y variedad de los formatos limita la aplicación de técnicas automáticas.

Métodos para la generación de metadatos

Una de las primeras referencias que aparecen sobre el tema en el contexto de la IG [16] propone cinco métodos para la generación de metadatos: manualmente (a través del teclado); extendiendo la información almacenada con valores obtenidos mediante consultas; medidas y observaciones automatizadas; extrayendo y calculando; y finalmente por inferencia en base a otros elementos. Otros autores [95], identifican dos métodos para la creación de metadatos automatizada: la extracción y la recopilación. El primero de ellos emplea técnicas de minería de datos para la recuperación de información. El segundo, pretende recopilar información ya existente.

Ampliando las ideas anteriores podemos enumerar los siguientes métodos o técnicas para la generación de metadatos:

1. Introducción manual por teclado
2. Extracción de metadatos del propio recurso
3. Extracción de metadatos a partir del contenido
4. Recolección en el proceso de creación de los datos
5. Aprovechamiento del contexto
6. Búsqueda (look-up) desde una tabla de referencia
7. Medición del valor
8. Computación del metadato
9. Inferencia del metadato

El primer método es bien conocido, la introducción manual de los metadatos a través del teclado es el método por defecto en la mayoría de los casos hoy en día: el usuario edita una ficha empleando un editor de metadatos más o menos sofisticado. El problema es la cantidad de tiempo y recursos necesarios que supone este método, por ello resulta un método poco eficiente [5]. Para entenderlo basta con imaginar una situación en la que se debe crear una ficha de cada uno de los libros de una biblioteca, aparte de lo laborioso de la tarea es posible encontrarse problemas como la falta de información [36] de alguno de los ejemplares o cometer errores a la hora de copiar los datos a la ficha [144].

El segundo método, la extracción de metadatos del propio recurso, resulta bastante obvio. Los propios datos pueden proporcionar gran cantidad de información sobre ellos mismos si se analizan correctamente. Por ejemplo, cualquier archivo de cualquier sistema operativo contendrá información como la fecha de creación y modificación o el tamaño que ocupa en disco.

El tercer método es la extracción de metadatos a partir del contenido. El contenido de los datos es una fuente de información muy importante. Analizando los datos se puede encontrar información relevante de forma explícita, por ejemplo, en un correo electrónico resulta fácil encontrar información como el remitente, el destinatario o la fecha de envío.

El cuarto método es la recolección de información durante el proceso de creación de los datos. Esta resulta ser una fuente de información “volátil”, pues solo se dispondrá de ella en el momento en que los datos son creados y por ello es necesario extraer y almacenar toda la información posible en el momento. Se considera que la información que se puede obtener del proceso de creación de los datos es muy importante y raramente tenida en cuenta. Mediante este método es posible averiguar con exactitud información relevante como el proceso de creación para poder replicar los resultados más adelante, costes (computacional, temporal, económico, etc.) o el autor de los datos.

El quinto método trata de aprovechar la información del contexto en que los datos son creados o explotados. Del contexto de creación se puede obtener información relevante, como la organización o empresa responsable de los datos y la temática de los mismos, ya que los datos que genera una empresa dedicada al análisis del estado de la bolsa probablemente serán de carácter económico. Podemos operar de forma similar con el contexto de explotación obteniendo información como la temática o la calidad que suelen tener los recursos que ofrece cierta empresa.

El sexto método supone que un elemento de metadatos se crea a través de una correspondencia con otro. Por ejemplo, se puede obtener el topónimo correspondiente a los datos usando las 4

coordenadas de su caja envolvente, a través de un servicio de nomenclátor (*Gazetteer*[38]).

El séptimo método implica que, en el proceso de creación de los datos, un sensor u otro mecanismo de medición puede proporcionar información relevante. Se podrían medir magnitudes como elevación, posición o temperatura, y colocar esos valores en la ficha de metadatos de forma automática. Podemos encontrar un buen ejemplo de este método en algunas cámaras digitales que lo utilizan para agregar, entre otros, la información que les proporciona su dispositivo GPS integrado a las imágenes en forma de etiquetas EXIF[39].

El octavo método se centra en al cálculo de un elemento de metadatos empleando los datos en sí o los propios metadatos. En este sentido hay muchas líneas de investigación abiertas que abarcan un amplio abanico de posibilidades. Se pueden encontrar desde diferentes técnicas para realizar un análisis/procesado de documentos de texto o páginas web para averiguar su tema principal, a otras técnicas que emplean los propios datos para, por ejemplo, determinar la provincia de un pueblo por cálculos topológicos.

El noveno y último método es la inferencia de metadatos a partir de otros metadatos o de los datos. Según algunos autores [16] supone el mejor método –de hecho en algunas situaciones el único– para la creación de metadatos a posteriori, es decir, documentando recursos ya existentes. Un ejemplo sería inferir la época de los datos por el metadato temperatura, a lo mejor recogido por el séptimo método que hemos explicado, de manera que una regla establecería que para temperatura inferior a 15 grados en Tenerife supongamos invierno. La creación de estos metadatos inferidos solapa ampliamente a los campos de investigación de la minería de datos y de la recuperación de datos [16] [91].

De entre las técnicas de minería de datos que sería posible aplicar [100] destacamos: el análisis exploratorio de los datos, cuyo objetivo es explorar los datos sin una idea clara de lo que buscamos; el modelado descriptivo, que pretende describir todos los datos, por ejemplo mostrando su distribución, agrupándolos o mostrando posibles relaciones entre variables; el modelado predictivo, cuyo

[38] http://en.wikipedia.org/wiki/Gazetteer
[39] http://www.exif.org

objetivo es construir un modelo que permita predecir una variable en base a los valores de otras variables conocidos; métodos de descubrimiento, basados en la detección de patrones, y cuya idea es identificar patrones, reglas o combinaciones de elementos que ocurren con frecuencia; la recuperación por contenido, se basa en la comparación de los datos de acuerdo al patrón de interés para buscar patrones similares.

Algunos trabajos basados en la idea de inferir información en base a técnicas relacionadas con la minería de datos proporcionan un *framework* para generar metadatos a partir de documentos electrónicos como imágenes o documentos de texto [123]. Otros generan metadatos para publicaciones a partir de sus referencias bibliográficas [57] y otros analizan fotografías para inferir información como si fue hecha de día o de noche, en un entorno natural o urbano, en interior o en exterior, etc. [28] [201].

Otros autores han explorado cómo las técnicas de minería de datos pueden extraer información de la IG. Algunos destacan que los metadatos conformes a los estándares de IG presentan algunos problemas a la hora de aplicar algoritmos de minería de datos, dado que son datos con muchas dimensiones, contando en muchos caso con cientos de atributos representados en varios formatos y normalmente con muchos atributos en blanco, especialmente por los campos opcionales de los estándares [203]. Sin embargo, otros autores [64], creen que los métodos de minería de datos pueden ser un buen medio para extraer información útil a partir de otros recursos. En este sentido, propone tanto la generación automática como la recuperación de metadatos mediante el empleo de estas técnicas.

Como se puede observar, todos estos métodos proporcionarán información importante sobre los recursos que hoy en día se está dejando escapar. Mientras la mayoría de los métodos son aplicables durante todo el ciclo de vida de los datos, otros métodos solo serán aplicables en el momento en que los datos son creados. Resultando de especial importancia estos últimos dado que la información que no se recoge en ese momento se pierde para siempre, y alguna de esta información puede resultar esencial para conseguir una correcta descripción de los recursos.

Workflows para la creación de metadatos

Los métodos propuestos pueden ser combinados para proporcionar una respuesta a las necesidades más completa.

Nuevamente, de acuerdo a la revisión de la literatura relacionada que se realiza en [143], algunos autores sugieren el uso de plantillas de metadatos como medida para facilitar el proceso de creación de los metadatos [102]. En este sentido, existen trabajos [158] que proponen un *framework* para la generación de metadatos cuyo punto de inicio es, como podemos ver en la Figura 6, la definición de una plantilla de metadatos para la entidad (a), personalizar dicha plantilla para el recurso específico (b), en base a ella, si los metadatos existen, se procesan para adaptarlos a la plantilla (c), y si no existen, se crean (d). A continuación, se añade toda la información relativa a su procedencia (e) y finalmente se sincroniza con otros metadatos provenientes de alguna herramienta comercial.

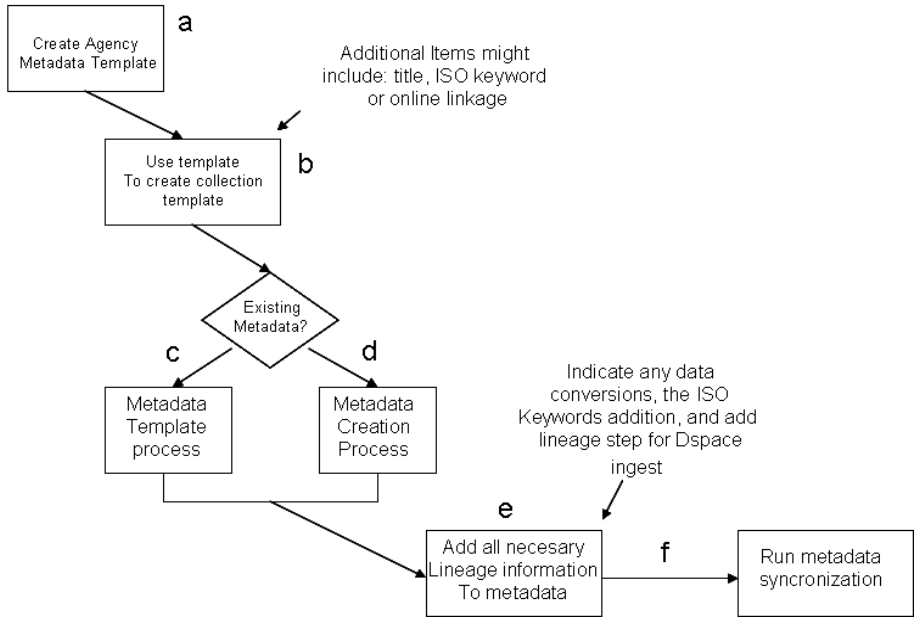

Figura 6: Ejemplo de *workflow* para la creación de metadatos [158]

Otros autores [203] proponen la generación de metadatos en tres pasos. El primer paso consiste en aplicar algunas técnicas de extracción de metadatos en base al formato de datos específico de la IG. El siguiente paso es la deducción automática de información sobre la calidad de los datos usando técnicas estocásticas, de comparación

o por fuerza bruta. Finalmente, propone aplicar técnicas de minería de datos que permitan refinar el conocimiento del recurso.

Por otra parte, en [175] se propone un marco para la automatización de la creación de los metadatos basado en la creación automática, el enriquecimiento y la actualización. Según los autores, la creación automática será necesaria cuando no existan metadatos asociados a la IG. El enriquecimiento implica la mejora del contenido de los metadatos mediante la monitorización de las etiquetas aplicadas por los usuarios cuando realizan las búsquedas de los recursos, y posteriormente la actualización automática de los metadatos en base a dichas etiquetas.

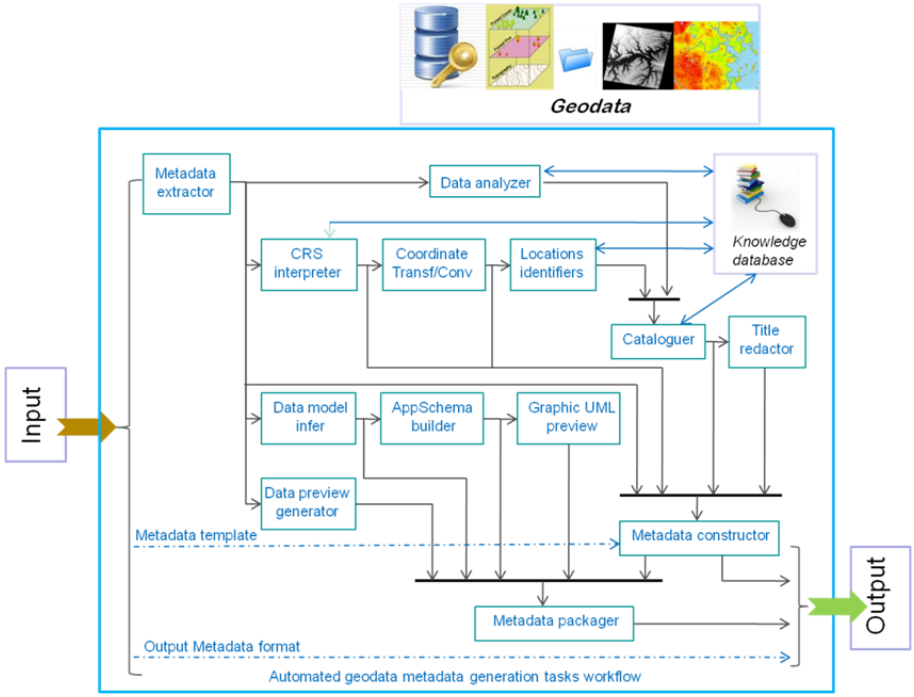

Figura 7: Ejemplo de *workflow* para la creación de metadatos [142]

Otros trabajos [64] presentan una arquitectura basada en la inferencia de metadatos a través de técnicas de minaría de datos para recursos dentro de la red de una organización. En esta arquitectura, en inicialmente es necesario seleccionar los datos, posteriormente estos datos son procesados para extraer atributos y alimentar la base de conocimiento. Tanto la extracción de atributos como el proceso de datos se aplican a todos los documentos disponibles (series temporales, datos espaciales y datos web) con el fin de alimentar los

procesos de minería de datos. Este conocimiento es enriquecido gracias a las palabras clave proporcionadas por los usuarios en el proceso de explotación.

Finalmente [142] propone otro *workflow* para la creación de metadatos. Como se puede ver en la Figura 7 en base a los metadatos extraídos del recurso, se analizan para inferir el tipo de datos que contienen, se tratan sus coordenadas para convertirlas a un sistema de referencia común y se analiza su formato de datos. Finalmente, con toda la información recopilada, se construye un registro de metadatos en base al formato de salida deseado.

Estándares de metadatos

Los estándares permiten la definición precisa de criterios comunes para una actividad o uso específico. Se considera un estándar de metadatos, a una forma genérica de organizar los metadatos, un modelo según el cual se pueden organizar los metadatos de un objeto. Los estándares para metadatos varían de uno a otro, básicamente, en la información que consideran importante, es decir, que consideran como un metadato.

Se han definido recomendaciones para la creación de metadatos, cuya finalidad principal es proporcionar una estructura "jerárquica y concreta" que permita describir exhaustivamente cada uno de los recursos a los que hacen referencia. Han sido creadas y aprobadas por organismos de normalización a partir de opiniones de expertos en esta materia. Estas recomendaciones, en forma de normas o esquemas de metadatos, proporcionan criterios para caracterizar sus recursos con propiedad.

Dentro de un sistema de información abierto, los metadatos deben ser representados en base a una norma común o estándar por cuestiones de interoperabilidad. Por lo tanto, tras ser creados, los metadatos deberán organizarse en base a un estándar que permita publicarlos. A nivel general el estándar de metadatos más conocido es *Dublin Core*[40] y en el contexto de la IG es ISO19115:2003[41] y sus perfiles derivados. Dado que realmente los estándares no influyen en la creación de los metadatos si no que son necesarios para su

[40] http://dublincore.org
[41] http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

publicación en un entorno interoperable, veremos más detalles sobre los diferentes estándares en el Capítulo 3 dedicado a la publicación.

Herramientas de creación de metadatos

Algunos autores [12] sugieren que el uso de entornos informáticos interactivos para la creación y visualización de metadatos facilitarían el trabajo. Otros [93], tras revisar diferentes herramientas de creación de metadatos, afirman que su uso permite dirigir los esfuerzos de los usuarios a aspectos que requieran inteligencia. Dependiendo del grado de automatización y de los requerimientos de participación por parte de los usuarios en el proceso de creación de los metadatos, se distingue entre generadores y editores en los cuales los procesos automáticos y humanos están integrados.

Normalmente las aplicaciones que nos permiten crear contenido textual, imágenes, audio o video, incorporan cierta funcionalidad de generación automática de metadatos para los recursos que se crean con ellas. Por ejemplo, *Microsoft Office*, además de algunos metadatos como la fecha de creación y modificación o el autor, incorpora al documento un título basado en el texto de la primera línea. Estos metadatos a menudo son utilizados por el sistema de archivos para indexar y clasificar el contenido. En este sentido, existen estudios [94] sobre la creación de metadatos para diferentes tipos de recursos y formatos específicos.

Como primer ejemplo de herramienta específica para la creación de metadatos podemos referirnos al proyecto *Apache Tika*[42]. Este es un proyecto de la Fundación Apache[43] distribuido bajo una licencia libre. El proyecto ofrece un conjunto de herramientas para la detección y extracción de metadatos y de contenido del texto estructurado de varios tipos de recursos reutilizando librerías de existentes para el acceso a los datos de formatos específicos. Actualmente soporta numerosos formatos, incluyendo varios formatos de texto, audio, imagen y video[44].

[42] http://tika.apache.org
[43] http://www.apache.org
[44] http://tika.apache.org/1.2/formats.html

Otro ejemplo de herramienta es la *Metadata Extraction Tool*[45], desarrollada por la *National Library of New Zealand*[46] para la extracción de metadatos que permitan la preservación de recursos de una amplia gama de formatos de archivos como documentos PDF, imágenes, archivos de sonido, documentos de Microsoft Office y muchos más.

En el contexto de la IG, como consecuencia de la creciente necesidad de metadatos para permitir la búsqueda de la enrome cantidad de recursos disponibles en entornos distribuidos como las IDEs, a lo largo de los últimos años se han venido desarrollando una gran cantidad de aplicaciones software que, ya sea como herramientas independientes o como *plug-ins*[47] dentro de otras aplicaciones, facilitan en gran medida la creación de dichos metadatos. La mayoría de estas aplicaciones son presentadas como editores de metadatos y, algunas de ellas, tienen capacidades para generar de forma automática algunos metadatos concretos. La cantidad de información que se puede extraer de un recurso depende fundamentalmente del modelo de representación utilizado y de su propio formato de archivo [147]. De esta forma, hay elementos que sólo pueden ser extraídos de ciertos tipos de datos y archivos, sin embargo otros, como el tamaño de los datos, pueden ser obtenidos en cualquier caso. A continuación se presentan las aplicaciones más conocidas y sus características principales son resumidas en la Figura 8.

MetaLite[48] es una de las herramientas de edición de metadatos más sencilla y, al mismo tiempo, más utilizada. Esta aplicación, desarrollada por el FGDC, es de libre uso y únicamente proporciona soporte para un conjunto de elementos de la norma. Está preparada para ejecutarse sobre sistemas Windows y, entre sus características destacan: intercambio de metadatos en formato HTML, TXT y SGML (XML), traducción de la aplicación a cuatro idiomas (español, inglés, francés y portugués), integración de pequeños diccionarios para completar las palabras clave en esos cuatro idiomas, etc.

---

[45] http://meta-extractor.sourceforge.net
[46] http://natlib.govt.nz
[47] Un *plug-in* es una aplicación complementaria que se relaciona con otra para aportarle una función nueva y generalmente muy específica, siendo ejecutada por la aplicación principal.
[48] http://edcnts11.cr.usgs.gov/MetaLite

M3Cat[49] es una aplicación desarrollada por la compañía canadiense *Intelec Geomatics Inc*. Se trata de una aplicación Web que almacena los metadatos siguiendo diferentes perfiles y estándares en una base de datos Access u Oracle. La principal característica de esta herramienta es el soporte a los niveles jerárquicos de los metadatos. Sin embargo, por el momento, este soporte se limita a la copia de la información contenida en el metadato del padre al metadato del hijo durante la creación de este último.

MetaD[50] es una herramienta para crear y editar metadatos, desarrollada por la Infraestructura de Datos Espaciales de Cataluña (IDEC), que permite la edición y exportación de metadatos siguiendo el perfil IDEC, subconjunto de la norma ISO 19115:2003, con su implantación ISO 19139:2007, destinado a describir la IG (gráfica, alfanumérica, etc.).

IME[51], elaborada por el Instituto Nacional de Técnica Aeroespacial, permite la definición de perfiles de ISO 19115:2003 mediante ficheros de configuración y su posterior edición. Además de la posibilidad que brinda al usuario de definir sus propios perfiles, presenta otras características relevantes, como la validación de los metadatos y la traducción de la aplicación a tres idiomas diferentes (inglés, español y polaco).

GeoNetwork[52] es una herramienta desarrollada por la FAO-UN, WFP-UN y UNEP. Se trata del catálogo web de metadatos geográficos por excelencia. Incluye un buscador de metadatos, herramientas de edición e importación, y utilidades de administración. Soporta metadatos creados conforme a los siguientes esquemas de codificación: ISO 19115:2003, CSDGM y Dublin Core. A partir de la versión 2.1 incluye también soporte para ISO 19139:2007.

CatMDEdit[53] es un software de código abierto que facilita la documentación de recursos, que presta especial atención en la descripción geográfica de los mismos [232]. Esta herramienta ha sido desarrollada por el Grupo de Sistemas de Información Avanzados de

[49] http://www.intelec.ca/html/en/technologies/m3cat.html
[50] http://www.geoportal-idec.cat/geoportal/cas/meta-d
[51] http://www.crepad.rcanaria.es/metadata/index.htm
[52] http://geonetwork-opensource.org
[53] http://catmdedit.sourceforge.net

la Universidad de Zaragoza y por la empresa GeoSpatiumLab S.L., con el patrocinio del Instituto Geográfico Nacional. Desarrollada en Java, destaca por sus características multiplataforma y multilingüe, proporcionando herramientas que facilitan la creación, manipulación y publicación de los metadatos para IG.

	Metalite	**M3CAT v1.6**	**MetaD v3.0.4**	**IME v4.1**	**GeoNetwork v2.2.0**	**CATMDEdit v.4.0.1**	**ESRI ArcCatalog v9.2**
Sistemas operativos	Windows	Aplicación Web	Windows	Multi-plataforma	Aplicación Web	Multi-plataforma	Windows
Lenguajes de programación	Visual Basic	ASP	Visual Basic	Java	Java, Javascript, XLS	Java, XLS	C++, XLS
Idiomas	EN, ES, FR, PT	FR, EN	CA, EN, ES, IT, EL	ES, EN, PL	EN, FR, ES, CN (others available but not in web interface)	EN, ES, FR, PL, PT, CS	EN
Open Source	•	•			•	•	
Normas soportadas	CSDGM	ISO19115:2003 CSDGM CSDGM_NBII GILS	ISO19115:2003 ISO19119:2005	ISO19115:2003	ISO19115:2003 ISO19119:2005 CSDGM Dublin Core	ISO19115:2003 (versión completa de la norma, y perfiles específicos: NEM, Core, INSPIRE, WISE). CSDGM Dublin Core	ISO19115:2003 CSDGM
Utilización de Tesauros		•	•		•	•	
Ayuda de la aplicación	Ayuda en línea	FAQ, manual de usuario y tutorial	FAQ, manual de usuario Ayuda en línea	FAQ disponible en la página web	Manual de usuario en PDF y HTML, menajes de ayuda en el editor	Manual de usuario en PDF, ayuda en línea	Manual de usuario
Generación automática de metadatos						•	•
Validación		ISO19115:2003	ISO19115:2003 conforme a ISO19139:2007 Plantillas Personalizadas	ISO19115:2003 conforme a ISO19139:2007 e ISO19119:2005; CDSGM y Dublin Core	ISO19115:2003 conforme a ISO19139:2007; Dublin Core y validación de perfiles	ESRI-ISO	
Configuración de la edición		•			•	•	•
Sistemas de búsqueda		•	•		•	•	•
Formatos de importación/ exportación	HTML, txt, sgml (XML)	XML, ASCII txt, zip	XML	XML, HTML, PDF, txt	XML, HTML, MEF, text, RSS, GeoRSS, OpenSearch DC interface	XML, RDF, HTML, Excel	XML, HTML

Figura 8: Tabla comparativa de herramientas de creación de metadatos para IG

ArcCatalog, desarrollada por ESRI como una funcionalidad añadida a *ArcGIS*[54], es quizá una de las aplicaciones de creación de metadatos más utilizadas. Esta herramienta permite la carga automática de ciertos elementos básicos y la actualización sincronizada de datos y metadatos. Los metadatos son almacenados en formato XML junto a los ficheros de datos que describen, o bien en una base de datos. Permite la personalización de los editores de metadatos y de los estilos de presentación en base a nuevas plantillas definidas por el usuario, y proporciona asistencia para la publicación de los metadatos.

En la Figura 8 se muestra un resumen comparativo de las características principales de las aplicaciones de edición de metadatos comentadas anteriormente.

2.1.4 Anotación de Recursos

En esta sección se presentan y analizan diferentes opciones para incorporar metadatos a los recursos, finalmente se discutirán sus ventajas y desventajas.

Técnicas para la georreferenciación de recursos

La georreferenciación de recursos puede realizarse usando formatos que soportan información de geolocalización de forma nativa como el JPEG2000 [202] [194] o la familia MPEG[55] en el contexto de la multimedia. No obstante, la existencia de este tipo de formatos no está presente en todas las áreas y es por ello que en muchas ocasiones se trata de forma separada los datos y sus metadatos o información de localización. Una de las causas principales para esta carencia es el uso de formatos antiguos e incapaces de incluir información sobre su geolocalización que no fueron diseñados para tal fin. A pesar de esto y debido a la gran cantidad de información disponible en estos formatos, distintas soluciones han aparecido para portarlos a entornos SIG y así relacionarlos con información geográfica.

[54] http://www.esri.com/software/arcgis
[55] http://www.mpeg.org

En base a esto, se consideran dos familias de soluciones. Por una parte, las soluciones orientadas a la modificación interna del recurso para añadir la información necesaria. Por otra parte, las soluciones basadas en la anotación externa del mismo y su posterior encapsulación junto a sus metadatos como una unidad tanto a nivel físico como lógico.

Técnicas orientadas a la modificación interna

Las técnicas basadas en la modificación interna del fichero para incorporar metadatos son extremadamente dependientes del formato del recurso que se intenta georreferenciar. Básicamente, la mayoría de las soluciones basadas en este tipo de técnicas añaden metadatos en ciertos sectores de los archivos, normalmente en las cabeceras, que se considera que no tienen una tarea asignada o no se usan.

Un buen ejemplo de técnica de modificación interna es Adobe XMP (*Extensible Metadata Platform*) [3]. Esta técnica de etiquetado permite la incorporación de metadatos dentro del propio recurso y proporciona una forma fácil de embeber información relevante directamente en el recurso. El problema de esta técnica es que sólo puede ser aplicada sobre ciertos formatos (TIFF[56], JPEG[57], PNG[58], GIF[59] y PDF[60]) por lo que no cubre las necesidades de los usuarios que requieren otros formatos (audio, video, etc.). Además, esta técnica, dependiendo del formato, incorpora los metadatos en diferentes partes del archivo, incluyendo segmentos de la cabecera del archivo o intenta aprovechar etiquetas diseñadas para otros propósitos. A pesar de que en un principio se asegura que esta técnica no daña el archivo [3], podemos encontrarnos en el caso en que los recursos modificados nos son válidos para algunas aplicaciones.

Una aproximación similar propone el uso de las etiquetas del *International Press Telecommunications Council* (IPTC) [134]. Este formato permite la inserción de metadatos en algunos tipos de archivos añadiendo etiquetas específicas a sus cabeceras. El método de las etiquetas IPTC representa un formato más antiguo y que poco a

[56] http://partners.adobe.com/public/developer/tiff/index.html
[57] http://www.jpeg.org
[58] http://www.libpng.org/pub/png
[59] http://es.wikipedia.org/wiki/Graphics_Interchange_Format
[60] http://www.adobe.com/devnet/pdf/pdf_reference.html

poco está siendo sustituido por XMP. Un esfuerzo de colaboración entre ambos ha producido el *IPTC Core Schema for XMP* permitiendo la combinación de las dos aproximaciones para incrustar los metadatos.

Otra solución interesante basada en cambios internos enfocada al mundo de las imágenes y ampliamente adoptada por los fabricantes de cámaras digitales es la incorporación de etiquetas *Exchangeable Image File Format*[61] (EXIF) [118] en archivos de imagen, especialmente aquellos que usan compresión JPEG. De nuevo el problema reside en el limitado número de formatos que soporta y los posibles daños causados al archivo.

Todas estas técnicas se basan en la incorporación de etiquetas en las cabeceras de un limitado grupo de formatos. Estas cabeceras varían de un formato a otro, demostrando la falta de consistencia entre ellos. Además, dichas cabeceras pueden ser ignoradas por algunas aplicaciones pero usadas para fines distintos en otras lo que acarrea el riesgo de crear archivos que no pueden ser utilizados en todas las aplicaciones capaces de interpretar la especificación original del formato en el que está representado el recurso.

Técnicas orientadas a la anotación externa

La mayoría de las técnicas actuales que evitan modificar el recurso están basadas en la anotación de los recursos desde archivos externos como en el caso de *Synchronized Multimedia Integration Language*[62] (SMIL) [33], o la generación de otros archivos estructurados como en el caso de los archivos *".world"* [63]. Usando estas técnicas, el recurso permanece intacto, pero se pierde la noción de unidad que integra datos y metadatos, dificultando su manejo y compartición.

Tratando de solucionar este problema están emergiendo formatos que encapsulan algunos de estos archivos externos con metadatos, junto con los datos y otros recursos en un sólo archivo. Este es el caso del *Metadata Exchange Format* (MEF) [170] o de KMZ[64], un

[61] http://www.exif.org
[62] http://www.w3.org/TR/REC-smil
[63] http://en.wikipedia.org/wiki/World_file
[64] http://www.google.com/earth/outreach/tutorials/kmz.html

formato ampliamente usado en aplicaciones de georreferenciación y servicios de mapas que explicaremos en detalle más adelante.

El formato MEF fue específicamente creado para el intercambio de datos y metadatos entre diferentes plataformas y especialmente entre nodos de *GeoNetwork*[65]. MEF se centra en facilitar tareas como el almacenamiento, la transferencia y migración de datos espaciales, metadatos, *thumbnails*, privilegios básicos y otra información relacionada. Su estructura interna está formada por un archivo XML que contiene los metadatos del recurso y otro archivo XML con un formato específico que especifica información adicional para GeoNetwork. Los archivos MEF permiten el transporte de archivos con IG, encapsulados en un único archivo, junto con sus propios metadatos. Este enfoque orientado al encapsulamiento facilita el intercambio, distribución y reutilización de los recursos.

Finalmente, la Web 2.0 promueve soluciones orientadas a georreferenciar algunos recursos a través de Internet. Este es el caso de algunas de las redes sociales más populares como *Flickr* o *Panoramio* que son utilizadas para compartir y georreferenciar imágenes o *Wikiloc*[66] utilizado con *tracks* de GPS. Estos *mashups* suelen almacenar los recursos usuario y anotar su geolocalización dentro de sus bases de datos. Los metadatos incluidos por este tipo de soluciones son muy limitados y, además, están orientados a un caso de uso y formato específicos, de modo que no son aplicables como una solución general.

Discusión sobre técnicas de anotación

Como se ha visto, las dos familias de soluciones de georreferenciación descritas ofrecen sus beneficios pero también sus desventajas.

La principal ventaja de las soluciones orientadas a la modificación interna del recurso es la integración total de los datos y sus metadatos en un mismo archivo. Así, se facilita en gran medida el transporte, la gestión y la difusión del recurso junto con sus metadatos. Sin embargo, esta familia de técnicas tiene varios inconvenientes derivados principalmente de la manipulación del formato original para

[65] http://geonetwork-opensource.org
[66] http://www.wikiloc.com

añadir más información. La desventaja más obvia es que la manipulación de los archivos originales puede suponer que queden inutilizables para otras aplicaciones. También tenemos que tener en cuenta que esta familia de técnicas incorpora metadatos adicionales en diferentes lugares dependiendo del formato del archivo original, siendo muy bajo el número de formatos que lo soportan. Además, la cantidad de información que se puede añadir es limitada porque, en la mayoría de los casos, la información se incorpora en segmentos o etiquetas con una capacidad limitada. Otra desventaja notable es la imposibilidad de añadir otros recursos tales como licencias o *thumbnails*. Asimismo, con el uso de técnicas de modificación interna dependemos de si el formato es conocido y abierto al cambio, y si hay controladores disponibles que nos proporcionen la capacidad de lectura/escritura para cambiar los metadatos. Es decir, no existe la misma libertad de manipulación para todos los formatos.

Por otra parte, las técnicas orientadas a la anotación externa combinadas con la encapsulación, posibilitan la integración de los datos y sus metadatos en una unidad y, además, permiten incluir otros recursos relacionados sin ninguna limitación. Otra ventaja de este tipo de técnicas es que no hay restricciones en el tamaño de los metadatos o datos; incluso podemos hacer uso de técnicas de compresión. Continuando con las ventajas, hay que señalar que estas técnicas son válidas para todos los formatos actuales y futuros, además, no alteran los archivos originales por lo que quedan exentos de este tipo de errores asociados. El principal inconveniente de esta familia de técnicas es que para operar con los archivos se requiere un preproceso para *desencapsular* o extraer los recursos, ya que no están directamente disponibles. Por otra parte, aunque los datos y los metadatos se encuentran encapsulados dentro de un mismo archivo, están separados por lo que son más difíciles de manejar.

En nuestra opinión, la integración en una sola unidad de datos y metadatos es imprescindible, dado que facilita su transporte e intercambio. Además, pretendemos conseguir una solución genérica, es decir, que la solución permita georreferenciar e incorporar metadatos a cualquier recurso actual y que pueda ser fácilmente aplicable a formatos de datos futuros. Por todas estas razones, como se verá en la Sección 2.3, se ha optado por una solución orientada a la anotación externa combinada con la encapsulación.

XML y RDF: lenguajes generales de representación de metadatos

Además de conocer las técnicas para la incorporación de los metadatos a los recursos, también debamos conocer algunos lenguajes que lo posibilitan. En este sentido, el W3C ha mostrado un fuerte interés en los metadatos al desarrollar lenguajes de representación como *eXtensible Markup Language*[67] (XML) [29] o *Resource Description Framework*[68] (RDF) [139].

Respecto a la utilización de XML como lenguaje de representación de metadatos en el contexto de los metadatos geográficos, cabe destacar que hay un consenso bastante generalizado. Tal como se menciona en [165], los sistemas de catalogación de metadatos deben soportar (reconocer) tres formatos de metadatos: el formato de implementación (dentro de una base de datos o sistema de almacenamiento), el formato de exportación o codificación (diseñado para la transferencia de metadatos entre distintos sistemas y computadores), y el formato de presentación (un formato apropiado para ser leído por las personas). Para los dos últimos formatos hay un consenso general respecto al uso de XML dado que es un lenguaje de marcado con reglas estructurales forzadas a través de un fichero de control (*XML-Schema*) que permite validar la estructura del documento, es decir, comprobar la conformidad respecto a una norma o a un estándar de metadatos. Además, a través de una especificación complementaria *eXtensible Stylesheet Language*[69] (XSL) [227], un documento XML puede ser usado junto a una hoja de estilo (expresada en XSL) para crear presentaciones o informes según los requerimientos del usuario.

En cuando a la codificación de los metadatos en XML, cabe destacar la especificación técnica ISO/TS 19139:2007 [114], la cual define la forma de convertir los modelos UML de la norma ISO 19115:2003 (y otras relacionadas) en la sintaxis adecuada sobre XML. Un archivo de intercambio de metadatos, acorde con la norma ISO 19115:2003 en formato XML, va a ser un documento XML que siga la sintaxis definida por la especificación técnica ISO 19139:2007. Es decir, esta especificación técnica define un conjunto de esquemas en

[67] http://es.wikipedia.org/wiki/XML
[68] http://www.w3.org/RDF
[69] http://www.w3.org/Style/XSL

XML que van a describir los metadatos asociados a cada nivel de información, permitiendo así su descripción, asegurando su validación y su posterior intercambio a través de archivos de metadatos. Estos esquemas XML se han generado a partir los modelos UML definidos en ISO 19115:2003 aplicando las reglas de codificación definidas en la norma ISO 19118 *Geographic Information-Encoding*. Esta norma establece un conjunto de reglas de codificación para transformar los esquemas conceptuales UML, descritos en cualquiera de las normas de la serie ISO 19100, en esquemas XML.

Por otra parte, si hablamos de lenguajes de representación de metadatos en contextos más generales, no podemos olvidarnos del uso extensivo de RDF. Seguramente, el uso de RDF como lenguaje de representación de metadatos se debe a su utilización para expresar modelos de metadatos muy extendidos como Dublin Core, y a su utilización como tecnología básica en la nueva concepción de la Web: la Web Semántica. Según [21], *"la Web Semántica es la extensión de la Web actual dentro de la cual la información recibe un significado bien definido, permitiendo que computadores y personas puedan trabajar en cooperación"*. Por lo tanto, RDF nos permite anotar semánticamente los recursos disponibles.

En conclusión, tanto XML como RDF son lenguajes generales que nos permiten la representación de metadatos. Además, permiten fijar la estructura de los documentos y su validación. De forma que, los estándares definen el contenido y la estructura de los metadatos mientras que estos lenguajes nos permiten representarlos físicamente y validar que cumplen dicho estándar.

En el contexto de la IG, XML proporciona el soporte necesario para la representación de los metadatos en base a los diferentes estándares definidos. Además, en este mismo contexto, han aparecido otros lenguajes o gramáticas basadas en XML que nos permiten la anotación de recursos georreferenciados.

Geography Markup Language

El *OpenGIS Geography Markup Language Encoding Standard*[70] (GML) es una iniciativa de OGC [177] y en 2007 pasó a ser también un estándar ISO (ISO 19136:2007). GML es una gramática XML para

[70] http://www.opengeospatial.org/standards/gml

expresar características geográficas. Sirve como un lenguaje de modelado para sistemas geográficos, así como un formato abierto de intercambio para las transacciones de IG en Internet.

Al igual que la mayoría de las gramáticas basadas en XML, hay dos partes de la gramática: el esquema que describe el documento y la instancia del documento que contiene los datos reales. Un documento GML se describe utilizando un esquema GML. Esto permite que los usuarios y desarrolladores recursos con IG genéricos que contengan puntos, líneas y polígonos. Algunas comunidades han definido diferentes esquemas de aplicación, que son extensiones especializadas de GML. Usando esquemas de aplicación, los usuarios pueden hacer referencia a caminos, carreteras y puentes en lugar de puntos, líneas y polígonos. Si todos los miembros de una comunidad se comprometen a utilizar los mismos esquemas pueden intercambiar datos fácilmente y estar seguros de que un camino es un camino aun cuando lo ven. El uso de GML es bastante extendido, por ejemplo los clientes y servidores con interfaces que implementan WFS[71] permiten leer y escribir datos representados en GML.

Keyhole Markup Language

El *Keyhole Markup Language*[72] (KML) es un lenguaje de marcado basado en XML para representar IG. KML es un estándar abierto inicialmente impulsado por Google, pero que en 2008 se convirtió en un estándar OGC[73] [178]. KML es un lenguaje centrado en la visualización geográfica, incluyendo la anotación de mapas e imágenes. La visualización geográfica no sólo incluye la presentación de los datos gráficos, sino también el control de la navegación del usuario en el sentido de dónde ir y dónde buscar.

El uso de KML está muy extendido para codificar los diferentes recursos web [13] [58]. Su popularidad probablemente se debe a su simplicidad y sus capacidades de visualización y anotación [225], que permiten que KML sea ampliamente soportado por las herramientas geoespaciales y servicios web cartográficos más comunes.

[71] http://www.opengeospatial.org/standards/wfs
[72] https://developers.google.com/kml/?hl=es
[73] http://www.opengeospatial.org/standards/kml

KML proporciona algunos mecanismos para georreferenciar recursos incorporándoles detalles de visualización y algunos metadatos. En cuanto a visualización, KML ofrece un rico conjunto de opciones ya sea en entornos 2D y 3D [43] que se suman al conjunto de geometrías primitivas básicas como puntos o polígonos. Sin embargo, hay algunas cuestiones conceptuales y aspectos técnicos a tener en cuenta a la hora de abordar la anotación de cualquier tipo de recursos georreferenciados.

2.2 Creación de Metadatos

Como hemos visto en las secciones anteriores, potencialmente, cualquier recurso puede ser georreferenciado e integrado con otra IG dentro o fuera del contexto de las IDEs. Por otra parte, los metadatos ayudan a las personas involucradas en el uso de IG a encontrar la información que necesitan y a determinar la mejor forma de usarla [165], es decir, resultan ser una herramienta efectiva para describir nuestros recursos, posibilitando su descubrimiento, evaluación y acceso. Esta sección presenta una nueva metodología para la creación de metadatos y, en base a ella, se exploran dos casos de estudio reales que implementan parte de dicha metodología.

2.2.1 Metodología Propuesta

Llegados a este punto, ya conocemos los principales aspectos sobre la creación de metadatos, así como los métodos y herramientas disponibles para ello, gracias a la sección anterior dedicada al estudio del arte sobre el tema. En esta sección pasaremos a detallar la metodología que se propone como contribución y que permite generar de forma automática metadatos completos y de una calidad razonable, evitando tanto como sea posible la participación del usuario.

La clave del éxito es la óptima combinación de diferentes métodos (ver Sección 2.1.3); no se trata de uno sobre otro, sino el empleo de varios o todos en armonía. Pero en primer lugar hay que entender el problema y las posibles soluciones. Para citar sólo un ejemplo, en [78], cuyos autores eran expertos en bibliotecas digitales pero evidentemente no en las nuevas tecnologías, lamentan que de una

foto se pueda inferir quienes son las personas retratadas, pero no cuando fue tomada la foto.

> *"This class of strategies is likely to be successful only for metadata that can be inferred from the object itself. For example, it may be possible to determine the names of those present from a picture, but it is likely to be impossible to determine the time and data at which the picture was taken"*

Pensando en el siglo XXI, las fotos ya no son elementos en papel sino ficheros digitales, y mientras el reconocimiento automático de formas y contenido (quienes son las personas en esta foto) se nos resiste de momento (aunque ya hay soluciones parciales [41]), la mayoría de las fotos digitales actuales llevan metadatos internos creados por la propia cámara, que indican entre otras cosas la fecha, hora, e incluso la localización geográfica de la foto.

En consecuencia, la metodología para la generación de metadatos que se propone es una combinación de todos los métodos descritos anteriormente orquestados de forma eficiente. La Figura 9 resume de forma gráfica los diferentes métodos que forman parte de la metodología propuesta y que se detallan a continuación.

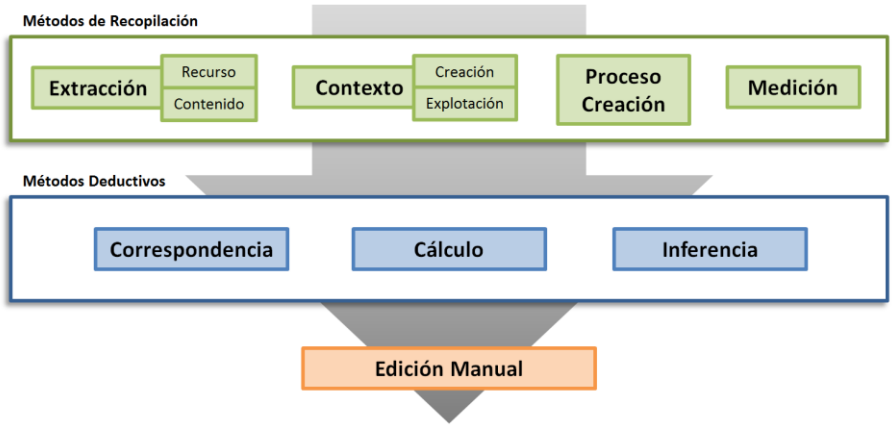

Figura 9: Metodología Propuesta para la Generación de Metadatos

Se empezará por obtener (extraer) toda la información relevante que se pueda obtener del propio recurso, por ejemplo el tamaño de los datos o las fechas de creación y modificación.

Después, se intentará extraer tanta información como sea posible del contenido. Debemos destacar que esta es una de las fuentes de

información más importantes, por lo que debemos prestar especial atención. La forma de analizar el recurso y la cantidad de información disponible dependerá completamente de la naturaleza del recurso y del formato mediante el que está representado. Mediante este método se puede extraer información explícita en los datos, por ejemplo, en un correo electrónico, es fácil encontrar información como el remitente, el destinatario o la fecha en la que fue enviado.

A continuación, se agrega la información común relativa al contexto de creación y de explotación del recurso. Del contexto de creación podemos obtener información relevante como la organización o la empresa responsable de los datos y el tema de los datos. Podemos operar de una forma similar con el contexto de explotación del recurso y obtener información como el tema o la calidad de los recursos ofrecidos por cierta organización. Toda esta información puede ser previamente establecida, revisada por el usuario, o automáticamente obtenida explorando el recurso y su contexto.

El siguiente paso es considerar la recolección de información del proceso de creación de los datos, obviamente si este existe. Debemos destacar que esta fuente de información es volátil dado que sólo estará disponible en el momento en el que el recurso es creado y por esa razón debemos obtener y almacenar toda la información posible en ese momento. Consideramos que la información que puede ser obtenida durante la creación de los recursos es muy importante y raramente tenida en cuenta. Mediante este método podemos obtener información de forma detallada sobre el proceso de creación para poder replicarlo más adelante, sobre los costes asociados (computacional, temporal, económico, etc.) o sobre el autor de los datos.

De forma adicional, durante el proceso de creación de los datos, un sensor u otro mecanismo de medición puede proporcionar información relevante. Es posible medir algunas magnitudes como elevación, posición o temperatura e incorporar estos valores a los metadatos de forma automática.

Una vez alcanzado este punto, se dispondrá ya de una base de información, y es en base a ella que aplicando el resto de métodos deductivos se podrá ampliar.

Un primer paso para deducir nuevos metadatos es crear nuevos elementos a partir de una correspondencia directa con otro elemento de metadatos ya existente. Por ejemplo, dado un recurso con cierto formato podemos afirmar fácilmente el tipo de datos que contiene, o podemos obtener el nombre del lugar que cubre el recurso consultando un servicio de nomenclatura en base a su caja envolvente.

El siguiente paso para deducir nuevos metadatos es el cálculo de un elemento de metadatos empleando los propios datos y/o metadatos. En este sentido hay muchas líneas de investigación abiertas que abarcan un amplio abanico de posibilidades. Se pueden encontrar desde diferentes técnicas para realizar un análisis/procesado del contenido textual del recurso para averiguar su tema principal, a otras técnicas que emplean los propios datos para, por ejemplo, determinar la provincia de un pueblo por cálculos topológicos.

El último paso para deducir nuevos metadatos es la inferencia de metadatos a partir de otros metadatos o de los datos. Se basa en deducir nueva información en base a la aplicación de reglas lógicas y otras técnicas de minería de datos sobre la base de información ya disponible.

Finalmente, nunca se debe olvidar el ofrecer al usuario la posibilidad de introducir o modificar la información, aunque la idea es que éste gane confianza en la metodología en base a la observación de resultados aceptables y acabe por no participar en el proceso de generación de metadatos.

Esta metodología, permitirá mejorar progresivamente la generación automática de metadatos y la calidad resultante de estos tal y como se vayan aplicando y mejorando los diferentes métodos que la componen. Además, la metodología propuesta tiene en cuenta e intenta recopilar información que actualmente pasa desapercibida y no por ello deja de ser importante, como es la que proviene del proceso de creación. En consecuencia, el resultado de aplicar esta metodología será obtener más metadatos, de mayor calidad y corrección, más completos y con reducida participación por parte del usuario. De este modo se está atacando directamente el mayor

problema de la generación de metadatos: lo tediosa y costosa que resulta esta tarea actualmente.

2.2.2 Generación de Metadatos en gvSIG

En esta sección se describe un caso de estudio en el que se ha implementado una primera aproximación de la metodología de generación de metadatos propuesta como prueba de concepto. En este sentido, se ha desarrollado un prototipo de gestor de metadatos, usando la funcionalidad y las posibilidades de extensión que ofrece el proyecto *gvSIG*[74].

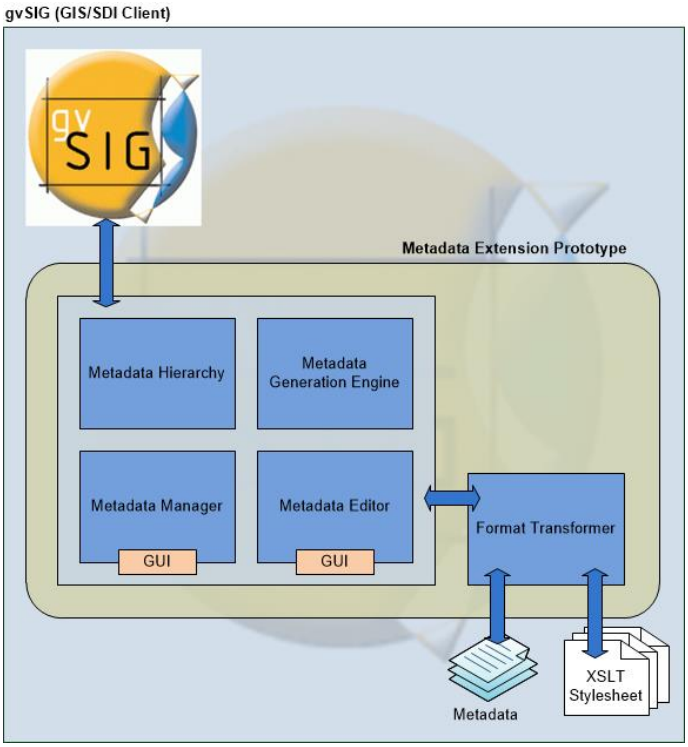

Figura 10: Arquitectura del prototipo de gvSIG

Se extendió la funcionalidad de *gvSIG* para permitir la creación y la gestión de metadatos de forma integrada y fácil. El prototipo interactúa con el núcleo de *gvSIG* para manejar los metadatos asociados a todos los recursos susceptibles de ser descritos mediante metadatos.

[74] http://www.gvsig.org

Además, proporciona funcionalidad para la extracción automática de metadatos explícitos de los recursos y su contenido. Dado que *gvSIG* es una herramienta que nos permite crear diferentes tipos de recursos que contienen IG, con esta solución integrada podemos obtener gran cantidad de información disponible en el proceso de creación de datos. El administrador de metadatos trabaja en segundo plano recopilando todos los metadatos mientras que los usuarios están trabajando con sus datos geoespaciales, cuando es requerido se aplica la metodología de generación de metadatos propuesta para obtener tanta información de los recursos como es posible sin interacción del usuario. Como valor añadido *gvSIG* utiliza estos metadatos internamente con propósitos de mejora de eficiencia: evitar la duplicación de tareas o nuevos cálculos y visualizar los recursos adecuadamente. La Figura 10 que representa la arquitectura general del prototipo de generación de metadatos en *gvSIG*.

Al incorporar a *gvSIG* el soporte de gestión de metadatos asociados a cualquiera de los datos que maneja la aplicación cualquier componente marcado como tal, podrá ser interrogado para obtener sus metadatos. A nivel de modelado, un metadato será una entidad con una serie de atributos y tipos definibles de forma dinámica, con una jerarquía de datos similar a la que define el modelo de objetos de *gvSIG*. Así, la estructura de un objeto se podrá definir de forma dinámica, y consultar los tipos de datos de sus elementos (*Metadata Hierarchy*).

El prototipo almacenará los metadatos en un archivo con formato XML junto a los datos para futuros usos. El encargado de todos los aspectos relacionados con la persistencia y la invocación y sincronización de tareas será el módulo *Metadata Manager*.

Cuando un recurso es creado, y por lo tanto aún no tiene metadatos asociados, se generarán todos los metadatos posibles aplicando la metodología propuesta a través del módulo *Metadata Generation Engine*. En este prototipo se extraen de forma automática los llamados metadatos explícitos del recurso (formato, resolución, sistema de referencia, fecha de creación, etc.) usando la información del sistema operativo y los controladores de *gvSIG* que son capaces de leer las cabeceras de los archivos de diferentes formatos para recolectar información. Como trabajo futuro se incluirán las técnicas que permiten la inferencia de nuevos metadatos de acuerdo a la

metodología propuesta de forma que el usuario encuentre un registro de metadatos completo.

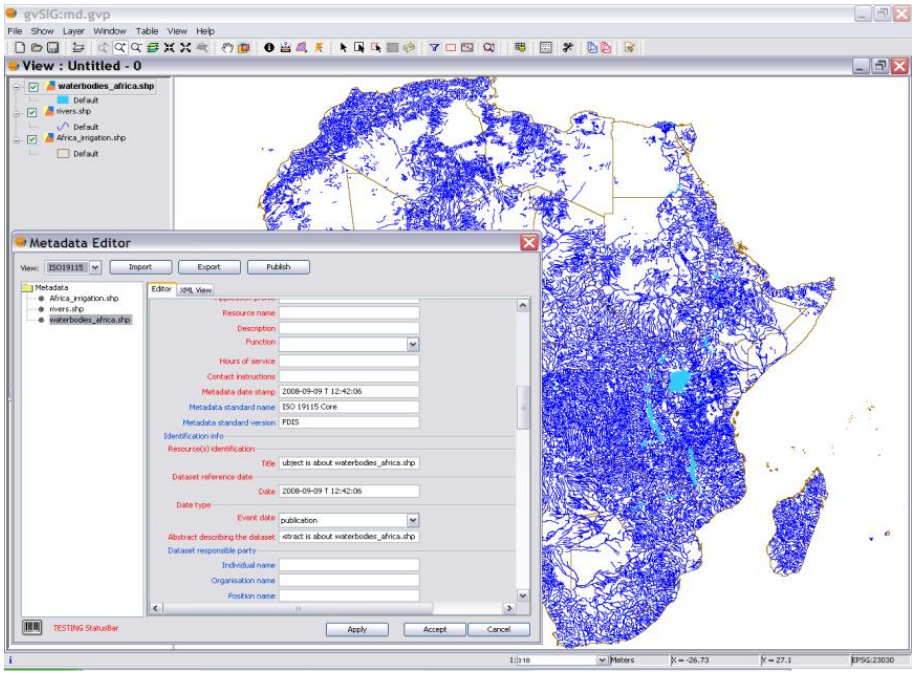

Figura 11: Editor de metadatos del prototipo de gvSIG

Finalmente, el módulo *Metadata Editor* proporciona los componentes necesarios para la implementación de un editor de metadatos que permite la visualización, edición y validación de los metadatos en base a diferentes estándares. Las transformaciones de los metadatos entre diferentes estándares y su validación se realizan a través del módulo *Format Transformer*. La Figura 11 muestra una captura de pantalla donde se puede ver parte de la interfaz gráfica de usuario del editor integrado en *gvSIG*.

Este prototipo está disponible desde octubre de 2008 como una extensión piloto de *gvSIG*. En resumen, el prototipo incluye un gestor de metadatos capaz de realizar la extracción automática de metadatos explícitos de los recursos para uso interno o para su exportación y proporciona un editor que nos permite visualizar, modificar y validar los metadatos en base a un estándar.

La funcionalidad de este prototipo es limitada dado que sólo es capaz de trabajar con el formato de archivo para datos vectoriales

shapefile[75] y el formato de metadatos estándar implementado y soportado es el núcleo del estándar ISO19115:2003. Sin embargo, su arquitectura ha sido diseñada para soportar toda la funcionalidad deseada. Así que, de alguna manera, se trata de una prueba de concepto de la funcionalidad completa.

2.2.3 Plataforma Común para la Generación de Metadatos

Actualmente, resulta muy complicado conseguir un sistema que permita la descripción de recursos totalmente autónomo, pues siempre será necesaria la participación del usuario para introducir o por lo menos validar los campos de metadatos menos intuitivos, hay que tener en cuenta que no todos los datos son fáciles de averiguar, por ejemplo el resumen o el título. En consecuencia, deberemos empezar por rellenar los campos básicos de descubrimiento de forma que se puedan ejecutar búsquedas mínimas con éxito, por ejemplo, en un catálogo. Más tarde podremos dedicar esfuerzos a completar rigurosamente el metadato. Es preferible tener todos los metadatos incompletos que "atascarse" intentando rellenar exhaustivamente uno de ellos.

En esta sección se describe otro caso de estudio en el que se ha implementado una aproximación de la metodología de generación de metadatos propuesta. En este sentido, el objetivo de este trabajo es desarrollar una plataforma común para generar la mayoría de las descripciones de metadatos de forma automática y con soporte para multitud de formatos. De modo que se puedan describir de forma completa y veraz los recursos para posteriormente poder ser publicados y conseguir así facilitar a los usuarios el acceso a los mismos.

Cuando un usuario se encuentra frente a un recurso desconocido en términos de formato, lo primero que tiene que hacer es examinar y extraer la mayor cantidad de información posible del recurso en sí mismo, de su contexto y, obviamente, de su contenido. Pero esto no siempre es fácil, dado que la extracción automática de metadatos implica el conocimiento de las estructuras internas de los formatos de almacenamiento de datos utilizados por los recursos geográficos

[75] http://en.wikipedia.org/wiki/Shapefile

[147]. Este proceso normalmente lleva a cabo una correspondencia entre las características extraídas de cada formato y los distintos elementos de metadatos descritos por alguno o algunos de los estándares existentes (Dublin Core, ISO19115, etc.). Sin embargo, el gran número de formatos de datos existentes para los recursos geográficos hace muy difícil que una sola aplicación pueda manejar todos ellos. Un enfoque alternativo es el desarrollo de soluciones integradas y flexibles basadas en la reutilización de librerías, herramientas o componentes que son capaces de leer múltiples formatos para extraer la información de metadatos.

Si aparte de describir los recursos geográficos, se pretende generalizar el proceso de extracción de metadatos para cualquier tipo de recurso multimedia, entonces el problema se agrava, dado que el número de posibles formatos con los que se va a tener que tratar aumenta considerablemente. Por esa razón, se busca una plataforma que permita acceder a los recursos heterogéneos y obtener información de una manera homogénea.

Discusión sobre herramientas y plataformas de acceso a datos y metadatos

La necesidad de acceder y de obtener información de tantos formatos como sea posible motivó un estudio que analiza y evalúa varias plataformas comunes que proporcionan acceso a información geográfica, así como varias soluciones de código abierto para la extracción de metadatos.

El objetivo es obtener descripciones de recursos basadas en la extracción de metadatos, por lo que las herramientas de extracción de metadatos deben ser consideradas en primer lugar. Analizando las capacidades de generación automática de metadatos de *CatMDEdit*, se puede ver que proporciona extracción de metadatos para varios de los formatos geográficos soportados [97]. Sin embargo, como se pretende generalizar el proceso de extracción de metadatos para cualquier tipo de recurso multimedia, se consideró la herramienta *Apache Tika* como una solución que se adecua a los objetivos marcados. Actualmente, *Apache Tika* no incluye soporte para formatos geoespaciales, sin embargo tiene una arquitectura extensible que permite añadir nuevos formatos de datos. Como la extensibilidad es un requisito fundamental en este enfoque para que se pueda dar

soporte a tantos tipos y formatos de recursos como sea posible, *Apache Tika* fue la herramienta de extracción de metadatos seleccionada.

Como paso previo a la extracción de los metadatos, la nueva solución tenía que ser capaz de acceder e interpretar los formatos de datos. A continuación se discute sobre las plataformas de acceso a datos geográficos analizadas, que son, junto con *Apache Tika*, el otro componente de la solución integrada. La primera plataforma que se analizó fue *GeoTools*[76] [214], las primeras pruebas de extracción de metadatos atrajeron inicialmente la atención como una buena solución, sin embargo, posteriormente resultó complicado ampliar la gama de formatos de recursos soportados como otras plataformas permiten, y que además ya dan soporte a más formatos. Por su parte, la capa de acceso a datos (DAL) de *gvSIG* [6] tiene como objetivo proporcionar a *gvSIG* una capa de abstracción que permite al núcleo de la aplicación operar de forma homogénea con diferentes fuentes de datos y formatos. Aunque DAL es conceptualmente compatible con la plataforma que se estaba buscando, todavía se encontraba en las primeras etapas de desarrollo. Las librerías *GDAL/OGR*[77] [219] proporcionan acceso a una gran cantidad de formatos geográficos ráster y vectoriales, con un bajo nivel de abstracción. Por lo tanto, serían un buen punto de partida si se hubiese deseado empezar a desarrollar una plataforma común para acceder a datos geoespaciales. Pero mediante una solución con un nivel de abstracción mayor se puede facilitar el trabajo y ahorrar una gran cantidad de tiempo y esfuerzo. Por último, el proyecto *OSGeo FDO*[78] posibilita el acceso a diversas fuentes de datos geoespaciales a través de un mecanismo común. Soporta una gran variedad de fuentes de datos, incluyendo formatos de archivos, bases de datos y servicios geoespaciales. Se consideró que *OSGeo FDO* puede ofrecer la funcionalidad deseada y es compatible con casi todos los formatos de información geográfica conocidos, por lo que fue la plataforma de acceso a datos seleccionada.

[76] http://www.geotools.org
[77] http://www.gdal.org
[78] http://fdo.osgeo.org

Integración de Apache Tika y OSGeo FDO

La aproximación que se considera aquí es la combinación de los proyectos *Apache Tika* y *OSGeo FDO* en una solución integrada que ofrece los beneficios de ambos proyectos. Mediante esta integración, se obtiene una poderosa herramienta de extracción de metadatos para gran variedad de tipos de recursos multimedia, con especial énfasis en recursos geoespaciales.

El prototipo funcional [19] permite la extracción de metadatos de una amplia gama de formatos de recursos multimedia. En concreto, esta herramienta soporta los tipos de recursos inicialmente soportados por *OSGeo FDO* (más de 150 formatos de IG)[79] y los tipos de recursos que soporta *Apache Tika* (más de 50 formatos multimedia)[80]. Por lo que la solución integrada es compatible con más de 200 formatos de recursos multimedia en su conjunto. El nivel de detalle de las descripciones que esta herramienta proporciona para cada tipo de recurso depende del tipo del propio recurso.

Etiqueta	Significado
Resource	Inicia la descripción de un recurso
FormatName	Formato en el que se encuentra el recurso analizado
ResourceType	Tipo de recurso
Provider	Proveedor de FDO utilizado para analizar el recurso
Source	Origen de los datos
ResourceName	Nombre del recurso
ConnectionString	Cadena de conexión al recurso
dateStamp	Fecha de la creación de los metadatos
keywords	Palabras clave en la descripción del recurso
SpatialContexts	Descripción de los contextos espaciales del recurso. Incluyendo nombre, sistema de coordenadas, extensión…
Schemas	Descripción de los esquemas de datos del recurso. Incluyendo nombre, atributos, *features*…
SchemaAttributes	Descripción de los atributos del esquema. Incluyendo nombre y valor de los mismos.
FeatureClasses	Descripción de los *features* del recurso. Incluyendo nombre, restricciones, sus propiedades…
BaseIdentityProperties y Properties	Descripción de las propiedades de cada *feature*. Incluyendo nombre, tipo y diferente información dependiente del tipo, como tamaño, valor por defecto, precisión...

Figura 12: Resumen de los metadatos extraídos gracias a OSGeo FDO

La nueva implementación basada en *OSGeo FDO* permite analizar los recursos geoespaciales independientemente del lugar donde se

[79] http://fdo.osgeo.org/OSProviderOverviews.html
[80] http://tika.apache.org/1.2/formats.html

almacenan y extraer descripciones uniformes de metadatos, independientemente del formato de datos del recurso. La Figura 12 recoge un conjunto seleccionado de los descriptores de metadatos que se extraen. Estas etiquetas son descriptores comunes e independientes del formato de los recursos. Sin embargo, las características específicas de cada formato pueden ser también capturadas.

Además de estas etiquetas, en el prototipo se han añadido a las descripciones de los servicios basados en estándares OGC toda la información procedente de sus respectivos *GetCapabilities*[81]. Cabe destacar que esta puede ser la principal fuente de información para este tipo de recursos, siempre y cuando sus responsables hayan dedicado cierto esfuerzo en completar sus metadatos. Por otra parte, en la solución también se ha incluido un módulo de configuración para especificar a priori algunas etiquetas de metadatos en un archivo XML que se incluirán automáticamente en todas las descripciones de metadatos. Por lo tanto, este módulo permite incluir información dependiente del contexto en cada descripción de metadatos de forma fácil y configurable. Los metadatos basados en el contexto pueden incluir información tan relevante como el autor del conjunto de datos o la empresa a la que pertenecen, entre otros.

La aplicación en este caso de estudio de la metodología de generación de metadatos propuesta (ver Sección 2.2.1), ha permitido desarrollar una plataforma que permite obtener información y acceder de forma homogénea a recursos heterogéneos en base a la integración de los proyectos *OSGeo FDO* y *Apache Tika*, las descripciones conseguidas para los recursos de IG a partir de los metadatos generados de forma automática son bastante completas. Si a esto se le suma la participación del usuario a la hora de incluir más metadatos, ya sean como metadatos relativos al contexto preconfigurados o rellenando a mano los metadatos menos intuitivos, se puede conseguir un sistema que facilite en gran medida las rutinarias y poco motivadoras labores de los creadores de metadatos, reduciendo además los errores que se producen al escribir directamente los metadatos.

[81] Todos los servicios basados en estándares OGC implementan la operación *GetCapabilities* que proporciona metadatos sobre el servicio, sus operaciones y parámetros soportados y sobre su contenido.

2.3 MIMEXT: Anotación de Recursos Heterogéneos

La anotación de recursos se refiere al hecho de incorporar o añadir metadatos a un recurso. En la Sección 2.1.1 de este capítulo hemos visto que cualquier recurso es susceptible de ser georreferenciado y que pueden ser descritos mediante el uso de metadatos. Para que un recurso sea georreferenciado, este debe contener o estar acompañado por, al menos, su localización espacial en base a un sistema de coordenadas determinado. Además, es deseable que los recursos estén acompañados por otros metadatos que los describen y permiten su correcta integración cualquier sistema de información. En esta sección, tras explorar diferentes opciones para incorporar metadatos a los recursos en la Sección 2.1.4, se propone una solución.

En este contexto, la solución que se propone pretende conseguir dos objetivos. Por una parte, que permita la georreferenciación de un gran número de tipos de recurso. Esto rompería con la actual tendencia de ofrecer soluciones específicas para un determinado tipo de recurso o escenario de aplicación. Por otra parte, que resulte un método para compartir de forma sencilla datos, metadatos y otros recursos relacionados encapsulándolos en un único archivo. Esta solución podría usarse por ejemplo para enriquecer las IDEs incorporando nuevos tipos de datos de otros dominios, ya documentados y georreferenciados y que no se han tenido en cuenta hasta la fecha (p. ej. hojas de cálculo, video o modelos 3D).

Solución propuesta: MIMEXT

Tras explorar diferentes alternativas actuales para georreferenciar los recursos, es decir para incorporar los metadatos a los recursos, la contribución que se presenta en esta sección pretende ser una solución lo más general posible, es decir, permite la georreferenciación de un gran número de tipos de recurso. Por otra parte, se pretende que resulte un método sencillo para compartir datos, metadatos y otros recursos relacionados encapsulándolos en un único archivo.

Debido a su simplicidad, sus capacidades de visualización y anotación y que la mayoría de herramientas geoespaciales y servicios

web los soportan, KML se ha convertido en un estándar popular y ampliamente utilizado.

Teniendo en cuenta todos los beneficios ofrecidos por KML el uso de un fichero codificado en este estándar facilitaría la anotación de los recursos incluyendo detalles sobre su visualización. Además, mediante el uso de KML se extiende el posible uso de la solución propuesta en todas aquellas aplicaciones que reconocen este formato. Por lo tanto, se propone el uso de archivos KML como elemento central para la georreferenciación de los recursos.

KML para la georreferenciación de recursos

KML puede representar una solución para la georreferenciación de recursos. Actualmente, este lenguaje soporta una serie de primitivas geométricas como son el punto, líneas, polígonos o incluso modelos COLLADA[82] heredados de GML. Estas primitivas se pueden asignar a elementos del lenguaje que contienen cierta información como pueden ser los *Placemark*. Dentro de estos elementos existen campos como *Description* que permiten la inclusión de código XHTML[83] que en muchas ocasiones se utiliza para embeber recursos como por ejemplo imágenes o vídeos. Así, embebiendo estos recursos en ciertas etiquetas se obtiene un método que, aunque indirecto, resulta eficiente para la georreferenciación de ciertos recursos.

Inclusión de metadatos en KML

KML ofrece la posibilidad de añadir metadatos mediante el uso de distintos elementos del lenguaje. El método más razonable consiste en el uso de la etiqueta *ExtendedData* que permite la inserción de código XML dentro de un archivo KML. Dependiendo del uso de otras etiquetas dentro de *ExtendedData*, existen tres formas de añadir dicho código XML que en nuestro caso puede representar los metadatos de un determinado recurso. Es posible añadir simples pares de tipo clave-valor, es posible definir un pseudo-esquema en el propio documento KML para especificar la estructura que va a seguir el código XML añadido y, por último, es posible importar esquemas complejos definidos de forma externa al archivo KML. Este último método es sin duda el más interesante pues permite por ejemplo

[82] http://collada.org
[83] http://www.w3.org/TR/xhtml1

importar esquemas de metadatos como el ISO19115 para describir los recursos definidos en el archivo KML.

Falta de funcionalidad en KML

Como ya se ha comentado, existe la posibilidad de georreferenciar ciertos tipos de recursos embebiéndolos dentro de etiquetas XHTML. Esta aproximación representa un uso impropio de elementos del lenguaje XHTML, que sólo puede ser aplicado a un pequeño conjunto de tipos de recursos (aquellos que pueden embeberse en código XHTML). De esta forma nos encontramos con que KML promete ser una posible solución para la georreferenciación de recursos de distinta índole, sin embargo el lenguaje no es lo suficientemente rico como para ser aplicado de forma más genérica.

La extensión KML MIMEXT

Se pretende georreferenciar una amplia gama de tipos de recursos, que no pueden ser anotados de forma efectiva dentro de un archivo KML para su procesamiento. Una situación similar tuvo lugar con HTML y los navegadores web, donde se requiere la instalación de diferentes *plug-ins* o aplicaciones para visualizar algunos tipos de recursos directamente en el navegador web.

El lanzamiento de la especificación HTML5[84] trata de evitar todas estas complicaciones presentes en HTML, introduciendo nuevas etiquetas multimedia para cubrir de forma nativa una gran cantidad de contenidos (video, audio, canvas, etc.). Esta solución toma un enfoque similar al extender KML para facilitar la integración de los recursos georreferenciados que KML no soporta todavía. La extensión KML propuesta, llamada MIMEXT (*MIME Extension*), mejora KML en dos aspectos: ofrece un mecanismo de georreferenciación mejor y da soporte a una gran variedad de tipos de recursos a través de la anotación.

Extendiendo el estándar KML

La extensión y la restricción del esquema del estándar KML para los propósitos específicos es posible a través de los llamados Perfiles de Aplicación (*Application Profiles*) [178]. Un perfil de aplicación debe

[84] http://www.w3.org/TR/html5

ceñirse a ciertas restricciones que aseguran que las nuevas etiquetas están correctamente derivadas de los elementos base de KML. La extensión de KML MIMEXT propuesta es u perfil de aplicación construido por herencia, es decir, los nuevos elementos y las etiquetas se derivan de los tipos básicos abstractos del núcleo de KML.

MIMEXT

La extensión MIMEXT consiste en añadir nuevos elementos a KML para georreferenciar los recursos y para anotar su tipo MIME. En cuanto al primer grupo, los elementos para la georreferenciación, derivan directamente de la etiqueta de KML <*Geometry*>. Esto permite asociar esta geometría de forma efectiva con todos los recursos que actualmente no están soportados por el estándar KML.

En cuanto al segundo grupo de elementos, añadidos para anotar los tipos MIME, deben registrar la información sobre el tipo MIME de un determinado recurso y su extensión acorde con el formato del archivo que lo contiene. Esta información puede ser útil, no sólo para el usuario final, sino también para que las aplicaciones cliente sean capaces de ofrecer la visualización directa de estos recursos.

La descripción sobre el tipo de recurso debe seguir las especificaciones MIME [27] [80] [81]. Que define una extensa lista de tipos MIME[85], especificando en cada caso el tipo de archivo, el subtipo y la extensión. Las descripciones sobre el tipo de recurso basadas en MIME han sido tomadas como base para nuestra aproximación con el fin de proporcionar información sobre el recurso de una manera lo más estandarizada posible.

La estructura de MIMEXT

Los archivos XML tienen una estructura bien definida, particularmente en lo referido a las estructuras anidadas y la sintaxis de los elementos, atributos y tipos. La Figura 13 muestra todos los elementos que componen esta nueva extensión y la jerarquía de los elementos.

[85] http://www.iana.org/assignments/media-types

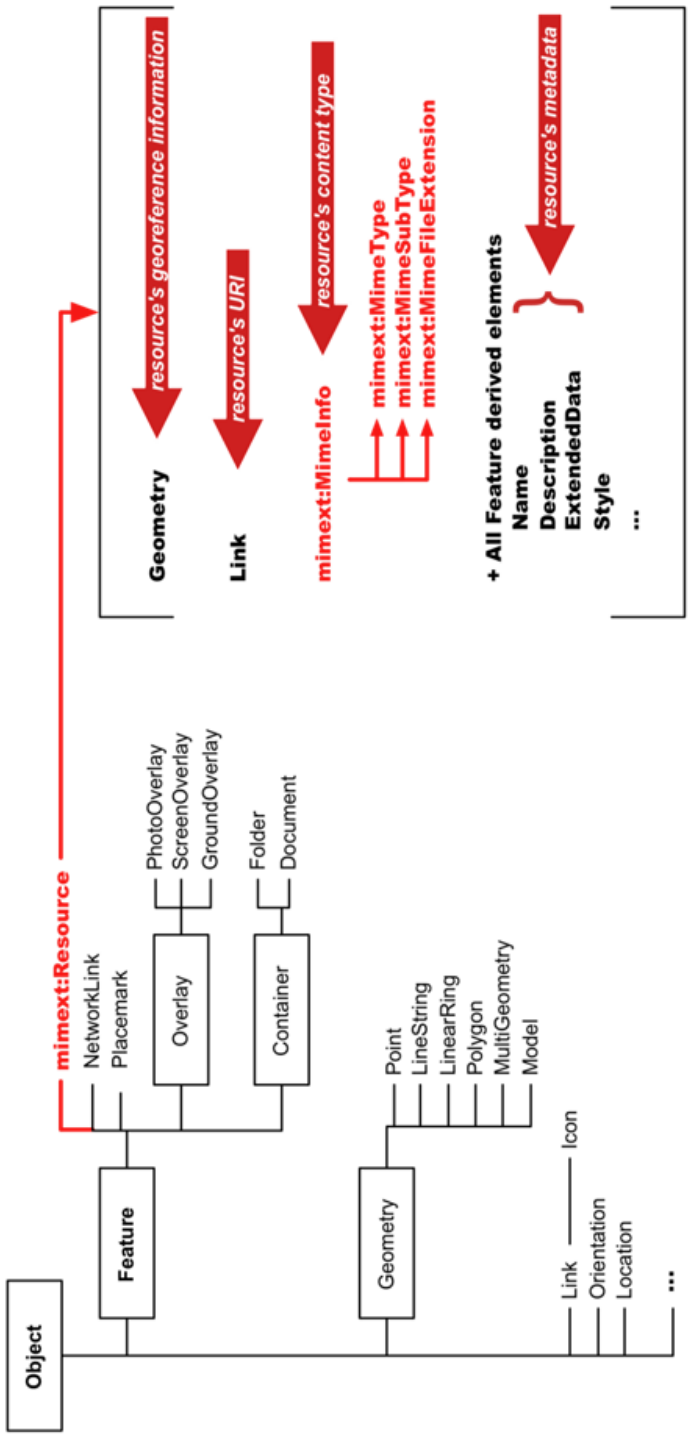

Figura 13: Jerarquía y relaciones entre los elementos de KML (negro) y los elementos de MIMEXT (rojo)

El elemento de la extensión MIMEXT <*Resource*> deriva directamente del elemento abstracto de KML <*Feature*> y, por herencia, contiene los mismos elementos que tiene <*Feature*>. Aparte de los elementos heredados de KML, el elemento <*Resource*> de MIMEXT incluye tres etiquetas específicas (Figura 13, derecha).

La etiqueta de KML <*AbstractGeometryGroup*> (denominada *Geometry* en la Figura 13) permite la asociación de cualquier tipo de geometría definida en KML al recurso que se está describiendo (<*Resource*>). Estas geometrías incluyen <*Point*>, <*LineString*>, <*LinearRing*>, <*Polygon*>, <*Multigeometry*> y <*Modelo*>. En lugar de georreferenciar un recurso únicamente como un punto, al permitir la asociación con geometrías más complejas, podemos añadir información adicional acerca de los recursos. Por ejemplo, se podría asociar a una pista de audio que describe una carrera una geometría lineal representando el recorrido de dicha carrera o asociar un polígono con un documento PDF que representa información catastral.

Por su parte, la etiqueta <*Link*> de KML permite especificar la ubicación (URI[86]) para un recurso determinado. Actualmente soportamos las referencias apuntando tanto a recursos locales como remotos. La etiqueta <*Link*> resulta esencial cuando se carga el contenido de servicios geoespaciales remotos como los *Web Mapping Services*[87] (WMS) o imágenes de otros servicios sociales como *Flickr*.

A diferencia de los dos anteriores, la etiqueta <*MimeInfo*> de MIMEXT, es un elemento totalmente nuevo creado específicamente. Su finalidad es aportar información acerca del tipo del recurso con el que se está trabajando. De modo que facilita la anotación del tipo MIME para cualquier recurso, indicando su tipo MIME, su subtipo y la extensión del archivo usando las etiquetas de MIMEXT <*MimeType*>, <*MimeSubType*> y <*MimeFileExtension*> respectivamente. La etiqueta <*MimeType*>, referente al tipo del recurso, puede contener valores como *audio*, *image*, *video* o *text*. En el caso del subtipo (<*MimeSubType*>) se podrían usar valores como *pdf*, *msword*, *x-latex* o *mpeg*. Finalmente en el caso de la extensión (<*MimeFileExtension*>) se usan valores como *gif*, *avi*, *odt* o *mp3*. Esta información pretende

[86] Un *Uniform Resource Identifier* (URI) es una cadena de caracteres corta que identifica inequívocamente un recurso, normalmente accesible en una red o sistema.
[87] http://www.opengeospatial.org/standards/wms

ser procesada y utilizada por aplicaciones como globos virtuales para visualizar y explotar de forma conveniente los recursos.

Por último, podemos mejorar las descripciones de los recursos mediante la incorporación de más metadatos en otros elementos KML derivados de <*Feature*>. Por ejemplo, dentro de la etiqueta de KML <*Description*> se podrían incorporar nuevas etiquetas con elementos de los metadatos basados en las características del recurso, o incluso se podría incorporar un registro completo de metadatos de acuerdo a un estándar como ISO19115 dentro de la etiqueta <*ExtendedData*> de KML.

La Figura 14 muestra un fragmento de MIMEXT como ejemplo de una descripción de un recurso de audio. En el podemos apreciar todas las etiquetas que hemos descrito anteriormente, destacando en rojo los elementos de la extensión MIMEXT.

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2"
  xmlns:mimext="http://www.geoinfo.uji.es/kml/ext/2.2">
<mimext:Resource>
<name>Audio Resource</name>
  <description>This is a test for audio resources</description>
  <Point>
    <coordinates>26.8694,37.2545,0</coordinates>
  </Point>
  <Link>
    <href>resources/audio_file.wav</href>
  </Link>
  <mimext:MimeInfo>
    <mimext:MimeType>audio</mimext:MimeType>
    <mimext:MimeSubType>wav</mimext:MiMeSubType>
    <mimext:MimeFileExtension>wav</mimext:MimeFileExtension>
  </mimext:MimeInfo>
</mimext:Resource>
</kml>
```

Figura 14: Ejemplo de descripción MIMEXT

Encapsulación

La anotación externa como método para la georreferenciación ofrece algunas ventajas. Sin embargo, el disponer de los datos y sus metadatos como distintas unidades físicas o lógicas puede acarrear problemas, especialmente para la correcta administración, actualización o sincronización de ambos. Es por ello que para mitigar estos efectos se utilizan técnicas como la encapsulación o agrupación

de ambos elementos en un único archivo para así componer una única unidad lógica y física.

El uso de KML facilita esta tarea introduciendo el uso de los archivos KMZ. Básicamente, un archivo KMZ es un archivo comprimido que contiene un archivo KML junto con una carpeta donde se alojan elementos referenciados en dicho documento KML, que suelen ser iconos o imágenes.

En nuestra aproximación, el uso de KMZ permite encapsular en un mismo archivo el recurso, su información de geolocalización y otros metadatos en formato KML así como otros recursos relacionados (p.e licencias de uso). La Figura 15 muestra la distribución lógica de los distintos componentes de nuestra propuesta al ser encapsulados en un único archivo KMZ.

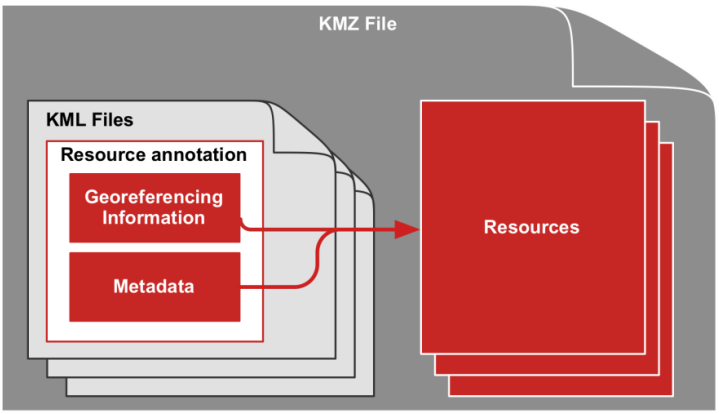

Figura 15: Encapsulación mediante KMZ

Gracias a su estructura simple, la mayoría de las aplicaciones que soportan KML también soportan KMZ haciendo de esta una solución ideal para transportar y compartir información geográfica. En este sentido, KMZ representa una solución simple pero potente para la encapsulación en un único archivo de cualquier tipo de recurso junto con su ubicación espacial y otra información descriptiva expresada en KML y, por lo tanto, en la extensión MIMEXT. Este enfoque ofrece las mismas ventajas enumeradas para KML, además de una forma sencilla de combinar el recurso con su información asociada.

2.4 *DESCaaS*: Servicios de Descripción

Los Sistemas de Información (SI) proporciona a los usuarios todas las funciones para realizar las tareas diarias, tales como el descubrimiento, acceso o descarga de los recursos que manejan. Siguiendo la tendencia de ofrecer a los usuarios funcionalidades como servicios, la contribución que se detalla en esta sección propone ampliar la funcionalidad de los SIs con un nuevo paradigma llamado *Description as a Service* (DESCaaS). Según nuestro concepto, los servicios DESCaaS pueden utilizarse para interactuar con otros servicios, permitiendo el encadenamiento de servicios, y ofrecer funcionalidades de descripción de recursos que facilitan la explotación, publicación y descubrimiento de nuevos contenidos.

Concepto

El concepto DESCaaS surge como una evolución de la combinación de los conceptos de "descripción", "servicio" y *Software as a Service*[88] (SaaS), siendo capaz de invocar componentes de software reutilizables y bien configurados a través de la red. Equivalente a *Model as a Service* (MaaS) [187], DESCaaS es un tipo de SaaS en la capa de servicios Web de los SI. Si consideramos que nuestro modelo es el proceso de generación de metadatos, DESCaaS podría ser visto como un caso de uso específico de MaaS. SaaS se define como un modelo de distribución de software donde el soporte lógico y los datos que maneja se alojan en servidores de una organización a los que los usuarios acceden bajo demanda a través de la red. El paradigma DESCaaS ha sido motivado por la importancia y el volumen de datos que, en nuestra sociedad, son accesibles pero no se han publicado correctamente y, en consecuencia, no son totalmente descubribles.

En consecuencia, el objetivo de DESCaaS es proporcionar a los usuarios descripciones homogéneas de recursos, cuyo volumen y complejidad aumenta día tras día. Para obtener las descripciones, DESCaaS ofrece acceso a procesos de generación de metadatos bajo demanda a través de la red. El propósito de este enfoque es explotar una infraestructura basada en la red para facilitar la integración de servicios de descripción desde y en cualquier SI.

[88] http://en.wikipedia.org/wiki/Software_as_a_service

La Figura 16 presenta una visión general del paradigma DESCaaS. Brevemente, los servidores DESCaaS reciben como entrada básica el propio recurso. Se pueden considerar otras entradas opcionales como metadatos adicionales o algunos parámetros de configuración. El recurso dado se analiza usando métodos de generación de metadatos de terceros proporcionados por el servicio (métodos de extracción de metadatos o traducción, modelos de minería de datos, etc.) para construir su descripción de acuerdo con el formato de salida solicitado.

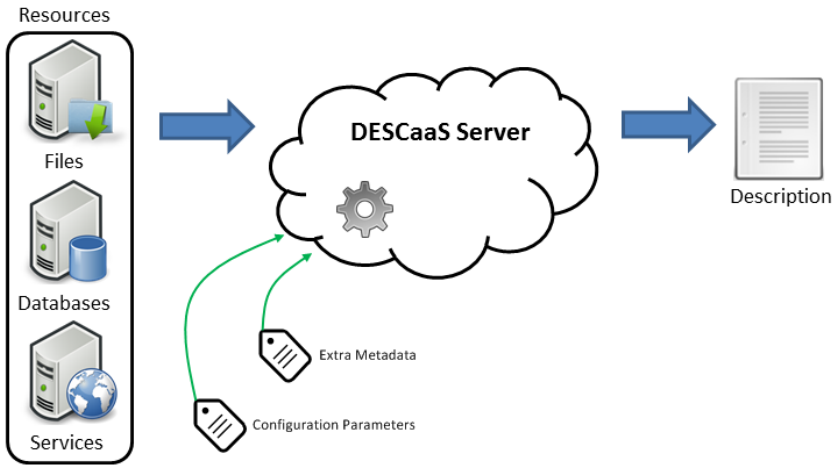

Figura 16: Visión general del paradigma DESCaaS

Especificación

Se define la entrada básica de DESCaaS simplemente como un "recurso" porque tratamos de definir el paradigma lo más genérico y abstracto como sea posible. Sólo un parámetro de entrada se especifica como obligatorio, una URL que apunte al recurso. No importa si es un archivo local o remoto, los datos obtenidos de una base de datos o como respuesta de otro servicio, sólo es necesario que sea accesible. Las restricciones sobre los tipos de recursos soportados provendrán de la implementación concreta, de acuerdo con las capacidades del proceso de descripción, por ejemplo, un método de generación de metadatos determinado basado en la extracción puede que solo sea capaz de analizar un tipo concreto de recursos. Además, algunos servicios de descripción pueden necesitar

algunos parámetros de configuración (*Configuration Parameters*) para configurar correctamente sus métodos de descripción. A través de la configuración del servicio, DESCaaS puede ser capaz de ofrecer diferentes niveles de descripción. Por otra parte, algunos casos de uso requieren que se proporcionen metadatos adicionales (*Extra Metadata*) para obtener descripciones personalizadas, por ejemplo información contextual.

El formato de salida de DESCaaS no se especifica para permanecer lo más genérico posible. Al final le corresponde a los desarrolladores decidir que formatos ofrecerá el servicio, es decir, metadatos sin procesar, XML o diferentes estándares.

El número de procesos de descripción que proporciona una sola implementación de DESCaaS tampoco está limitado. Una implementación DESCaaS debería ofrecer su funcionalidad de descripción a través de una interfaz estandarizada, es decir, se permiten múltiples procesos de descripción gestionando diferentes tipos de datos y que resultan en diferentes formatos de descripción.

Operaciones

De acuerdo con esta especificación conceptual del paradigma sugerimos tres operaciones obligatorias que deben ser implementadas por un servicio DESCaaS[89]. La operación *GetCapabilities* debe devolver un documento de metadatos sobre el servicio describiendo brevemente las capacidades de la implementación específica del servicio. Una explicación detallada sobre un proceso de descripción concreto, incluyendo sus entradas requeridas y opcionales (incluyendo los formatos de recursos admitidos) y los parámetros de salida (incluyendo los formatos de salida ofrecidos) son proporcionados por la operación *DescribeProcess*. Por último, la operación *Execute* debe ejecutar el proceso de descripción especificado y devolver la descripción obtenida en el formato requerido.

Beneficios

DESCaaS permite la publicación y la reutilización multilingüe y a alto nivel de los procesos de generación de metadatos. Además, al proporcionar descripciones de acuerdo a formatos estándar,

[89] Los nombres de las operaciones se han inspirado en las especificaciones de OGC.

DESCaaS promueve la interoperabilidad y mejora el descubrimiento de los recursos.

Una gran cantidad de trabajos científicos están dedicados a mejorar el descubrimiento de recursos en línea (para el intercambio de datos y su reutilización) mediante el uso de bibliotecas digitales o de IDEs en el contexto de la IG. En ambos casos, los sistemas se basan en la indexación de los recursos en base a sus descripciones homogéneas. DESCaaS apoya y facilita el proceso de indexación en dichas bibliotecas digitales. Además, DESCaaS ayuda a los productores de contenido durante la publicación, proporcionando herramientas para obtener descripciones de recursos homogéneas y basadas en estándares. Asimismo, DESCaaS proporciona un mecanismo para ayudar a los consumidores de datos a descubrir y explotar nuevos recursos mediante la obtención de sus descripciones. Para decirlo en pocas palabras, el paradigma DESCaaS permite la publicación y el descubrimiento de los recursos de una forma más eficiente.

Todos estos beneficios potencian la formación de comunidades de usuarios, facilitando el proceso de publicación de su contenido en una infraestructura de recursos compartidos. Al facilitar el proceso de publicación instamos a los usuarios a compartir sus datos. DESCaaS es especialmente interesante para los enfoques con una perspectiva de abajo hacia arriba (*bottom-up*[90]). De esta manera estamos promocionando los sistemas y comunidades basados en *Public Participation GIS* (PPGIS)[91], *Volunteered Geographic Information*[92] (VGI) y contenido generado por usuarios[93] en general. Por lo tanto, la cantidad de datos disponibles se va a incrementar y se estimula la formación de comunidades con un contexto de intereses comunes.

Como se ve, DESCaaS promete una gran mejora en la obtención de descripciones de recursos al permitir la publicación de los procesos de generación de metadatos. Sin embargo, todavía tienen que

[90] En el diseño *bottom-up* las partes individuales se enlazan para formar componentes más grandes, que a su vez se enlazan hasta que se forma el sistema completo. Esta aproximación también es aplicable a la forma de recopilar los datos, que en este caso provendrían de los usuarios para componer el sistema.
[91] http://en.wikipedia.org/wiki/Public_participation_GIS
[92] http://en.wikipedia.org/wiki/Volunteered_geographic_information
[93] http://en.wikipedia.org/wiki/User-generated_content

resolverse los principales problemas de la generación de metadatos. Como se refleja en secciones anteriores, el principal problema se refiere a la falta de información sobre algunos recursos para generar una descripción adecuada. La validez y la completitud de las descripciones siempre dependen de la cantidad de información que dispone el recurso y de la implementación específica del proceso de descripción. Por otra parte, otro problema es la necesidad de tratar con gran variedad de formatos para poder extraer metadatos y dar soporte a los diferentes estándares de metadatos para ser ofrecidos como salida.

Casos de uso

Debido a su definición general y abstracta, el paradigma DESCaaS pueden encajar con muchos casos de uso. El caso de uso más común es ofrecer una descripción de un recurso que no tiene. Esto supone una gran ayuda para el usuario para descubrir, explotar o publicar el recurso. Resultando especialmente interesante para los archivos binarios, dado que DESCaaS puede proporcionar una descripción textual de este tipo de recursos que permita su indexación en función de sus propiedades y su posterior descubrimiento. Otro caso de uso de este paradigma es la transformación de las descripciones entre diferentes formatos o estándares. Por otra parte, la comunidad que trabaja en diferentes aspectos de la Semántica también puede beneficiarse del enfoque que ofrece DESCaaS. Vincular representaciones de conocimiento existentes con los recursos que describen es una de las cuestiones abiertas en el campo de la semántica. Donde, sobre todo los recursos no estructurados como imágenes plantean problemas. En este sentido, DESCaaS ofrece descripciones semiestructuradas que pueden ser fácilmente anotadas semánticamente y, por tanto, supone una solución útil para añadir descripciones semánticas a los recursos [151]. Finalmente, el paradigma DESCaaS se puede aplicar a muchos casos de uso más específicos en los que las descripciones de recursos son necesarias con el fin de mejorar el descubrimiento, explotación, publicación e interoperabilidad de los recursos.

Generando descripciones de recursos NetCDF (NetCDF2NcML)

Como se ha comentado en el capítulo de introducción, en el contexto de iniciativas como GEOSS o GMES emergen retos para

abordar la recogida, análisis, y la distribución de los recursos para facilitar la generación y la comunicación de información sobre el medio ambiente. El proyecto ENVISION [150] está clasificado dentro de estas iniciativas y cubre algunos de sus propósitos. El objetivo general del proyecto es una plataforma de apoyo a las decisiones medioambientales basada en la Web, que ayude a los usuarios sin conocimientos en TIC a crear y ejecutar modelos ambientales como servicios habilitados semánticamente. A continuación se presenta un caso de uso real, como ejemplo de servicios de descripción DESCaaS, que proporciona descripciones de archivos NetCDF en el contexto del proyecto ENVISION.

El proyecto ENVISION define tres demostradores, un piloto sobre deslizamientos de tierra, un piloto sobre derrames de petróleo y un piloto sobre inundaciones. Dentro del piloto sobre derrames de petróleo hay una gran demanda de información adicional acerca de los recursos utilizados. La mayoría de los datos utilizados en este piloto son series de tiempo codificados como NetCDF[94] [186] con tres o cuatro dimensiones. Las series de tiempo son secuencias de puntos de datos, medidos típicamente en momentos sucesivos espaciados a intervalos de tiempo uniformes. El compositor del escenario de una deriva aceite tiene que crear el *workflow* y calibrar los parámetros de entrada. Un paso importante es la selección de los datos actuales de viento y mar para las próximas 60 horas. Tales datos son proporcionados por el Instituto Meteorológico Noruego y codificados en el formato NetCDF. Es en este punto donde el paradigma DESCaaS entra en juego. Con el prototipo implementado, el compositor del escenario puede analizar fácilmente el contenido de los archivos NetCDF mediante las descripciones NcML[95] [163] proporcionadas por el servicio DESCaaS para comprobar su idoneidad. Una visión adicional es incluir servicios DESCaaS en las composiciones de servicios. De este modo, los servicios DESCaaS representan un servicio de datos que contiene la descripción de los recursos y ofrece punteros a los recursos originales. Por otra parte, el compositor del escenario de la deriva de aceite también es responsable de la visualización de los resultados del modelo. Las predicciones sobre la deriva del aceite y su efecto sobre la población de bacalao también están codificadas en NetCDF. Obteniendo las

[94] http://www.unidata.ucar.edu/software/netcdf
[95] http://www.unidata.ucar.edu/software/netcdf/ncml

descripciones de los resultados del modelo a través de un servicio DESCaaS se puede mejorar la visualización y la explotación de estos resultados.

3. Publicación

La publicación es la acción que consiste en revelar, manifestar o difundir una determinada información o un determinado contenido, en conocimiento general del público[96].

Figura 17: Visión general del Capítulo 3

La publicación de los recursos catalogándolos o indexándolos de acuerdo a sus características y su contexto (metadatos), permite y facilita su posterior descubrimiento [199]. Por ello, en este capítulo se presentan conceptos teóricos y técnicos relevantes para la publicación de los recursos. Con el fin de acotar nuestro alcance, nuevamente, nos centraremos en la publicación de información geoespacial o recursos georreferenciados. La Figura 17 representa la estructura del capítulo y su posición dentro del *workflow* general.

[96] http://es.wikipedia.org/wiki/Publicación

En este capítulo en primer lugar, se revisan diferentes opciones actuales para la publicación de recursos georreferenciados prestando especial atención a aquellas que nos permiten hacerlo de una forma interoperable. Seguidamente se revisan diferentes estándares existentes que indican que información deben contener los metadatos que describen los recursos para ser publicados. De esta forma se responde a las preguntas: *¿Qué es publicar?*, *¿Por qué publicar?*, *¿Cómo se pueden publicar recursos georreferenciados?*, *¿Cómo se puede hacer de forma interoperable?* y *¿Qué información debe incluirse?*

A continuación, en la Sección 3.1.3 se presentan los catálogos como el método más utilizado actualmente para la publicación de contenido geográfico en el campo de la IG. Posteriormente, en la Sección 3.2.1, como contribución, se propone una metodología que permite la publicación de forma integrada y automatizada en el flujo de trabajo. Se acompaña con dos casos de estudio reales en los donde se han implementado, en base a la metodología propuesta, sendas soluciones para la publicación de metadatos en catálogos, incluyendo en el segundo caso un nuevo mecanismo para publicar las anotaciones semánticas. Respondiendo a las preguntas: *¿Qué son los catálogos de metadatos?*, *¿Cómo funcionan?, ¿Cómo facilitar y mejorar el proceso de publicación?* y *¿Cómo pueden publicarse recursos en ellos?*

Finalmente, abstrayendo la funcionalidad de indexación de los catálogos, en la Sección 3.1.4 se ha realizado un análisis de distintos tipos de indexación sobre metadatos, así como la integración de estos índices dentro de un sistema de recuperación de información. Concretamente, se analizan diferentes alternativas tanto para la indexación textual de metadatos como para su indexación espacial. Posteriormente, en la Sección 3.3, como contribución, se explora la posibilidad de combinarlos. De forma que respondemos a las preguntas: *¿Qué son los índices?*, *¿Cómo funcionan los índices textuales?*, *¿Cómo funcionan los índices espaciales?* y *¿Cómo pueden combinarse?*

3.1 Estado del Arte

> *"Pero tan pronto como hube adquirido algunas nociones generales de física creí que conservarlas ocultas era grandísimo pecado, que infringía la ley que nos obliga a procurar el bien general de todos los hombres, en cuanto ello esté en nuestro poder."*
>
> *Descartes (Discurso del método, 1637)*

Publicar los recursos supone compartir la información que contienen. El hecho de compartir hace referencia al disfrute en común de un recurso. El principal beneficio de compartir recursos es que estos pueden actualizarse, corregirse, modificarse y reutilizarse por parte de usuarios con intereses similares, de forma que evitamos duplicar esfuerzos tanto físicos como económicos. Pero para que todo esto sea posible es necesario que los recursos sean accesibles, es decir, que estén correctamente publicados, catalogados o indexados de forma que puedan ser encontrados y recuperados posteriormente.

Actualmente la publicación de los recursos y sus metadatos ya no solo es deseable, si no que requerida por iniciativas como INSPIRE, que además exige que se haga de forma interoperable.

La interoperabilidad se define como la capacidad de intercambiar y compartir datos entre dos sistemas o componentes informáticos sin la intervención de un tercer sistema, de modo que la información o datos compartidos puedan ser utilizados sin requerir una comunicación previa [146] [112] [204]. Y puede ser analizada a distintos niveles: tecnológico, sintáctico y semántico [212] [205] [117].

La interoperabilidad técnica es aquella que posibilita la interconexión de los sistemas a nivel de protocolos y el intercambio de información en su nivel más básico [212]. La interoperabilidad sintáctica es aquella que posibilita el intercambio de información en un formato común, incluyendo en este tipo de interoperabilidad aspectos como los formatos estandarizados de datos que intercambian los sistemas [205]. La interoperabilidad semántica es aquella que posibilita el intercambio de información, utilizando un vocabulario común y compartido que evite las inexactitudes en la interpretación del significado de los términos [117].

En esta sección, por una parte, se presentan diferentes especificaciones actuales para la publicación de recursos

georreferenciados que permiten la interoperabilidad técnica y sintáctica entre sistemas. Por otra parte se presentan diferentes estándares de metadatos que posibilitan la interoperabilidad sintáctica y semántica entre sistemas.

3.1.1 Publicación de Recursos Georreferenciados

En el mundo de la IG normalmente la publicación de los datos y los metadatos se realiza por separado. Por una parte, los datos son publicados (puestos accesibles) directamente en servidores especializados como *MapServer*[97] o *GeoServer*[98] que, entre otras, ofrecen interfaces estándar para su visualización (WMS[99]) y descarga (WFS[100]). Por otra parte, la publicación de los metadatos consiste en ponerlos a disposición de los usuarios en un servicio de catálogo para posibilitar el descubrimiento de los recursos que describen. Dado que el propósito perseguido es mejorar el descubrimiento de los recursos, en esta sección nos centraremos en la publicación de los metadatos.

Respecto a la publicación de metadatos en el contexto de la IG, OGC lanzó la especificación *Catalogue Service for the Web*[101] (CSW). CSW define una interfaz de catálogo estándar que permite la publicación y búsqueda de metadatos de datos geoespaciales y servicios. La publicación de los metadatos es posible a través de su perfil transaccional (CSW-T) que especifica operaciones para insertar, actualizar y borrar registros del catálogo de metadatos.

Sin embargo, los estándares de OGC no son el único mecanismo para la publicación y búsqueda de IG o recursos georreferenciados, existen mecanismos más simples. El estándar Z39.50[102] (ISO 23950), ampliamente utilizado en bibliotecas digitales, incluye un perfil GEO[103] que permite extraer metadatos en formato XML a través de Z39.50 y cuyo contenido está basado en estándar CSDGM.

---

[97] http://mapserver.org
[98] http://geoserver.org
[99] http://www.opengeospatial.org/standards/wms
[100] http://www.opengeospatial.org/standards/wfs
[101] http://www.opengeospatial.org/standards/cat
[102] http://www.loc.gov/z3950/agency
[103] http://www.fgdc.gov/standards/projects/GeoProfile

Un mecanismo más general, pero cuya filosofía y funcionamiento puede ser adaptado al campo de los recursos georreferenciados es la iniciativa *Open Archives Initiative*[104] (OAI). Esta iniciativa desarrolla y promueve estándares de interoperabilidad que tienen como objetivo facilitar la difusión eficiente de contenidos. OAI tiene sus raíces en los movimientos de acceso abierto e interoperabilidad de los repositorios institucionales. Con el tiempo, sin embargo, el trabajo de la OAI se ha expandido para promover un amplio acceso a recursos digitales para *eScholarship*, *eLearning* y *eScience*.

OAI proporciona la especificación *Open Archives Initiative Object Reuse and Exchange*[105] (OAI-ORE) que define estándares para la descripción e intercambio de agregaciones de recursos Web. Estas agrupaciones, a veces llamadas objetos digitales compuestos, pueden combinar los recursos distribuidos con múltiples tipos de medios, incluyendo texto, imágenes, datos y video. El objetivo de estos estándares es exponer el contenido de estas agregaciones a aplicaciones que soportan la creación, el depósito, intercambio visualización, reutilización y preservación de contenidos digitales.

Por otra parte, pueden explorarse los mecanismos de publicación más generales e intentar dotarlos con un contexto espacial como propone la *Geospatial Web* [103] [193]. En este sentido es posible publicar recursos georreferenciados directamente en servicios o redes sociales impulsadas por la Web 2.0 que soporten este tipo de información, usando los metadatos para documentarlos de forma apropiada. Por ejemplo, es posible publicar fotografías en *Flickr*, pequeños textos en *Twitter* o rutas de GPS en *Wikiloc*. La publicación de recursos en este tipo de servicios debe realizarse a través de la interfaz (API) específica de cada uno de ellos.

Otra estrategia de publicación consiste en poner los recursos disponibles para que sean indexados por los robots de los motores de búsqueda de Internet, esta forma de publicar recursos se realiza de forma efectiva con la mayoría del contenido disponible en Internet como las páginas HTML. Para ello, podemos explorar diferentes posibilidades que van desde simplemente dejar los recursos

[104] http://www.openarchives.org
[105] http://www.openarchives.org/ore

disponibles a crear un archivo KML asociado con los metadatos que los describa [1].

La especificación de estándares para la publicación de recursos georreferenciados, cuya implementación sea viable desde un punto de vista técnico y económico, resulta esencial para el progreso de la tecnología y los servicios.

3.1.2 Estándares de Metadatos para IG

Los elementos de metadatos agrupados en conjuntos diseñados para un propósito específico, por ejemplo, para un dominio específico o un tipo particular de recurso, se denominan esquemas de metadatos. Los esquemas de metadatos que son desarrollados y mantenidos a partir de opiniones de expertos en la materia por organizaciones de normalización u organizaciones que han asumido tal responsabilidad se denominan estándares de metadatos.

Los estándares de metadatos son requisitos que tienen por objeto establecer un entendimiento común del significado o semántica de los datos, para garantizar el uso y la interpretación correcta y adecuada de los datos por parte de sus propietarios y los usuarios. Para lograr este entendimiento común es necesario definir una serie de características o atributos de los datos.

Los estándares de metadatos a menudo empiezan como esquemas desarrollados por una comunidad de usuarios particular para permitir la mejor descripción posible de un tipo de recurso de acuerdo a sus necesidades. El desarrollo de estos esquemas tiende a ser controlado a través de un consenso comunitario combinado con procesos formales para la presentación, aprobación y publicación de nuevos elementos. Por lo general, estos esquemas, contienen definiciones semánticas de los elementos y formas estandarizadas de representarlos en formatos digitales, tales como bases de datos o XML. Como hemos visto en el capítulo anterior, este último se está convirtiendo rápidamente en muchas comunidades en el estándar *de facto* de los lenguajes de marcado. Las definiciones semánticas de los elementos contemplan aspectos tanto de su estructura como de su contenido. Asegurar una estructura consistente permite el intercambio de datos y la consulta, gestionar el proceso de creación, registrar la

procedencia y los procesos técnicos y la gestión de permisos de acceso. Mientras que definir el contenido de los elementos pretende garantizar búsquedas eficaces a través de una entrada de datos consistente y la inclusión de los puntos de acceso utilizando vocabularios controlados, como archivos autoritativos, tesauros o esquemas de codificación.

Por lo tanto los estándares de metadatos resultan esenciales para la correcta publicación y descubrimiento de los recursos dentro de un sistema de información interoperable. A continuación se presenta el estándar de metadatos más conocido a nivel general y seguidamente, centrando la atención en el mundo de la IG, se describen distintas recomendaciones surgidas a nivel internacional, europeo o nacional. Finalmente, se realiza una pequeña discusión sobre estas normas y recomendaciones.

Dublin Core

A nivel general, la alternativa de metadatos para la catalogación más utilizada a nivel mundial es el esquema *Dublin Core*[106] (DC), creado y mantenido por la *Dublin Core Metadata Initiative* (DCMI). Más concretamente el esquema usado es el *Dublin Core Metadata Element Set v1.1* o su versión más reducida, la norma ISO 15836:2003 [116], lo que se conoce habitualmente como *"DC simple"* y que incluye sólo 15 elementos de metadatos. Esos quince elementos básicos para describir cualquier recurso, se presentan habitualmente divididos en tres grupos que indican la clase o alcance de la información incluida en ellos, y que responden, en cierta medida, a las expectativas que tiene el usuario cuando se enfrenta a la información en la red. Algunas de las fortalezas de este esquema de metadatos son:

- Su simplicidad
- La independencia sintáctica.
- Alto nivel de normalización formal: ANSI/NISOZ39.85-2001, ISO 15836-2003.
- Crecimiento y evolución del estándar a través de una institución formal constituida en consorcio: la DCMI.
- El conjunto de elementos DC se ha convertido en una infraestructura operacional del desarrollo de la Web Semántica.

[106] http://dublincore.org

Content Standard for Digital Geospatial Metadata (CSDGM)

La norma *Content Standard for Digital Geospatial Metadata* (CSDGM) [76] [75] fue creada en 1994 por el FGDC de EE.UU. para dar soporte a la construcción de su IDE nacional. Aunque es una norma a nivel nacional, fue la primera en aparecer y se ha difundido a nivel internacional dada su integración en diversas herramientas SIG o su utilización en redes distribuidas de catálogo a nivel internacional.

El objetivo de esta norma es proporcionar un conjunto común de terminología y definiciones para la documentación de datos geoespaciales digitales. Define la información requerida por los usuarios de información espacial para determinar la disponibilidad del conjunto de datos, la propiedad para un uso determinado, el acceso al conjunto de datos y la forma de transferencia de los datos.

ISO19115:2003

La *International Organization for Standardization*[107] (ISO) creó en 1992 el comité ISO/TC 211 con responsabilidades en Información Geográfica y Geomática. Esté comité se ha encargado de preparar una familia de normas en este contexto, y entre ellas, un conjunto de normas relacionadas con metadatos para la descripción de recursos geoespaciales [129].

La norma internacional ISO 19115:2003 [115] para metadatos de IG se aprobó en mayo de 2003 y define elementos que permiten describir, entre otros, la identificación, extensión, calidad, el esquema de representación espacial, los sistemas de referencia utilizados, y la forma de distribución de los datos. Aunque esta norma está principalmente orientada a la catalogación de conjuntos de datos geográficos (incluyendo también series o fenómenos geográficos individuales) en formato digital, también se puede extender a otras formas de datos geográficos como mapas, documentos textuales, datos no geográficos o servicios.

Esta norma de metadatos es muy compleja e incluye una extensa serie de elementos de metadatos, unos obligatorios y otros opcionales. El documento consta de 140 páginas, incluye un total de 409 ítems y define 27 listas controladas, mediante las que se definen

[107] http://www.iso.org

los posibles valores de ciertos campos. Para su elaboración fue necesaria la colaboración de 33 países miembros de ISO/TC211 y un total de 16 países que aportaron expertos al Grupo de Trabajo encargado de su definición. En 1996 se disponía ya de un primer borrador, en el año 2003 se aprobó el texto definitivo como Norma Internacional de metadatos que fue adoptada como Norma Europea por CEN/TC 287 en 2005. AENOR (Asociación Española de Normalización) ha decidido también su adopción como Norma Española (UNE-EN ISO19115).

Aunque esta norma define un extenso número de elementos de metadatos, establece un “conjunto mínimo” de metadatos (el núcleo o *Core*) a considerar para todo el rango de aplicaciones de los metadatos (desde mapas en formato papel a conjuntos de datos en formato digital, como imágenes satélite, modelos digitales del terreno, etc.). Con este conjunto se pretende establecer unos mínimos para facilitar el descubrimiento, el acceso, la transferencia y la utilización de los datos. Este núcleo está formado por elementos obligatorios y otros opcionales que, usados todos ellos, aumenta la interoperabilidad de los datos y permite a los usuarios entenderlos sin ambigüedades (ver Figura 18).

Figura 18: Elementos del Núcleo de ISO19115

Especificaciones de metadatos en el ámbito de OGC

Open Geospatial Consortium (OGC) es una organización internacional comprometida en un esfuerzo cooperativo para crear especificaciones informáticas abiertas en el área de

geoprocesamiento. Como parte de su borrador *"OpenGIS Abstract Specification"*, OGC dedica una sección al registro de metadatos para datos espaciales. OGC está colaborando estrechamente con FGDC e ISO/TC211 para generar estándares de metadatos espaciales globales. En una reunión plenaria en Viena (Austria), en marzo de 1999, ISO/TC211 recibió con satisfacción la realización del acuerdo de cooperación entre OGC e ISO/TC 211 y aceptó los términos de referencia para un grupo de coordinación ISO/TC211/OGC.

Estos grupos han tenido diferentes ideas acerca de qué características hay que incluir. El proveedor de datos necesita un gasto considerable de tiempo y recursos si quiere hacerse con la información, y para el usuario de datos el detalle puede ser mayor del que necesita para una investigación inicial. Por consiguiente, en muchas situaciones se necesitan definir diferentes niveles de metadatos, con capacidad para llegar a niveles crecientes de detalle. Así pues, los metadatos deben variar de acuerdo con el propósito.

Las organizaciones ISO y OGC firmaron en 1999 un acuerdo de cooperación para un consenso técnico en sus respectivos desarrollos. Esto se realiza por medio de revisiones mutuas y el desarrollo de borradores. Cuando se identifica un punto discordante se toman las medidas necesarias para llegar a un punto de vista común y consensuado. Este acuerdo ha permitido la definición de varios estándares adoptados directamente de entre las numerosas especificaciones de OGC. De manera similar, OGC adoptó el Esquema Espacial ISO 19107:2003 para su especificación de geometría y topología (*Simple Features Profile*) y un formato de archivo XML llamado GML (*Geographic Markup Language*).

Entre los beneficios del acuerdo entre estas dos organizaciones se pueden mencionar:

- Aquellos productos conformes a especificaciones OGC también se ajustarán a los estándares ISO.
- Tanto los estándares de ISO como los de OGC se verán beneficiados por la estrecha colaboración que conlleva a mejores estándares.
- Se logrará un reconocimiento internacional de excelencia en las disciplinas GIS.

Normas de metadatos en el ámbito europeo

En el ámbito europeo hay que mencionar el trabajo desarrollado en la última década. En 1998 el *European Commitee for Standardization* (CEN) elaboró a través de su grupo de trabajo CEN/TC287 (con responsabilidad para estándares de información geográfica) una norma europea voluntaria ENV (*European Norm Voluntary*) 12657 que llevaba por título *"Geographic Information-Data Description Metadata"*. Dicha norma fue adaptada al contexto español para definir los metadatos que debían acompañar a los datos que, hasta entonces, cumplían con la norma española MIGRA (Mecanismo de Intercambio de Información Geográfica Relacional formado por Agregación) [9].

También cabe mencionar las implicaciones de la aprobación de la directiva INSPIRE en el impulso a la creación de metadatos en los distintos estados miembros de la Unión Europea. La Directiva 2007/2/CE de la Unión Europea INSPIRE, aprobada el 14 de marzo de 2007, coordina y gestiona datos geoespaciales y la interoperabilidad de los servicios de datos en Europa. Para su desarrollo se está trabajando en la definición de normas de ejecución (*implementing rules*) [72], y una de ellas hace referencia a la creación de metadatos [73] [179]. Estas normas de ejecución para metadatos definen a un nivel abstracto aquellos descriptores que resultan esenciales para el descubrimiento de datos y servicios. Estos descriptores o elementos de metadatos constituyen el conjunto mínimo de los elementos necesarios para cumplir con la Directiva INSPIRE, pero no se descarta la posibilidad de que las organizaciones documenten sus recursos más extensamente con elementos adicionales derivados de diferentes normas internacionales. Además estas normas de ejecución van acompañadas de guías que establecen la correspondencia entre los elementos de metadatos definidos en las normas de ejecución y las normas de metadatos internacionales como ISO19115:2003 o Dublin Core.

Normas de metadatos en el ámbito español

En el contexto español hay que mencionar el desarrollo del Núcleo Español de Metadatos (NEM), que nace gracias a las opiniones, comentarios y aportaciones de un grupo abierto de expertos en metadatos pertenecientes a diferentes organizaciones e instituciones en el ámbito nacional, autonómico y local.

Es una recomendación de metadatos aprobada por el Consejo Superior Geográfico, a través de su Comisión de Geomática. NEM se define como un conjunto mínimo de metadatos entendidos como un perfil de ISO 19115:2003 de acuerdo con el concepto de perfil definido en la Norma ISO 19106:2004 "*Geographic Information-Profiles*", es decir, es un modo particular y concreto de aplicar y utilizar una Norma, seleccionando un conjunto de ítems y un conjunto de parámetros opcionales. Para ello este perfil va a tener en cuenta otras iniciativas y acciones relevantes que en la actualidad se están desarrollando en materia de metadatos.

Este perfil constituye por lo tanto un núcleo (*Core*) o conjunto de metadatos mínimo aconsejable por su utilidad y relevancia que va a permitir realizar, entre otros, búsquedas y comparaciones a partir de metadatos que proceden de diferentes fuentes, sobre distintos conjuntos de datos, de una manera rápida, práctica, fácil y fiable. Fue definido para ser utilizado por todos los catálogos generados en las diferentes organizaciones relacionadas con la información geográfica de manera que se consiga la interoperabilidad de metadatos en toda España.

No es, por lo tanto, un perfil normativo o restrictivo. No se pretende que se implemente directamente sino que se aconseja su utilización. Cada institución u organismo debe estudiar cuales son los metadatos que considera adecuados para satisfacer sus necesidades y, una vez establecidos, se recomienda incluir al menos los ítems que establece el perfil NEM, garantizando así la compatibilidad con el resto de iniciativas.

Discusión sobre estándares de metadatos

A lo largo de los años han ido surgiendo a nivel nacional, europeo o dentro de un dominio específico, un conjunto de iniciativas para normalizar la creación de metadatos. Sin embargo, todas estas iniciativas han ido derogándose en busca de la armonización con la norma internacional ISO 19115:2003. Incluso, la nueva versión de la norma americana CSDGM convergerá con la norma internacional.

Un aspecto importante relativo a los esquemas de metadatos es su nivel de detalle, que viene definido mediante la elección de la propia norma y la creación de extensiones especiales y perfiles. En primer lugar la norma elegida define un conjunto más o menos grande de

elementos con diferente condicionalidad: obligatorios, obligatorios si aplicable y opcionales. Una extensión de la norma consiste habitualmente en la adición de nuevas restricciones (ej., conversión de elementos opcionales en obligatorios), ampliación de listas de códigos y la creación de nuevos elementos y entidades. ISO 19115:2003 y CSDGM proporcionan métodos dentro de la propia norma para la extensión de los metadatos. Y si esas características adicionales son muy amplias (involucran la creación de un número considerable de elementos), ISO 19115:2003 recomienda la petición formal de creación de un perfil de aplicación específico para la comunidad de usuarios que lo requiera.

Sin embargo, aunque los perfiles específicos y la condicionalidad de los elementos facilitan cierta flexibilidad de los metadatos geográficos, hay que reconocer que resultan todavía muy detallados y complejos de manejar. CSDGM e ISO 19115:2003 definen más de 350 elementos cada uno distribuidos en múltiples secciones jerárquicas. Esta complejidad implica que para completar los metadatos geográficos haya que dedicar gran cantidad de tiempo y recursos humanos altamente cualificados. Este problema está provocando que muchas organizaciones se planteen el uso de otras normas de metadatos más sencillas y de propósito general (por ejemplo Dublin Core), focalizando los esfuerzos disponibles en mantener al día al menos los metadatos de descubrimiento que se mencionaban con anterioridad.

3.1.3 Servicios de Catálogo

Los servicios de catálogo son actualmente el método más utilizado para la publicación de contenido geográfico en el campo de la IG. Estos sistemas realizan todas las operaciones que define el término indexación cuando se habla de sistemas de búsqueda. Por lo tanto, los catálogos recogen, procesan y almacenan los metadatos de los recursos para su posterior recuperación. El proceso de recogida por lo general requiere de la interacción del usuario con el sistema ya que, en la mayoría de los casos, los metadatos deben ser cargados manualmente. Estos metadatos son el elemento clave en el proceso de indexación ya que es la información utilizada para la recuperación de los recursos y no los recursos en sí.

Las aplicaciones de catálogo normalmente implementan la especificación estándar CSW de OGC. Los proveedores de datos simplemente suben sus metadatos directamente a la aplicación de servicio de catálogo con el fin de ponerlos a disposición de los demás. Entonces, los metadatos son procesados y almacenados en la aplicación a la espera de ser recuperados a través de las interfaces de descubrimiento expuestas. Este proceso implica la inmediata publicación, almacenamiento y accesibilidad del contenido, pero también implica interacción humana y control durante el proceso.

Los catálogos son herramientas muy útiles y potentes para el descubrimiento de contenido geográfico. Sin embargo, su uso está más centrado en los usuarios profesionales ya que los catálogos requieren un cierto grado de conocimiento o experiencia para la creación de los metadatos y su publicación.

A continuación se presentan las aplicaciones de catálogo más populares [207]:

GeoNetwork

GeoNetwork[108] es una aplicación de catálogo basada en estándares, de código abierto y libre para la gestión de recursos georreferenciados a través de la web. Proporciona potentes funciones de edición de metadatos y de búsqueda, así como un visor de mapas interactivo embebido en la propia página web. Es la implementación de referencia del estándar CSW.

GeoNetwork ofrece un entorno de gestión de información espacial estandarizado y descentralizado, diseñado para permitir el acceso a bases de datos georreferenciadas, productos cartográficos y metadatos relacionados de una gran variedad de fuentes, mejorando el intercambio y la compartición de información espacial entre las organizaciones y su audiencia, utilizando las capacidades de Internet. Esta aproximación de la gestión de la información geográfica tiene como objetivo facilitar a una amplia comunidad de usuarios de información espacial el acceso rápido a los datos espaciales disponibles y a los mapas temáticos existentes, en base a los que se pudiera apoyar la toma de decisiones.

[108] http://geonetwork-opensource.org

El principal objetivo de este software es mejorar la accesibilidad de una amplia variedad de datos, junto con la información asociada, a escala diferente y desde fuentes multidisciplinarias, organizados y documentados de una forma estandarizada y consistente. El reto es aumentar el intercambio y la compartición de datos entre las organizaciones para evitar la duplicación, incrementar la cooperación y la coordinación de esfuerzos en la recopilación de datos y ponerlos disponibles para beneficiar a todos, ahorrando recursos y al mismo tiempo preservando los datos y su información asociada.

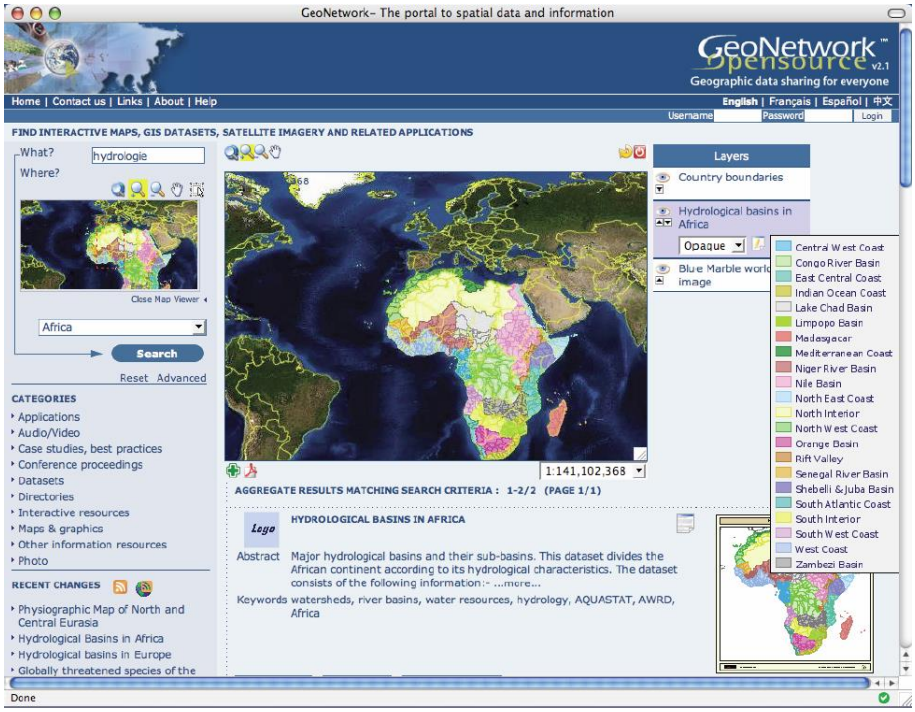

Figura 19: Página principal de *GeoNetwork*

Este servidor de catálogo se ha convertido en el más popular de su clase especialmente debido a su apuesta por el software libre, por su atractiva interfaz de usuario (ver Figura 19), por su interoperabilidad y por el estricto cumplimiento de los estándares tanto en los referentes a formatos de metadatos como los que se refieren a la implementación de los servicios.

Esri Geoportal Server

Esri Geoportal Server[109] es una aplicación de catálogo de código abierto, que permite acceder a recursos como *datasets*, raster y servicios web a través del estándar CSW. Además, ayuda a las organizaciones a gestionar y publicar metadatos para sus recursos geoespaciales, permitiendo a los usuarios finales descubrirlos y acceder a ellos.

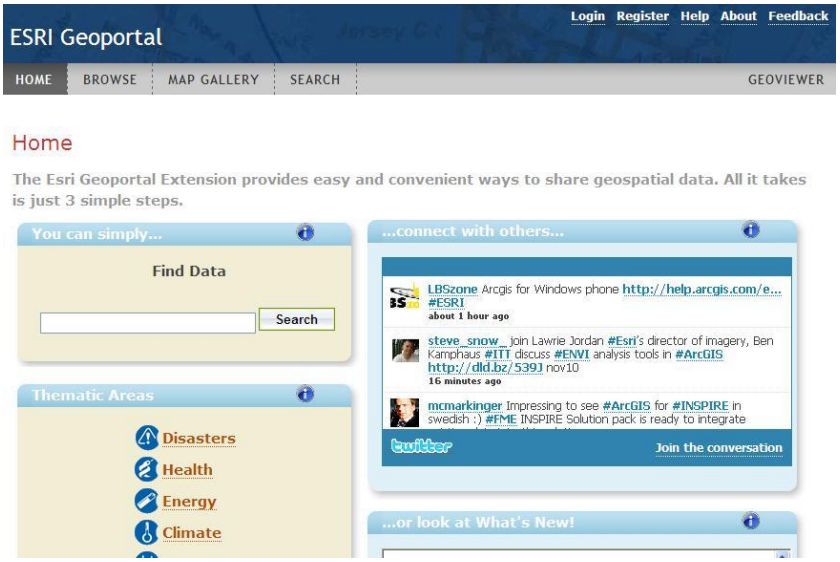

Figura 20: Página principal de *Esri Geoportal Server*

Geoportal ayudar a mantener la calidad, actualidad y disponibilidad de los recursos registrados, proporcionando herramientas para evaluar las nuevas entradas, controlar el acceso a los metadatos y recursos y, además, está integrado con otros productos de la compañía para facilitar la publicación y acceso a los recursos.

Los productores de datos publican sus recursos en el geoportal registrando los metadatos del recurso en el servicio de catálogo. También incluye herramientas sencillas para la generación y registro de metadatos, de forma que es posible tanto crear metadatos directamente en la aplicación web como cargar metadatos existentes. Por otra parte, *Geoportal* proporciona una interfaz sencilla para que los usuarios puedan descubrir y acceder a los recursos publicados (ver Figura 20).

[109] http://www.esri.com/software/arcgis/geoportal

deegree

deegree[110] es un software de código abierto para las IDEs y la web geoespacial. *deegree* incluye componentes para la administración de datos geoespaciales, incluyendo acceso a datos, visualización, descubrimiento y seguridad. Está basado en los estándares abiertos de OGC y de ISO/TC 211. Entre sus componentes incluye un servidor de catálogo[111], que implementa el estándar CSW incluyendo su perfil transaccional.

eXcat

eXcat[112] es una herramienta muy eficaz para la publicación de metadatos usando el estándar CSW de OGC. *eXcat* está escrito en Java y consta de un servidor y un cliente CSW. Destaca de otros programas similares porque está desarrollado sobre el sistema de base de datos XML de código abierto *eXist*[113], que utiliza tecnologías web estándar como *XPath*[114] y *XQuery*[115]. *eXcat* es muy fácil de usar y permite que los metadatos sean almacenados directamente en formato XML (por ejemplo, ISO 19115/19139) y que sean publicados y consultados a través del estándar CSW. Además, esta aplicación de catálogo, presenta los metadatos convenientemente a los usuarios y de una manera uniforme.

3.1.4 Indexación

Con la vista puesta en el descubrimiento de los recursos, en esta sección vamos a estudiar el funcionamiento de diferentes técnicas de indexación, también usadas internamente en los catálogos, con el fin de proponer soluciones más automatizadas. Por lo que entramos de lleno en el campo de la recuperación de información (*Information Retrieval*[116]).

---

[110] http://www.deegree.org
[111] http://wiki.deegree.org/deegreeWiki/deegree3/CatalogueService
[112] http://gdsc.nlr.nl/gdsc/en/tools/excat
[113] http://exist.sourceforge.net
[114] http://www.w3.org/TR/xpath
[115] http://www.w3.org/TR/xquery
[116] http://en.wikipedia.org/wiki/Information_Retrieval

La recuperación de información es la actividad consistente en obtener recursos relevantes a una necesidad de información a partir de una colección de recursos disponibles. Las búsquedas pueden estar basadas en el propio contenido del recurso si, por ejemplo, indexamos por completo el contenido de documentos de texto o basadas en los metadatos que describen a los recursos.

La Figura 21 extraída de [11], libro de referencia en el campo de la recuperación de información, muestra una arquitectura *software* simple y genérica que posibilita el proceso de recuperación. Debemos tener en cuenta que esta arquitectura está diseñada para documentos textuales, sin embargo, se podría adaptar fácilmente al contexto de la IG, perfeccionando el proceso de indexación para incluir la localización de los recursos (como veremos en este capítulo) y ofreciendo servicios de búsqueda especializados (como veremos en el capítulo siguiente).

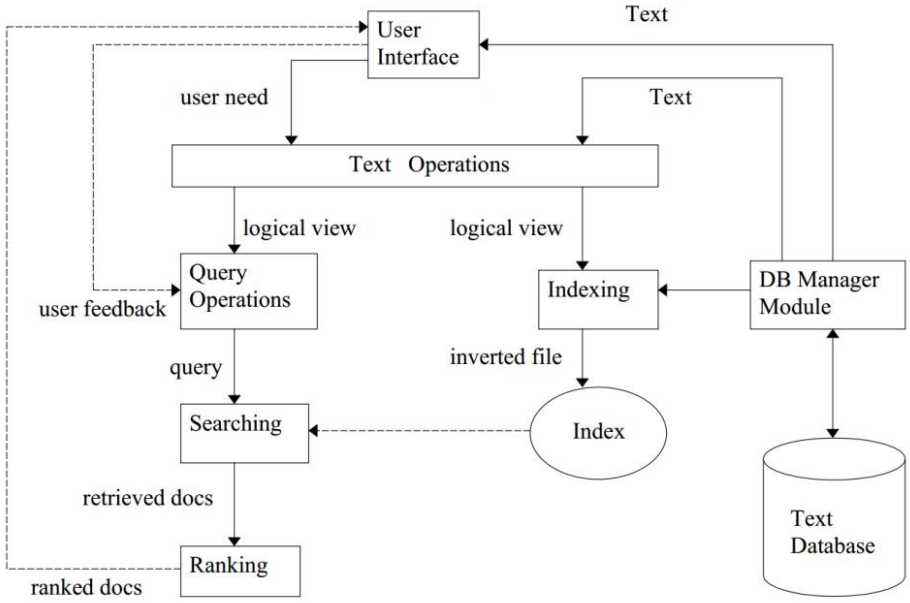

Figura 21: El proceso de recuperación de información [11]

El proceso de recuperación implica un preproceso en el que se recopilan los recursos (*Text Database*) y se analizan para obtener una vista lógica (*logical view*) de ellos, por ejemplo generando metadatos que describan dichos recursos de forma homogénea. A partir de la vista lógica se construye un índice (*Index*), un índice es una estructura

de datos crítica porque permite la búsqueda rápida sobre grandes volúmenes de datos. Dependiendo de la naturaleza de los recursos se pueden usar diferentes estructuras de índices, en la Figura 21 aparece representada la más popular a nivel textual: los índices invertidos (*inverted file*) [240].

Una vez indexados los recursos, el proceso de recuperación puede iniciarse. Por ello, en esta tesis se considera la indexación como un proceso de publicación de los recursos. Diferentes aspectos sobre el proceso de recuperación serán tratados en el siguiente capítulo dedicado al descubrimiento de recursos georreferenciados, pero para ponernos en contexto, de forma muy resumida, el usuario en primer lugar deberá expresar sus necesidades de información en forma de consulta (*query*). Esta consulta será procesada para obtener unos recursos como resultado (*retrieved docs*). El procesado rápido y eficaz de las consultas se realiza gracias al índice previamente generado.

Por ello, se ha realizado un análisis de distintos tipos de indexación sobre metadatos, así como la integración de estos índices dentro de un sistema de recuperación de información. Concretamente, se analizan diferentes alternativas tanto para la indexación textual de metadatos como para su indexación espacial.

Índices Textuales

Como hemos visto en el capítulo anterior, los recursos estarán descritos por metadatos representados en lenguaje XML, que son los elementos que vamos a indexar. La mayor parte de los metadatos que describen los recursos pueden ser considerados como texto.

En esta sección exploraremos algunos índices que nos permiten realizar consultas basadas en términos o palabras clave, que son las realizadas típicamente en la web. Este tipo de consultas incluye aquellas que se realizan empleando términos clave de forma individual, las que usan frases formadas por términos clave y las que emplean operadores lógicos para combinar términos o frases.

Índices invertidos

Como hemos dicho anteriormente, el índice invertido (*inverted file*) es el método de indexación textual más conocido. En [240] se puede encontrar una buena descripción de esta estructura y de los avances

relacionados con ella que se han realizado en los últimos años. El índice invertido es muy fácil de mantener y permite resolver de manera eficiente consultas basadas en términos clave, sobre todo cuando se buscan los términos clave de manera individual.

Un índice invertido es una estructura de datos que almacena una correspondencia del contenido (como palabras o números) con sus ubicaciones en un archivo de base de datos, en un documento o en un conjunto de documentos. El propósito de un índice invertido es permitir búsquedas textuales rápidas, pero tiene un coste de preprocesado mayor cuando un documento es añadido. Es el método más popular en sistemas de recuperación de información, por ello es ampliamente usado en motores de búsqueda.

Existen dos variantes principales de índices invertidos: los que contienen una lista de referencias a documentos para cada palabra (*inverted index*) y los que además contienen la posición de esa palabra dentro del documento (*full inverted index*). Estos últimos índices ofrecen más funcionalidad, como la búsqueda de frases, pero requieren más recursos espaciales y temporales para ser creados.

R_0: “eso es lo que es”
R_1: “que es eso”
R_2: “eso es queso”

inverted index

es	{0, 1, 2}
eso	{0, 1, 2}
lo	{0}
que	{0, 1}
queso	{2}

full inverted index

es	{(0, 1), (0, 4), (1, 1), (2, 1)}
eso	{(0, 0), (1, 2), (2, 0)}
lo	{(0, 2)}
que	{(0, 3), (1, 0)}
queso	{(2, 2)}

Figura 22: Ejemplo de índices invertidos

A modo ilustrativo, la Figura 22 muestra un pequeño ejemplo de la construcción de los índices invertidos. En primer lugar podemos observar el conjunto de recursos textuales que va a indexar el sistema (R_0, R_1 y R_2). Posteriormente observamos el índice invertido (*inverted*

index) correspondiente, donde los enteros en la notación de conjuntos se corresponden con los subíndices de los recursos que los contienen. Al realizar una consulta de búsqueda sobre los términos *"que"*, *"es"* y *"eso"* sobre este índice, nos devolvería el conjunto *{0, 1}* ∩ *{0, 1, 2}* ∩ *{0, 1, 2} = {0, 1}* que indica que sólo los dos primeros recursos contienen los términos especificados. A continuación, en base a los mismos recursos, encontramos el índice invertido a nivel de palabra (*full inverted index*), donde además de indicar en qué recursos aparecen las palabras también indicamos su posición. Entonces, *"queso": {(2, 2)}* significa que la palabra *"queso"* aparece en el tercer recurso y es la tercera palabra dentro de ese recurso (en ambos casos se empieza a contar desde cero). Si lanzamos la consulta de búsqueda sobre la frase *"que es eso"* sobre este índice obtenemos que todos los términos aparecen en los recursos R_0 y R_1 como en el caso anterior, pero sólo lo hacen de forma consecutiva en R_1.

A lo largo de los años, se han ido proponiendo variantes sobre esta estructura básica para mejorar la eficiencia de determinados tipos de consultas y para disminuir el espacio necesario para su almacenamiento [11].

Alternativas

Aunque el índice invertido es la técnica más eficiente para resolver búsquedas de términos clave de manera individual, cuando se realizan otros tipos de consultas menos habituales, como pueden ser las búsquedas de frases, existen otras estructuras como los *arrays de sufijos* [138] que son más rápidas. Sin embargo, tienen el inconveniente de ser más complicados de construir y de mantener.

Por otra parte, existen técnicas de compresión de textos que resultan útiles no sólo para ahorrar espacio en disco sino también para mejorar los tiempos de procesamiento, de transmisión y de transferencia a disco a la hora de realizar búsquedas [213] [159]. El principio de estas técnicas de compresión es la sustitución de las palabras de los recursos por códigos, normalmente de longitud variable, que suelen ser más cortos a medida que aumenta la frecuencia de aparición del carácter o de la palabra y más largos a medida que disminuye la misma.

Existen muchas otras técnicas de indexación textual, pero se considera que los índices invertidos son suficientes para el alcance de

esta tesis. Además, en las siguientes secciones veremos técnicas de indexación espacial y cómo combinar ambos tipos de indexación.

Implementaciones

A continuación se presentan un conjunto de implementaciones actuales que proporcionan funcionalidades de indexación. Todas ellas utilizan algoritmos de indexación basados en los índices invertidos.

Lucene[117] es un API de código abierto para recuperación de información promovido por la *Apache Software Foundation*[118]. Es útil para cualquier aplicación que requiera indexado y búsqueda a texto completo. Lucene ha sido ampliamente usado por su utilidad en la implementación de motores de búsquedas. La base de la arquitectura de Lucene es el concepto de Documento (*Document*) que contiene Campos (*Fields*) de texto. Esta flexibilidad permite a Lucene ser independiente del formato del fichero. Textos que se encuentran en ficheros PDF, XML, páginas HTML, documentos de Microsoft Word, así como muchos otros pueden ser indexados siempre que se pueda extraer información textual de ellos.

Xapian[119] es una librería de código abierto para la recuperación probabilística de información, publicada bajo licencia GPL. Es decir, es una librería completa que sirve como motor de búsqueda de texto. Está escrito en C++, pero proporciona enlaces para permitir su uso desde otros lenguajes como Perl, Python, PHP, Java, etc. Xapian es altamente portable y multiplataforma. Xapian está diseñado para ser un conjunto de herramientas altamente adaptables que permiten a los desarrolladores agregar fácilmente indexación y búsqueda avanzada en sus propias aplicaciones.

Managing Gigabytes for Java[120] *(MG4J)* es un motor de búsqueda textual diseñado específicamente para grandes colecciones de recursos. Es altamente personalizable y ofrece un buen rendimiento. Ofrece todas las funcionalidades más comunes. MG4J utiliza algoritmos de compresión de cadenas [24] [223], lo que lo hace eficiente en la gestión de repositorios voluminosos. El problema de

[117] http://lucene.apache.org
[118] http://www.apache.org
[119] http://xapian.org
[120] http://mg4j.di.unimi.it

estos algoritmos es que no están especialmente orientados al manejo de ficheros en formato XML, sino que están planteados para ser utilizadas en documentos de texto plano.

XQEngine[121] es una implementación libre de un motor de búsqueda basado en el lenguaje de consulta *XQuery*[122] definido por W3C. Resulta interesante porque es capaz de indexar los documentos completos, es decir, indexa todas sus etiquetas. Esto permite realizar búsquedas por cualquier propiedad del metadato. Como contrapartida requiere una gran cantidad de memoria en ejecución, lo que no lo hace adecuado para grandes volúmenes de recursos. Otra de las desventajas que tiene es que las consultas deben realizarse usando el nombre exacto de la etiqueta. Esto significa que debemos conocer el estándar del metadato para poder consultar adecuadamente.

Sphinx[123] es un motor de búsqueda abierto diseñado con el fin de indexar contenidos de bases de datos. Actualmente soporta de manera nativa *MySQL*, *PostgreSQL* y bases de datos ODBC. Otras fuentes de datos pueden ser indexadas mediante el apropiado filtro XML. Soporta tanto indexación incremental como en lote. Ofrece una gran variedad de funciones de procesamiento de texto que permiten afinar Sphinx para los requerimientos particulares de la aplicación, y también una serie de funciones de relevancia que permiten ajustar la calidad de búsqueda.

Discusión

Se estudia la creación de índices textuales para acelerar las búsquedas y la devolución de metadatos de IG, por ello, se han realizado varias pruebas con estas implementaciones [48]. Se evaluaron el tiempo necesario para la creación de los índices sobre diferentes cantidades de descripciones de recursos y el tiempo de respuesta a diferentes tipos de consultas.

Concluyendo que, pese a tener unos tiempos de indexación mayores a otras implementaciones, la opción que mejor se comporta es la del proyecto Lucene, ya que los tiempos de respuesta a consultas son siempre similares, independientemente de la

[121] http://xqengine.sourceforge.net
[122] http://www.w3.org/TR/xquery
[123] http://sphinxsearch.com/about/sphinx

complejidad de las mismas y el número de metadatos indexados. Además de su rendimiento y escalabilidad, Lucene es un proyecto mucho más maduro que otros y ofrece una amplia funcionalidad. Adicionalmente, al ser un proyecto Apache se asegura una numerosa y activa comunidad de desarrolladores, mucha documentación y un buen soporte.

Por otra parte, existen estudios más amplios como [152] donde se comparan hasta 29 motores de búsqueda, incluyendo algunos de los que se han presentado en esta sección, y se analizan aspectos como el tiempo de indexado, el tamaño del índice, consumo de recursos, tiempo de búsqueda, precisión, etc. Sobre una colección de recursos HTML. Estos estudios también dejan a Lucene en una buena posición dado su buen tiempo de indexación y consulta, el pequeño tamaño de sus índices y su reducido consumo de recursos. Aunque resaltan que la elección final depende de las necesidades específicas y objetivos particulares de cada usuario.

Índices espaciales

Dada la naturaleza de los recursos que manejamos en el contexto de la IG, como hemos visto en el capítulo anterior, gracias a los metadatos que los describen podemos conocer la ubicación espacial de lo que representa el recurso.

Por otra parte, la recuperación de recursos dentro de un sistema de información puede ser costosa por la gran cantidad de información que contiene. En nuestro contexto resulta muy interesante poder recuperar los recursos teniendo en cuenta su localización, por ello en esta sección estudiaremos algunas estrategias de indexación que aceleren la búsqueda y devolución de los recursos respecto a consultas de tipo espacial.

A lo largo de los años, se han propuesto muchas estructuras diferentes para lograr este objetivo. Dichas estructuras se pueden clasificar de manera general en métodos de acceso a puntos o *Point Access Methods* (PAM) y métodos de acceso espacial o *Spatial Access Methods* (SAM). Los métodos del primer grupo se caracterizan por mejorar el acceso a recursos representados por puntos espaciales. En cambio, los métodos del segundo grupo son más generales ya que trabajan con recursos representados por todo tipo de objetos geográficos (puntos, líneas, polígonos, etc.). Durante los

últimos años se han propuesto muchos PAMs y SAMs diferentes. En [83] y [189] se puede encontrar una buena descripción de las estructuras más relevantes en cada categoría. Consideramos importante la posibilidad de incluir geometrías complejas, por lo que nos centraremos en los SAMs.

Estrategias de indexación espacial

A continuación se presentan las dos estrategias de indexación espacial más comunes, a partir de las cuales derivan la mayoría de las técnicas existentes actualmente:

Los índices *Quadtree* [190] están basados en una estructura de datos jerárquica cuya propiedad es que está basada en el principio de descomposición recursiva del espacio. Es una estructura de tipo árbol en la que cada nodo interno tiene exactamente cuatro hijos, obtenidos al subdividir el espacio de forma recursiva en cuatro cuadrantes o regiones. Este tipo de índices permiten solventar consultas espaciales de una manera eficaz, pero con la limitación de que las consultas deben restringirse a un área de forma rectangular. La Figura 23 muestra de forma gráfica como se construyen estos índices.

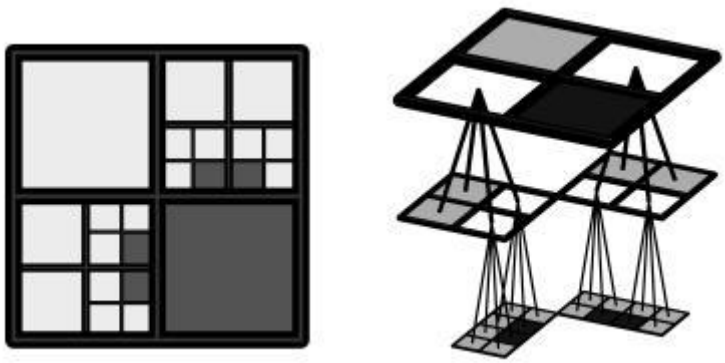

Figura 23: Ejemplo de *Quadtree*

Los índices *R-tree* [98] son uno de los métodos de la categoría SAM más populares. Esta estructura se basa en un árbol balanceado derivado del *B-tree* que divide el espacio en rectángulos de cobertura mínima o *Minimum Bounding Rectangles* (MBR) agrupados jerárquicamente y que pueden solaparse entre ellos o no. El árbol se mantiene balanceado dividiendo aquellos nodos que tienen un número de descendientes por encima del umbral de carga máxima y

combinando aquellos otros que tienen un número de descendientes por debajo del umbral de carga mínima. Cada nodo hoja tiene asociado un MBR que delimita el área del espacio que cubre ese nodo. Además, los nodos internos también almacenan un MBR que delimita el área que cubren todos sus descendientes.

Típicamente es el método preferido a la hora de indexar IG. Los objetos espaciales (polígonos, líneas y puntos) son agrupados y asignados a un MBR dentro del índice, por lo que permiten consultas espaciales sobre geometrías complejas. La Figura 24 representa ejemplos simples de índices *R-tree* para rectángulos en 2 dimensiones (izquierda) y cubos en tres dimensiones (derecha), estas imágenes han sido extraídas de la Wikipedia[124].

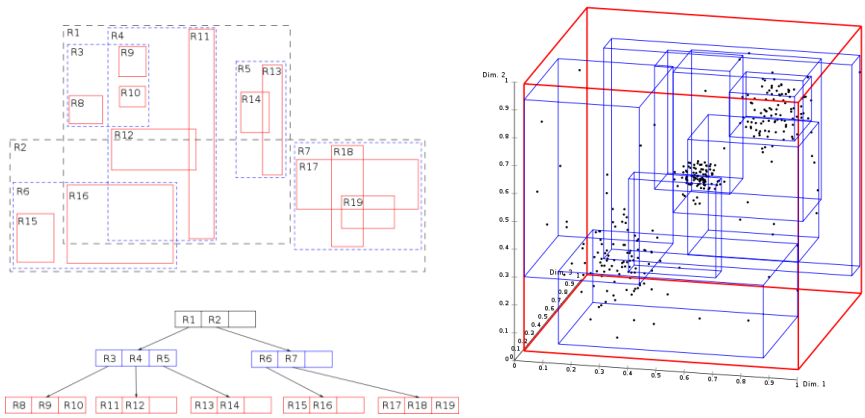

Figura 24: Ejemplos de *R-tree*

A partir de la propuesta original de *R-tree* han aparecido muchas variaciones que pretenden mejorar su eficiencia en aplicaciones específicas, puede verse un resumen de ellas en [140].

En base a pruebas realizadas [127] [48] se considera que los índices *R-tree* resultan más apropiados por su tiempo de respuesta más rápido y por permitir la indexación de geometrías complejas con la consiguiente reducción de falsos positivos.

[124] http://en.wikipedia.org/wiki/R-tree

3.2 Publicación en Servicios de Catálogo

Como hemos visto en la Sección 3.1.3, los servicios de catálogo son actualmente el método más utilizado para la publicación de contenido geográfico en el campo de la IG. A continuación se propone una metodología para publicar recursos georreferenciados en servicios de catálogo de forma integrada en el flujo de trabajo y se presentan dos casos de estudio reales en los que se han implementado sendas soluciones para la publicación de metadatos en catálogos.

3.2.1 Metodología Propuesta

Actualmente la creación y publicación de recursos continúa existiendo en paralelo en lugar de coexistir de una forma más integrada. Se es consciente de la importancia de generar metadatos que describan el contenido de datos y servicios, pero no suele existir un puente entre los recursos y los catálogos que facilitan su evaluación y acceso. Salvo raras excepciones implementadas de una forma un tanto *ad hoc*, es difícil encontrar servicios de catálogo que faciliten la descripción de un conjunto de datos y que a continuación ofrezcan el acceso al servicio de mapas (WMS) que permita la visualización de esos datos on-line. O a la inversa, tampoco es habitual que desde un cliente de visualización de mapas se pueda consultar la descripción exacta del origen de datos que se está sirviendo en línea.

Una vez disponemos de una descripción de los recursos, el siguiente paso es la publicación de estos metadatos. Para ello, la metodología que se propone como contribución, pretende que este proceso se realice de forma integrada y automatizada en el flujo de trabajo, de forma que los usuarios reciban asistencia para la publicación automática de los metadatos. Con ello se pretende incrementar la cantidad de metadatos publicados, ya que se reduce drásticamente el esfuerzo necesario para describir correctamente los recursos y publicarlos.

En un primer momento, la metodología para la publicación de metadatos fue implementada con la idea de publicar metadatos en

servicios de catálogo. Por ello, como veremos en las dos siguientes secciones se utilizó el protocolo CSW. Sin embargo, esta metodología también ha sido incorporada en la implementación de GeoCrawler (ver Sección 5.2) donde los recursos recopilados son publicados de forma automática e integrada indexándolos en base a sus descripciones.

3.2.2 Publicación de Metadatos en gvSIG

Este caso de estudio pretende dar una primera aproximación la metodología propuesta en la sección anterior desarrollando un primer prototipo de una extensión para el manejo de metadatos dentro de la aplicación de escritorio gvSIG. La idea que se persigue es vincular la producción de datos con la de metadatos y con la publicación de estos últimos, de manera que un experto creador y proveedor de datos tiene la facilidad de publicar su dato geográfico en una IDE a través de un servidor de mapas y simultáneamente, con los metadatos generados, y posiblemente editados, publicarlos en un Servicio de Catálogo incluyendo la información del servicio de mapas anterior.

En el capítulo anterior se ha propuesto una metodología para la generación de metadatos (ver Sección 2.2.1) y se ha presentado una primera aproximación a ella implementada como un prototipo sobre gvSIG (ver Sección 2.2.2). La publicación de metadatos es el siguiente paso de la metodología propuesta. Una vez que los metadatos se han generado, los usuarios podrán contar con asistencia para la publicación automática de estos metadatos. Podemos dar a los usuarios la posibilidad de publicar estos metadatos automáticamente de forma integrada en el flujo de trabajo. Con suerte, esto conducirá a aumentar la cantidad de metadatos publicados, ya que estamos reduciendo drásticamente el esfuerzo necesario para describir correctamente los recursos (mediante la generación automática de metadatos) y publicarlos.

La Figura 25 muestra la arquitectura general del prototipo de generación de metadatos en gvSIG presentado en el capítulo anterior extendida con los componentes necesarios para la publicación de los metadatos en un servidor de catálogo.

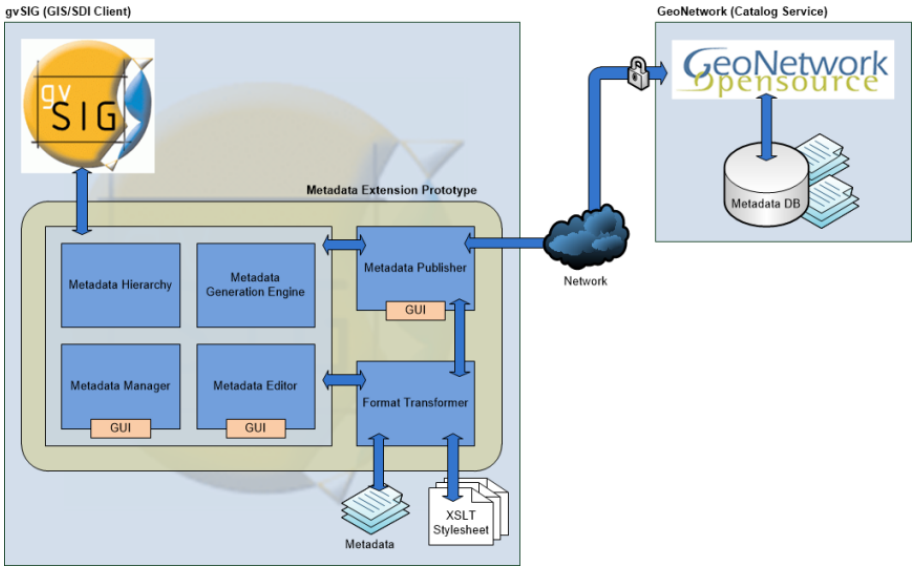

Figura 25: Arquitectura del prototipo de gvSIG (extendida)

Además de la funcionalidad para generar metadatos de forma automática y visualizarlos, editarlos y validarlos en base a los diferentes estándares soportados que se vio en el capítulo anterior, el prototipo permitirá al usuario publicar sus metadatos en un servidor de catálogo. Como servidor de catálogo se eligió *GeoNetwork* por ser la implementación más madura y popular de este tipo de aplicaciones.

Cuando el usuario requiere publicar los metadatos, el gestor de metadatos (*Metadata Manager*) obtendrá y validará los metadatos existentes para los datos. Dicha validación, realizada a través del módulo *Format Transformer*, consiste en confirmar que los metadatos que se quieren publicar contienen al menos el subconjunto de metadatos obligatorios según el estándar seleccionado. En el caso de que la validación falle, el usuario podrá utilizar el editor interno de metadatos para completar los metadatos (*Metadata Editor*). Una vez el conjunto obligatorio de metadatos esté completo, el usuario podrá mediante el módulo de publicación (*Metadata Publisher*) publicar su conjunto de datos en un servicio de catálogo disponible.

En este caso, la publicación de metadatos significa publicar los metadatos en un servicio de catálogo. Para ello usaremos el estándar CSW, específicamente su perfil transaccional CSW-T, que nos permite la publicación y búsqueda de metadatos de datos geoespaciales y servicios.

El módulo encargado de la publicación de los metadatos (*Metadata Publisher*) podrá ser invocado tanto desde el visor/editor de metadatos como directamente desde una opción en el menú, de forma que a través de un asistente el usuario puede elegir la capa actual para ser publicada o bien elegir un fichero de metadatos existente en su sistema de ficheros.

El asistente guiará al usuario para realizar la conexión al servicio de catálogo elegido, realizando una validación de usuario para comprobar que se tienen suficientes permisos sobre el servicio de catálogos. Tras la elección del estándar correspondiente se publicarán los metadatos, obteniendo el estado de la petición e informando al usuario del éxito o fracaso de dicha publicación. En la Figura 26 se puede observar una captura de pantalla de este asistente de publicación tras haber finalizado con éxito una operación de publicación.

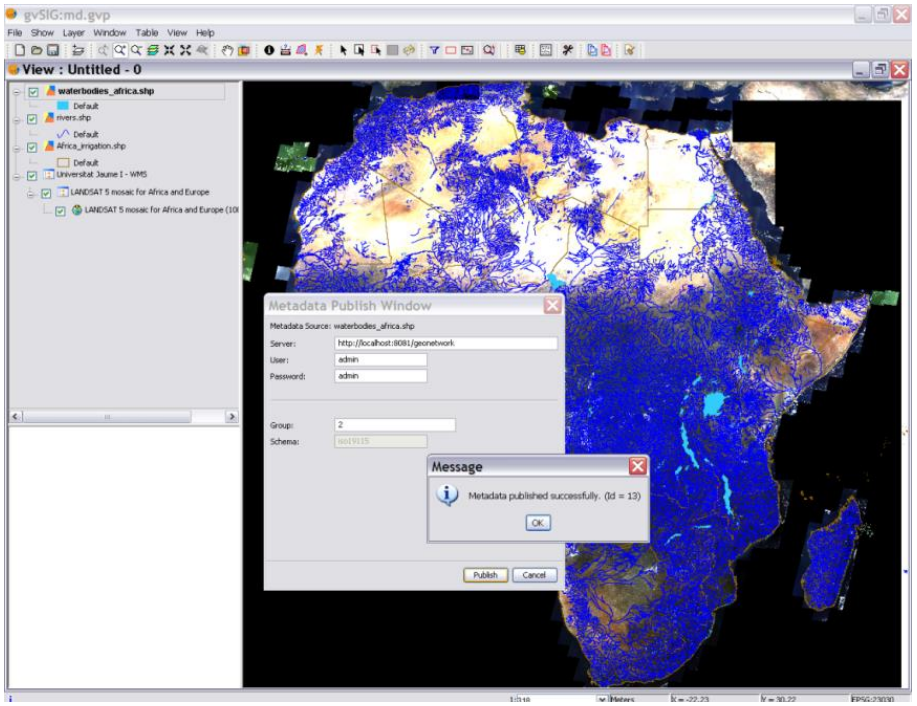

Figura 26: Asistente de publicación de gvSIG

Para comprender mejor el funcionamiento y la utilidad de este prototipo veamos un caso de uso típico. Un técnico usando gvSIG ha combinado los datos geoespaciales básicos, incluyendo datos sobre el terreno, como la pendiente y aspecto, con datos de vegetación para

crear un mapa de riesgos de incendio en un bosque. Suponiendo que tiene permiso para compartir este nuevo recurso, realizará el proceso de publicar el mapa de riesgos en un servidor de mapas, y también le gustaría (o se le exige) publicar su descripción en un servicio de catálogo, como el que se encuentra disponible actualmente en el geoportal de INSPIRE[125].

En nuestro caso de uso el recurso resultante, el mapa de riesgo, tiene asociado un objeto de metadatos que será creado por el prototipo. El último paso en el *workflow* es publicar los metadatos en un servicio de catálogo, entonces, tras comprobar la validez de los metadatos en base al estándar seleccionado, se llevará a cabo la interacción pertinente con el servidor para establecer la conexión y publicar los metadatos.

De esta forma, utilizando esta solución integrada, el usuario puede cerrar el ciclo de vida de los metadatos [10] dentro de la misma aplicación. Puede crear, modificar y publicar los metadatos utilizando el prototipo, y más tarde descubrir y recuperar los metadatos utilizando el cliente de catálogo integrado en gvSIG que, además, permite recuperar los recursos vinculados del Servicio de Catálogo.

3.2.3 Publicación de Metadatos en ENVISION

ENVISION ofrece descubrimiento semántico a través de catálogos, por ello, una vez los expertos han anotado semánticamente los recursos, debe ofrecer mecanismos para publicar tanto los propios recursos como las anotaciones semánticas para que posteriormente otros usuarios puedan encontrarlos. En este sentido, siguiendo la metodología propuesta en la Sección 3.2.1, se ha implementado dentro de la plataforma del proyecto un nuevo componente que, de forma muy resumida, inicialmente obtiene los metadatos necesarios de los servicios Geoespaciales. A continuación crea una descripción de estos recursos basada en el estándar ISO 19119:2005[126] (*Geographic information - Services*) y añade las anotaciones semánticas disponibles. Finalmente publica esta descripción en un catálogo *GeoNetwork* mediante un cliente CSW con su perfil

[125] http://inspire-geoportal.ec.europa.eu
[126] http://www.iso.org/iso/catalogue_detail.htm?csnumber=39890

transaccional CSW-T. De esta forma tanto el recurso como sus anotaciones semánticas quedan disponibles para que los usuarios puedan encontrarlos y utilizarlos de forma correcta. Esta contribución proporciona una solución que habilita la búsqueda semántica de los recursos publicados en el catálogo mejorando su descubrimiento.

Las descripciones semánticas de servicios OGC están basadas en la información proporcionada por su operación *GetCapabilities*, un documento estándar que cada servicio debe proporcionar, y que contiene metadatos describiéndose a sí mismo.

Sin embargo, los servicios de catálogo CSW no pueden almacenar estas descripciones directamente. Esto es debido a que el catálogo, como se define en el estándar, tiene un uso más amplio, pues debe permitir almacenar descripciones no solo de servicios si no de cualquier tipo de recurso geoespacial. Las descripciones almacenadas en el catálogo en forma de metadatos deben cumplir con alguno de los estándares soportados por el catálogo, en nuestro caso el estándar ISO 19119:2005 diseñado para describir servicios geoespaciales.

La publicación de servicios en el catálogo se realiza usando estas descripciones, por lo que para integrar completamente nuestra aproximación semántica, los metadatos en formato ISO deben referenciar las descripciones anotadas semánticamente, mientras se preserva la compatibilidad con el estándar. Esto se consigue añadiendo un elemento *onlineResource* a la descripción del servicio.

Estos metadatos deben ser creados para cada servicio que se quiera publicar en el catálogo. Existen diferentes implementaciones para transformar los documentos *GetCapabilities* en metadatos ISO, pero están basados en el criterio del desarrollador y no hay una forma estandarizada de hacerlo. Además, las anotaciones semánticas se perderían en esta transformación dado que no existe una relación directa con las propiedades de ISO.

Por ello, se dotó al proyecto ENVISION de un conjunto de componentes que realizan estas transformaciones de forma automática, de forma que los usuarios finales no necesitan crear estos metadatos ya que se generan de forma automática. Estos componentes forman parte de la infraestructura de gestión de recursos (*Resource Management*), cuyo propósito principal en lo

referente al descubrimiento es permitir a los usuarios publicar servicios anotados semánticamente.

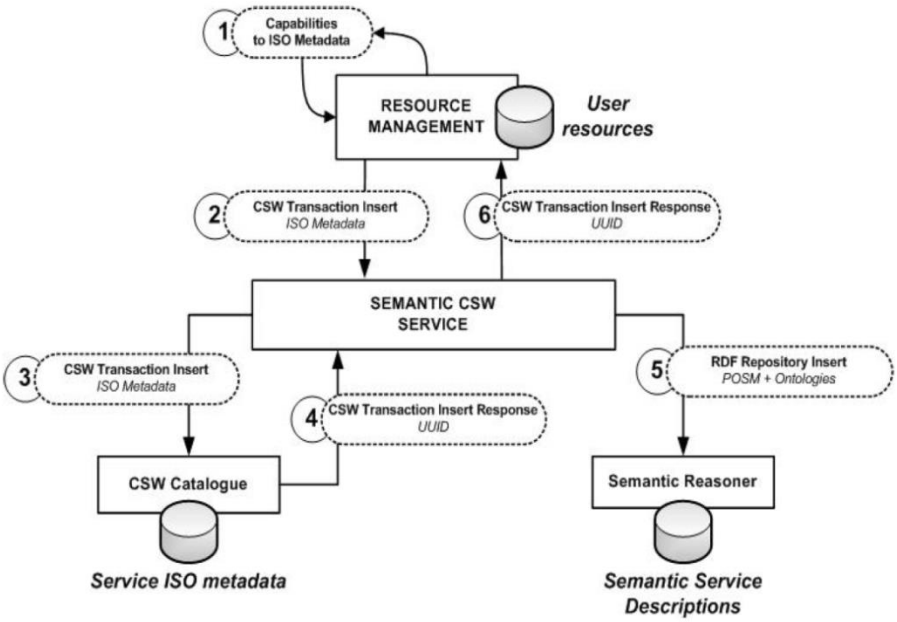

Figura 27: *Workflow* de publicación en ENVISION

El proceso de publicación de servicios OGC anotados semánticamente presentado como contribución e ilustrado por la Figura 27, se realiza de la siguiente forma:

1. La infraestructura de gestión de recursos (*Resource Management*) transforma el documento *GetCapabilities* anotado semánticamente del servicio en sus correspondientes metadatos ISO para ser publicados.
2. La infraestructura de gestión de recursos (*Resource Management*) envía una petición CSW-T *Insert* con los metadatos ISO al servicio CSW semántico (*Semantic CSW Service*).
3. El *Semantic CSW Service* en primer lugar redirige la transacción al servicio CSW subyacente (*CSW Catalogue*) usando el mismo mensaje, actuando como *proxy*[127].

[127] Un *proxy* es un programa o dispositivo que realiza una acción en representación de otro

4. Si la publicación en el catálogo se realiza correctamente, se recibirá un UUID que identifica al nuevo recurso añadido.
5. A continuación, las anotaciones semánticas son extraídas siguiendo los enlaces presentes en los metadatos. Todos los modelos y ontologías son recopilados y almacenados en la base de conocimiento del razonador semántico (*Semantic Reasoner*), de forma que puedan ser usados en la fase de descubrimiento.
6. El UUID que identifica al nuevo registro es retornado como parte de la respuesta CSW estándar.

En este caso de estudio se presenta una aproximación para mejorar el descubrimiento de servicios geoespaciales basada en la publicación en servicios de catálogo de descripciones anotadas semánticamente. Al añadir semántica a las descripciones de los recursos, estos pueden ser descubiertos en base a consultas sobre el significado real del contenido del servicio. Basando nuestro enfoque en los estándares OGC, garantizamos la interoperabilidad y la integración con las soluciones ya existentes.

3.3 Indexación Combinada

En la actualidad existen una gran cantidad de herramientas basadas en web y servicios diseñados para crear, modificar, compartir y visualizar recursos georreferenciados. Estas herramientas junto con la proliferación en el uso y disponibilidad de dispositivos de posicionamiento, como receptores GPS, crea un escenario ideal para los nuevos usuarios. Los usuarios no expertos ahora pueden crear y compartir contenido geográfico, una tarea reservada anteriormente a los profesionales. A pesar de que aún quedan algunas cuestiones por resolver, como la calidad de los datos, el contenido heterogéneo generado por los usuarios aparece en grandes cantidades y rápidamente. Estos factores limitan el uso de catálogos como soluciones eficaces para gestionar la continua proliferación de nuevos recursos.

Por ello, en la Sección 3.1.4 se ha realizado un análisis de distintos tipos de indexación sobre metadatos, así como la integración de estos índices dentro de un sistema de recuperación de información. Concretamente, se analizan diferentes alternativas tanto para la

indexación textual de metadatos como para su indexación espacial y en la presente sección se explora la posibilidad de combinarlos.

En nuestro contexto, para poder satisfacer las necesidades de información de los usuarios de sistemas que contienen recursos georreferenciados, las estructuras de indexación desarrolladas deben permitir la resolución de consultas que tengan tanto una componente textual como una componente espacial. En la Sección 3.1.4 se han analizado algunos de los métodos de indexación textual más relevantes y algunas implementaciones que nos ofrecen esa funcionalidad. Por otra parte, en la misma sección se han revisado las estructuras de indexación espacial más populares a la hora de indexar IG. En esta sección, como contribución, veremos cómo combinar estructuras de ambos tipos, permitiendo así resolver consultas mixtas.

En los últimos años han aparecido propuestas de estructuras de indexación [215] [42] que se basan en la combinación de un índice invertido con una estructura de *grid* [167]. La estructura *grid* es uno de los índices espaciales más sencillos y es el precursor de los índices *Quadtree*. Otras propuestas combinan los índices invertidos con los índices *R-tree* [236] [42]. Finalmente, otras propuestas combinan índices invertidos con los índices *Quadtree* [133]. Además de la complejidad de mantener ambos índices, todos los trabajos, en base a sus experimentos, concluyen que mantener por separado los índices es menos eficiente, lo que implica mayores tiempos de respuesta.

Por otra parte, al combinar búsquedas sobre índices textuales y espaciales se plantea el problema de cómo mezclar y ordenar los resultados provenientes de diferentes orígenes. Sería necesario implementar un mecanismo que aplicara un factor común a la relevancia de los resultados de cada una de las listas de manera que fueran comparables y ordenables. Por lo que utilizar un modelo que combine ambos timos de índices implica una complejidad lógica añadida que resultará costosa en cuanto a tiempo, lo que nuevamente implica mayores tiempos de respuesta.

Dados los problemas al combinar estructuras de indexación textual y espacial, se plantea la posibilidad de integrar los índices espaciales sobre los índices textuales. Además, es necesario indexar todos los metadatos que se puedan generar, por lo que elegir como base la indexación textual tiene sentido. Por ello, en base al algoritmo utilizado

por la aplicación de catálogo *GeoNetwork*, se ha trabajado en un algoritmo de resolución de consultas espaciales que se puede incorporar al motor de indexación textual *Lucene*, permitiendo de este modo la resolución conjunta de operaciones espaciales y textuales.

El algoritmo que se plantea en este apartado se basa en la comparación entre los puntos de los extremos de dos MBRs, uno correspondiente al recurso indexado (MBR_1) y el otro al proporcionado en la consulta (MBR_2). Para comparar otro tipo de formas geométricas (polígonos, líneas o puntos) será suficiente con calcular sus MBRs para poder aplicar el algoritmo. La Figura 28 muestra un ejemplo de MBR con sus correspondientes coordenadas.

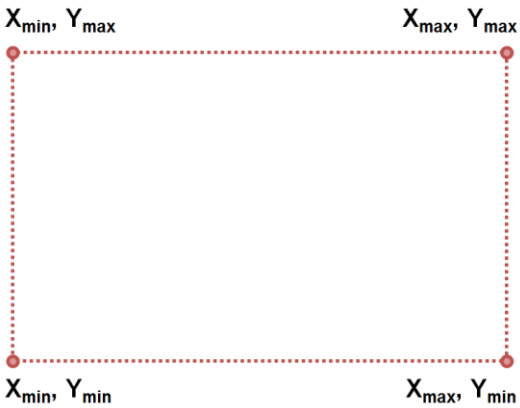

Figura 28: Ejemplo de MBR

El objetivo del algoritmo será devolver todos los recursos cuyo MBR_1 intersecte con el MBR_2. Se entenderá que un MBR_1 intersecta con un MBR_2 si cualquiera de sus límites se cortan o si MBR_1 está incluido o cubre a MBR_2. Resulta más fácil de entender el caso opuesto, podemos afirmar que MBR_1 **no** intersecta con MBR_2 si el primero se encuentra completamente fuera de los límites del segundo. Para ello basta con comprobar que:

$$[MBR_1(X_{min}) > MBR_2(X_{max})] \text{ OR } [MBR_1(X_{max}) < MBR_2(X_{min})] \text{ OR}$$
$$[MBR_1(Y_{min}) > MBR_2(Y_{max})] \text{ OR } [MBR_1(Y_{max}) < MBR_2(Y_{min})]$$

Volviendo a la indexación que es el tema que nos atañe en este capítulo, básicamente, en la fase de indexación deben obtenerse los valores X_{max}, X_{min}, Y_{max} e Y_{min} correspondientes al MBR que contiene al recurso y sobre los que posteriormente se realizarán las consultas, e

incluirlos en el índice. Debemos tener en cuenta que los valores de las coordenadas estén representados mediante un sistema de referencia común y aplicable a todo el globo terráqueo, como WGS84[128].

Esta solución supone un tiempo de indexación menor dado que no penaliza un nuevo tiempo de indexación ya que la indexación espacial se realiza integrada con la textual. Por otra parte, para consultas puramente espaciales la solución *R-tree* es más rápida cuando el número de recursos indexados es pequeño, sin embargo, conforme crece entra en juego el tamaño de la lista de resultados y la nueva solución resulta más rápida. Finalmente, para consultas mixtas la nueva solución resulta más rápida en todos los casos, debido a la complejidad de las soluciones combinadas que hemos comentado anteriormente.

[128] http://es.wikipedia.org/wiki/WGS84

4. Descubrimiento

Según la RAE[129] el descubrimiento es el *hallazgo, encuentro, manifestación de lo que estaba oculto o secreto o era desconocido*. Para los propósitos de este trabajo podemos resumirlo en encontrar y manifestar aquello que era desconocido. Siendo un poco más específicos, podemos decir que el descubrimiento implica encontrar los recursos (que eran desconocidos) que resultan interesantes de acuerdo a nuestras necesidades de información.

Anteriormente hemos visto cómo podemos describir los recursos (Capítulo 2) y seguidamente hemos visto cómo ponerlos a disposición de los usuarios (Capítulo 3), sin embargo la finalidad de describir y publicar los recursos no es otra que permitir y facilitar su descubrimiento, nuestro objetivo final. El descubrimiento o recuperación de información es un área de investigación muy amplia, por lo que nuevamente, con el fin de acotar nuestro alcance, nos centraremos en el descubrimiento de información geoespacial o recursos georreferenciados.

Aparte de ofrecer métodos de descubrimiento a los usuarios, cualquier sistema de información necesitará previamente descubrir los recursos que incorporará, por ello se ha incluido en este capítulo la fase de recopilación del *workflow* general. Por otra parte, entendemos que la visualización de los recursos, junto con interfaces gráficas de búsqueda, supone un factor fundamental para su descubrimiento por

[129] http://lema.rae.es/drae/?val=descubrimiento

lo que también es considerada en este capítulo. La Figura 29 representa la estructura del capítulo y las partes que cubre dentro del *workflow* general.

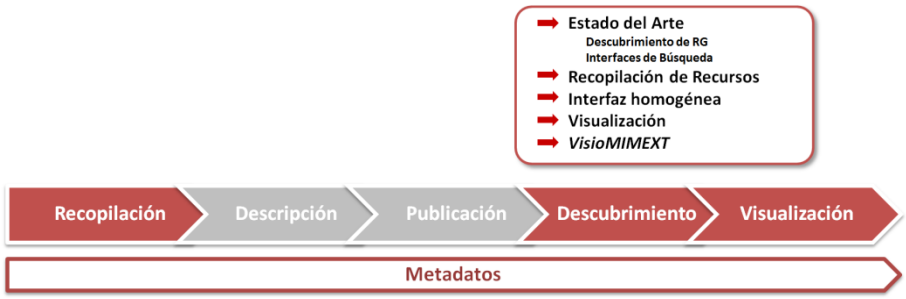

Figura 29: Visión general del Capítulo 4

En este capítulo, en primer lugar se revisan diferentes opciones actuales para el descubrimiento de recursos georreferenciados prestando especial atención a la funcionalidad que ofrecen los motores de búsqueda. De esta forma se responde a las preguntas: *¿Cuál es el proceso seguido para descubrir recursos?*, *¿Cómo se realiza el descubrimiento de recursos georreferenciados?*, *¿Qué papel juegan los motores de búsqueda?* y *¿Cómo influye la calidad de los metadatos en el descubrimiento?*

En segundo lugar, se plantea la pregunta: *¿Cómo descubre o recopila el sistema los recursos que incorpora?* Desde el punto de vista del sistema, existen dos opciones para recopilar los recursos: o los proporciona el usuario directamente al sistema, representando la forma en la que los usuarios publican sus recursos en los catálogos; o se automatiza su recolección, representando la forma en la que los motores de búsqueda recopilan los recursos mediante *crawlers*.

A continuación, se pretende responder a las preguntas: *¿Cómo expresar las necesidades de información?*, *¿Cómo se formulan consultas con un contexto geográfico?* y *¿Cómo realizar consultas sobre diferentes servicios de forma homogénea?* Para ello, en la Sección 4.1.2 se explora cómo los usuarios pueden expresar sus consultas a través de diferentes interfaces de búsqueda que ofrecen los servicios de descubrimiento. Empezando por interfaces de búsqueda específicas del contexto geográfico y, posteriormente, presentando opciones generales pero que permiten realizar consultas filtrando los resultados en base a un contexto espacial.

Posteriormente, en la Sección 4.3 como contribución se propone una interfaz de búsqueda que permitirá que los usuarios realicen consultas espaciales sobre diferentes redes sociales y servicios de una manera homogénea.

Seguidamente, como respuesta a las preguntas: *¿Qué papel juega la visualización en el descubrimiento?* y *¿Cómo visualizar recursos georreferenciados?* se exploran diferentes aspectos sobre la visualización de datos en el contexto científico, prestando especial atención a la IG. La visualización juega un papel importante en el descubrimiento de recursos, ya que nos permite examinarlos y comprenderlos de una forma más sencilla. De este modo, dada la naturaleza de la IG, se presentan los globos virtuales como una útil herramienta que permite la visualización e integración de datos para cualquier tipo de usuario.

Finalmente, se quiere explorar *¿Cómo encaja todo esto?* y *¿Resultan útiles las aplicaciones basadas en un globo virtual?* Por ello, como prueba de concepto, se presenta la aplicación *VisioMIMEXT* que integra un conjunto de componentes que permiten el descubrimiento de diferentes tipos de recursos georreferenciados a través de interfaces de búsqueda *OpenSearch* y la visualización de los resultados sobre un globo virtual, gracias a la extensión MIMEXT que facilita la georreferenciación de los recursos.

4.1 Estado del Arte

Cuando alguien va a una librería buscando un buen libro de suspense, probablemente empezará buscando en la sección específica de este tipo de libros. Entonces, empezará a revisar los títulos de los libros buscando uno que encaje con sus preferencias en base a, por ejemplo, en el periodo histórico o lugar dónde tendrá lugar la historia que narra. Es probable que a primera vista la información proporcionada por el título no sea suficiente, por lo que será necesario leer el resumen de la parte trasera del libro que ofrece más información sobre el libro y en ocasiones sobre el autor o la crítica. Si esto no fuera suficiente, quizá el potencial comprador pueda abrir el libro y mirar los detalles del editor o el año de publicación y, en última

instancia, empezar a leer el libro para ver si es el que mejor se adapta sus gustos.

Este ejemplo cotidiano puede servir para ilustrar los pasos seguidos por alguien que intenta descubrir algún tipo de recurso. El primer paso es decidir dónde empezar a buscar. En el ejemplo el usuario está buscando un libro de suspense, por lo que es lógico empezar a buscar en la sección correspondiente de la librería. Su equivalente en el mundo de la informática probablemente sería acceder a un motor de búsqueda, como *Google*, *Yahoo* o *Bing*, o servicios de búsqueda más especializados dependiendo de las preferencias del usuario y sus conocimientos. El segundo paso del proceso sería comprobar la información de los libros en base a ciertos criterios, lo que implica el uso de información sobre su contenido. En el caso de los libros resulta bastante fácil ya que podemos simplemente leer la información, pero este no es siempre el caso en el entorno de recursos heterogéneos en el que nos movemos, lleno de contenido no estructurado como imágenes, audio o video. Dado que la mayoría de interfaces de búsqueda actuales están basadas en rellenar ciertos criterios de búsqueda, será necesario describir estos recursos adecuadamente para poder encontrarlos. Como hemos visto en el Capítulo 2, las descripciones de los recursos, en forma de metadatos, son el elemento clave en el descubrimiento de cualquier tipo de recurso, especialmente en entornos distribuidos y heterogéneos.

Por otra parte, en el Capítulo 3 hemos definido el descubrimiento de información como la actividad consistente en obtener recursos relevantes a una necesidad de información a partir de una colección de recursos disponibles. En dicho capítulo, como un paso previo al proceso de recuperación de información, hemos explorado diferentes opciones para publicar los recursos, quedando estos disponibles para su descubrimiento.

El proceso de descubrimiento ilustrado por el ejemplo anterior puede ser fácilmente extrapolado al proceso que muestra la parte derecha de la Figura 21, donde vemos que el usuario deberá expresar sus necesidades de información (*user need*) a través de la interfaz de usuario (*User Interface*) en forma de consulta (*query*), que será procesada (*Searching*) para obtener unos recursos como resultado (*retrieved docs*).

De este modo, la informática resulta un campo imprescindible para el desarrollo de la sociedad al cubrir la necesidad de gestionar la ingente cantidad de información que manejamos hoy en día. A lo largo de los años se han propuesto gran cantidad de sistemas, arquitecturas, estructuras y otras componentes con el objetivo de permitir el descubrimiento y el acceso eficiente a los recursos almacenados en enormes repositorios. El interés en el descubrimiento de recursos ha sufrido un incremento espectacular motivado por el crecimiento de Internet y la necesidad de realizar búsquedas en la web.

4.1.1 Descubrimiento de Recursos Georreferenciados

Centrándonos en el descubrimiento de recursos georreferenciados, tradicionalmente el descubrimiento de IG se ha realizado desde una perspectiva local y de forma centralizada. Sin embargo, en la última década, la noción de IDE [148] ha aparecido como una red de nodos interconectados para construir infraestructuras de información espacial a diferentes escalas y niveles (local, regional, nacional, Europeo e incluso global) [113]. Como hemos visto en el capítulo anterior, el descubrimiento de IG a través de esta inmensa red de nodos IDE ha sido delegado al uso de servicios de catálogo. Estos servicios especializados están basados en el uso de metadatos que describen los recursos que registran y ofrecen interfaces de búsqueda efectivas a los usuarios o las aplicaciones cliente.

Sin embargo, la apertura, sencillez y la facilidad de compartir que ofrecen las nuevas herramientas y servicios de la Web 2.0, también facilitan la creación de contenidos georreferenciados lo que resulta en un espectacular aumento de este tipo de recursos. La variedad y la velocidad a la que aumenta la cantidad de recursos, junto con la naturaleza de la mayoría de los creadores, sin profundos conocimientos técnicos, hacen que el uso de catálogos resulte una solución ineficaz para su gestión, debido a sus estrictos requisitos de metadatos y su proceso de publicación dirigido por el usuario. Por esta razón, son necesarios nuevos métodos para descubrir dichos recursos.

Durante la evolución de la Web, su tamaño empezó a alcanzar una dimensión que resultaba problemática a la hora de buscar la información contenida en directorios que incluían vínculos a los recursos. En ese caso, los motores de búsqueda aparecieron con resultados exitosos convirtiéndose en una herramienta esencial [4]. Quizás estas aplicaciones pueden representar de nuevo una solución para el descubrimiento de recursos georreferenciados en la red actual.

Motores de búsqueda

Los motores de búsqueda (*Search Engines*) son el mecanismo de descubrimiento más usado para búsquedas de propósito general. De hecho, si consultamos alguna herramienta de análisis de tráfico Web como *Alexa*[130] y examinamos los sitios con más visitas nos daremos cuenta del impacto y relevancia de este tipo de servicios. Entre los diez sitios más visitados encontramos cinco motores de búsqueda: *Google*, *Yahoo!*, *Baidu*[131], *Bing* y *SOSO* a través de *QQ*[132].

Los motores de búsqueda son servicios que aparecieron hace años para resolver el problema de encontrar recursos e información en una Web en expansión. Cada vez tienen un papel más importante ya que la cantidad de recursos disponibles aumenta en una proporción tal que hace la tarea de encontrar información específica casi imposible sin el uso de estas herramientas. Los motores de búsqueda realizan tres tareas principales [52]: recopilar los recursos, que veremos en más detalle en la siguiente sección; indexarlos, tal y como hemos visto en el capítulo anterior; y ofrecer servicios de búsqueda. Básicamente toda la información relevante se extrae de los recursos y se almacena convenientemente en una base de datos o índice para su rápida recuperación. Finalmente, la tarea de recuperación o búsqueda se realiza a través de interfaces específicas, que, por lo general, permiten la inserción de texto o palabras clave que se buscará en el índice.

En la Sección 3.1.4 se han presentado varias implementaciones de motores de búsqueda de propósito general y se han analizado sus capacidades de indexación textual basadas en los índices invertidos. En esta sección revisaremos las funcionalidades que ofrecen algunos motores de búsqueda sobre recursos georreferenciados.

[130] http://www.alexa.com/topsites
[131] http://www.baidu.com
[132] http://www.qq.com

A través de los años, con la popularización de nuevos tipos de datos y formatos, estas herramientas se han adaptado para buscar e indexar este tipo de recursos. Y, más recientemente, algunos motores de búsqueda como *Google* han incorporado capacidades para manejar IG expresada en KML o KMZ[133], permitiendo realizar búsquedas específicas sobre este tipo de recursos [52].

Por otra parte, la última versión de *Lucene* (4.1.0) publicada mientras se escribía esta tesis, incorpora de forma experimental un nuevo módulo espacial[134] (*Spatial Module*) que implementa algunas capacidades para la indexación y búsqueda basada en formas geométricas. Según indican, el rendimiento y las características dependerán de la estrategia que se elija. El módulo incorpora diferentes estrategias como un vector de puntos (*PointVectorStrategy*), que indexa únicamente puntos y permite consultas basadas en un rectángulo o un círculo, o un árbol de prefijos (*PrefixTreeStrategy*) que permite indexar cualquier forma y permite consultas basadas en la intersección.

A parte de estos, los intentos para proporcionar un contexto espacial a los motores de búsqueda son relativamente escasos [209]. Una de las razones puede ser la gran diversidad y heterogeneidad de los recursos georreferenciados. Algunas excepciones, sin embargo, demuestran que los motores de búsqueda que permiten consultas con una componente espacial, combinados con herramientas de visualización conducen a aplicaciones completas y útiles. *Alkemis*[135], por ejemplo mezcla los contenidos con información de otros dominios como el tráfico y el clima para ofrecer servicios personalizados basados en la localización [25].

Recuperación de Información Geográfica

La reciente convergencia de la recuperación de información con los sistemas de información geográfica ha dado lugar a la Recuperación de Información Geográfica o *Geographic Information Retrieval* (GIR) [184], un nuevo campo de investigación que ha despertado gran interés en la comunidad científica, debido a que un alto porcentaje de la información tratada por instituciones y empresas públicas o privadas

[133] http://support.google.com/webmasters/bin/answer.py?hl=es&answer=35287
[134] http://lucene.apache.org/core/4_1_0/spatial
[135] http://local.alkemis.com

tiene, en alguna medida, relación con datos espaciales. Y, en consecuencia, la toma de decisiones depende en gran parte de la calidad, exactitud y actualidad de esta información espacial y, por ende, de la calidad de los sistemas de informáticos que tratan esa IG.

El objetivo de GIR es la recuperación de información que incluya algún tipo de referencia espacial. Teniendo en cuenta la cantidad de recursos que contienen algún tipo de referencia espacial, las referencias geográficas pueden jugar un papel importante para la recuperación de información. Por ejemplo al realizar una consulta del tipo: *"noticias sobre disturbios cerca de Madrid"*. Para ello, GIR propone tareas como: la traducción de localizaciones, ya que muchos documentos contienen georreferencias expresadas en varios idiomas que pueden o no ser el mismo que el lenguaje de consulta; resolución de ambigüedad en georreferencias, por ejemplo *"Juan Madrid"* es un escritor español y no hace referencia a la ciudad, o *Islas Canarias* e *I. Canarias* hacen referencia al mismo lugar; y resolución de ambigüedades espaciales, por ejemplo la ciudad de *Valencia* en España o en Venezuela.

El hecho de que la mayoría de la IG esté encerrada en documentos de texto no estructurados ha dado lugar a muchas investigaciones por parte de diferentes comunidades científicas. De este modo, los proyectos tienen por objetivo analizar periódicos actuales e históricos, medios sociales y, cada vez más, documentos oficiales que se están publicando abiertamente. Normalmente encontramos la IG de forma implícita en el texto no estructurado y no de forma explícita. En paralelo, también se han desarrollado métodos basados en las nociones de la web semántica que intentan encapsular IG y otros datos contenidos en los recursos. De este modo han surgido vínculos entre las dos áreas de investigación, por ejemplo en el desarrollo de ontologías geográficas. Sin embargo, el campo es relativamente joven, y hasta la fecha no se ha demostrado plenamente su potencial. Por ejemplo, los talleres GeoCLEF[136] (*Evaluation of multilingual Geographic Information Retrieval Systems*), que buscan comparar los sistemas GIR con los métodos estándar de recuperación de información para consultas geográficas multilingües, únicamente han mostrado unas ventajas muy limitadas al usar enfoques con contexto espacial.

[136] http://www.uni-hildesheim.de/geoclef

La mayoría del trabajo realizado en el campo de investigación GIR está basado en la extracción de referencias espaciales explícitas a partir de las referencias espaciales implícitas presentes en los recursos analizados e intentar solucionar los problemas derivados de este proceso como la traducción entre diferentes idiomas y la resolución de ambigüedades. En el contexto de esta tesis podemos decir que esto queda fuera de nuestro alcance dado que los recursos considerados en este trabajo disponen (originalmente o tras ser anotados) de referencias espaciales explícitas. Sin embargo GIR es considerada una línea de investigación muy interesante y que podría proporcionar muchos recursos nuevos al sistema, por lo que será explorada en el futuro.

Calidad de metadatos y servicios de búsqueda

El término GIGO es famoso como abreviatura del dicho inglés *"Garbage In, Garbage Out"* (Entra Basura, Sale Basura). Usado normalmente en informática para describir el problema derivado de utilizar datos erróneos como entrada de un sistema, cuya consecuencia es que la salida será inevitablemente inexacta o incorrecta. Es decir, el uso de información de poca calidad limita el rendimiento de un sistema. En consecuencia, es imprescindible que los sistemas de recuperación cuenten con metadatos de calidad que describan los recursos con fidelidad para que puedan indexarse de forma exacta y que posteriormente puedan descubrirse resultados realmente relevantes.

Según la RAE[137], la calidad es una *propiedad o conjunto de propiedades inherentes a una cosa que permiten apreciarla como igual, mejor o peor que las restantes de su especie*. Por otra parte, según [107] un registro de metadatos de buena calidad se define como *"un registro que es útil en un número de contextos diferentes, siendo útil también respecto a las estrategias de búsqueda y términos que se pueden emplear para localizarlo"*. Otras definiciones son menos ambiciosas y simplemente hablan de adecuación al propósito perseguido [99]. Siguiendo estos razonamientos, podemos decir que los metadatos serán de una calidad suficiente si describen de forma fiel los recursos y esas descripciones son útiles para su propósito.

[137] http://lema.rae.es/drae/?val=calidad

La evaluación de la calidad de los metadatos implica realizar dos tipos de validaciones, la primera validación se preocupa por la estructura de los elementos de los metadatos y trata de determinar en qué medida los registros de metadatos cumplen con la norma o el estándar empleado, teniendo en cuenta, además, el formato definido para cada elemento. La segunda validación, se relaciona con la fidelidad y completitud con la que los metadatos describen el recurso. En esta validación influyen aspectos subjetivos como la coherencia entre el elemento y la información que contiene, evitar la duplicación de información, omitir contradicciones, la precisión de la información y la homogeneidad en diferentes registros de metadatos y por último, el usar un lenguaje natural que reduzca al mínimo la ambigüedad.

La primera validación resulta sencilla, ya que comprobar que la estructura de un registro de metadatos se ajusta a su respectivo estándar es automatizable gracias a los esquemas que proporcionan los estándares de metadatos. Sin embargo, la segunda validación sobre la calidad de un recurso está basada en la percepción que el usuario tiene del mismo, es una fijación mental del consumidor que asume conformidad con dicho recurso y la capacidad del mismo para satisfacer sus necesidades. Por lo tanto, la calidad es una cuestión subjetiva, influenciada por juicios personales. Debido a estas razones, la comunidad científica reconoce que la evaluación de la calidad de los metadatos no está exenta de dificultades.

En los últimos años se han presentado diferentes trabajos relacionados con la calidad de los metadatos. Estos trabajos abordan el tema desde diversas perspectivas, tratando de cubrir la mayor parte de sus aspectos. Algunos autores [161] [82] [31] proponen llevar a cabo un análisis estadístico sobre una muestra de registros de metadatos de diferentes repositorios y evaluar el uso del estándar. De forma que designan los campos más frecuentemente utilizados y sus valores atribuidos. Aunque no está directamente asociado a la calidad, los índices estadísticos producidos proporcionan una idea de la eficacia de los repositorios examinados. En este sentido, en [96] se presenta un estudio que examina la capacidad de los autores de los recursos para crear metadatos de calidad aceptable mediante la evaluación manual por parte de expertos. Otros autores [67] estudian el tema de la calidad identificando las deficiencias que la degradan y proponen el uso de una herramienta gráfica para visualizar estas deficiencias a nivel de repositorio. En otros trabajos [14] [99] [53] se

aborda el control de la calidad con el fin de satisfacer los requisitos funcionales de la aplicación en la que se utilizan. Para garantizar la calidad, en [105] se discute la contribución de los perfiles de aplicación como medio para exponer y hacer cumplir la calidad de los metadatos.

Con la introducción de *frameworks* genéricos para la evaluación de la calidad de los metadatos se consigue una visión más sistemática y organizada. En [156] se presenta un *framework* para la evaluación de los registros de metadatos mediante un conjunto de 23 criterios de evaluación. En [85] se presenta otro *framework* basado en los conceptos e ideas de la calidad de la información en general, identificando hasta 32 parámetros. Por otra parte, [30] profundiza en siete características sobre la calidad de los metadatos: integridad, precisión, procedencia, cumplimiento de las expectativas, consistencia lógica y coherencia, oportunidad y accesibilidad. En base a estas características, otros trabajos [111] [174] intentan llevar a la práctica la medición de la calidad. Otros autores [54] emplean técnicas de aprendizaje automático para clasificar los recursos, en base a unos indicadores. Del mismo modo, [235] también propone un conjunto indicadores de calidad y estudia la relación entre la calidad de los metadatos y la precisión obtenida en las respuestas de un servicio de catálogo de la comunidad OAI. Siguiendo estas líneas, algunos trabajos [59] [168] [206] también han analizado la calidad de los metadatos y su repercusión en los procesos de recuperación de IG.

La influencia de la calidad de los metadatos en la eficiencia de los sistemas de recuperación de información ha sido establecida en la mayoría de estos trabajos, destacando los riesgos que puede implicar el uso de metadatos de escasa calidad. Desde el punto de vista del descubrimiento, entre los problemas que surgen debido a la baja calidad de los metadatos, destacan la poca exhaustividad (*recall*), la poca precisión y la ambigüedad e inconsistencia de los resultados de las búsquedas. Por lo tanto podemos concluir que la calidad de los metadatos es un factor muy importante e influyente en el proceso de descubrimiento de los recursos.

4.1.2 Interfaces de Búsqueda

Anteriormente, hemos revisado el importante papel de los motores de búsqueda y otros mecanismos de recuperación de información, también hemos revisado algunas de las técnicas que utilizan para la publicación e indexación de los recursos. Sin embargo, no debemos olvidar que el objetivo de estos sistemas es buscar sobre los recursos que recopilan aquellos que son más adecuados de acuerdo a las consultas de los usuarios.

La forma de buscar más común es emitir una consulta inicial, revisar una lista de respuestas sugeridas y seguir los enlaces a recursos específicos. Si esta primera aproximación no conduce al descubrimiento de recursos útiles, el usuario refina o modifica la consulta mediante el uso de funciones avanzadas de consulta como la restricción del dominio de búsqueda o forzar la inclusión u omisión de términos de búsqueda específicos. En este modelo de búsqueda, una necesidad de información está representada por una consulta, y el usuario puede emitir varias consultas para cubrir su necesidad de información. De forma que los usuarios esperan ser capaces de recuperar recursos relevantes de acuerdo con los términos de las consultas que formulan.

Con motores de búsqueda típicos, dado que el contenido que albergan es en su mayoría texto no estructurado, la gran mayoría de las necesidades de información se presentan como consultas sobre un conjunto de palabras clave [240]. Algunas consultas de este tipo son en realidad expresiones tales como nombres propios y otras contienen frases completas marcadas de forma explícita, entre comillas. Otro método común es usar operadores booleanos como AND, OR o NOT para restringir las respuestas, exigiendo que todos, alguno o ninguno de los términos de la consulta estén presentes en una respuesta.

Sin embargo, en nuestro contexto, gracias a los metadatos disponemos de atributos estructurados útiles para categorizar los recursos que describen y filtrar los resultados. Uno de estos atributos es la localización del recurso, que nos permitirá filtrar los resultados en base a un contexto espacial.

Esta sección explora cómo los usuarios pueden expresar sus necesidades de información en forma de consultas a través de las interfaces de búsqueda que ofrecen los servicios de descubrimiento.

Catalogue Service for the Web

En general, los servicios de descubrimiento de recursos basados en las necesidades del usuario siempre han sido un reto, y tener en cuenta aspectos geoespaciales de esos recursos no facilita la tarea. Por ello, OGC define una arquitectura Web estandarizada para el descubrimiento recursos geoespaciales llamada *Catalogue Service for the Web* (CSW). CSW define de forma abstracta la operación del proceso de descubrimiento, estrechamente unido con un conjunto de estándares de metadatos, especificando sus capacidades de consulta y detallando la forma de interactuar con el sistema.

La especificación CSW [164] permite a diferentes entidades publicar información (metadatos) sobre recursos georreferenciados y establece los procedimientos para consultar o recuperar esta información. El mecanismo de descubrimiento es formalmente definido por (1) el esquema de los metadatos (cómo describir los recursos), (2) el lenguaje de consulta (cómo definir las consultas) y (3) la interfaz de descubrimiento (cómo invocar el proceso). El estándar CSW define estos tres aspectos como sigue:

1. Con el fin de promover la interoperabilidad, como hemos visto en el Capítulo 2, OGC recomienda el uso de estándares de metadatos ya definidos proporcionados por algunos organismos de normalización (por ejemplo: ISO 19115), que definen los esquemas para la representación de IG. Sobre la base de estos esquemas comunes, existe un conjunto predefinido de propiedades principales, a partir de las cuales las consultas pueden ser creadas (por ejemplo: tema, título, resumen, formato, identificador, MBR, etc.).
2. La especificación también define el *OGC_Common Catalogue Query Language* [164] y proporciona un conjunto mínimo de tipos de datos y operaciones de consulta que todo catálogo OGC debe implementar. El lenguaje de consulta soporta de forma predefinida algunas operaciones de filtrado (booleano, correspondencia textual, temporal, operadores geoespaciales, etc.) y proporciona un *framework* para posibles extensiones.

3. Para la interfaz de invocación, se definen un conjunto de operaciones abstractas que dan soporte al descubrimiento. Como vimos en el Capítulo 3, la publicación se define por un conjunto de operaciones incluidas en el perfil transaccional CSW-T (insertar, eliminar y actualizar). Por otra parte, el descubrimiento cubre diferentes operaciones de consulta basadas en servicios que son descritas como estructuras de registros: *DescribeRecord* proporciona información acerca del modelo de los registros; *GetRecords* buscará registros aplicando los filtros especificados y devolverá sus identificadores; y *GetRecordsById* devolverá los registros especificados por su identificador. Más detalles acerca de estas operaciones se pueden encontrar en la especificación [164].

Siguiendo estos principios, las consultas pueden ser formuladas conteniendo palabras clave (texto arbitrario) y/o filtros espaciales mediante el uso de operadores espaciales.

Podemos encontrar disponibles varias implementaciones de los servicios de catálogo OGC CSW. En [207] se han analizado las implementaciones más populares según los requerimientos del proyecto ENVISION.

OpenSearch

OpenSearch[138] se ha convertido rápidamente en un exitoso mecanismo de búsqueda sobre miles de sitios web, servicios y repositorios que están, cada vez más, adaptándose para exponer de una forma estándar y simple sus interfaces de búsqueda. A continuación se describe la interfaz básica de *OpenSearch* y su extensión geográfica [210].

Patrón de consulta basado en palabras clave

OpenSearch proporciona una funcionalidad de búsqueda y de recuperación mínima que cualquier repositorio o servicio debe soportar. La especificación *OpenSearch* describe un patrón de búsqueda básica que encaja muy bien en las interfaces de búsqueda mínimas que han identificado a la mayoría de servicios Web 2.0. Un servicio habilitado para *OpenSearch* expone su interfaz para informar a las aplicaciones cliente, como la herramienta *"búsqueda*

[138] http://www.opensearch.org

personalizada" descrita anteriormente, sobre la forma en que deben emitir las simples consultas HTTP GET ampliándolas con los parámetros de consulta específicos. Como resultado, las respuestas normalmente son codificadas en formatos de datos ligeros tales como *GeoRSS*[139], *Atom* [171] o *KML*.

La interfaz de búsqueda *OpenSearch* sólo tiene un parámetro de consulta obligatorio llamado *"searchTerms"* que permite a las aplicaciones cliente recuperar los recursos que están relacionados con una o más palabras clave. Otros parámetros de consulta, como los relativos a la paginación de los resultados (*"count"*, *"startIndex"*, *"startPage"*), son opcionales. En el lado del servidor, el servicio que implementa la interfaz *OpenSearch*, realiza una búsqueda basada en texto sobre el repositorio de recursos que alberga, normalmente considerando los siguientes descriptores de metadatos para cada recurso objetivo: título, autor, descripción y etiquetas definidas por el usuario. Cabe señalar que la lista de descriptores de metadatos objetivo depende del servicio específico. Por ejemplo, *Twitter*, *Flickr* y *YouTube*[140] pueden compartir algunos descriptores pero seguramente otros son diferentes.

Patrón de consulta espacial

Aunque el patrón de consulta basado en palabras clave funciona bien en muchas situaciones, es posible especificar criterios de búsqueda avanzados a través de extensiones o perfiles especializados. En la comunidad *OpenSearch*, los perfiles de búsqueda específicos se describen mediante la ampliación de las capacidades de básicas de *OpenSearch*. Entre las extensiones disponibles[141], la extensión *Geo* [210] define una lista de parámetros de consulta para habilitar el filtrado de resultados por contexto geográfico. Por lo tanto, las consultas con base espacial son soportadas a través de la combinación adecuada de la interfaz de búsqueda de *OpenSearch* y su extensión *Geo* (*OpenSearch-Geo*).

La extensión *OpenSearch-Geo* está definida sobre la especificación básica de *OpenSearch*, por lo que todos los parámetros de consulta obligatorios y opcionales mencionados anteriormente están también

[139] http://georss.org
[140] http://www.youtube.com
[141] http://www.opensearch.org/Specifications/OpenSearch/Extensions

disponibles. Además de estos parámetros básicos, *OpenSearch-Geo* define algunos parámetros de consulta específicos y opcionales. El parámetro *"box"* filtra los resultados a partir de un área rectangular. Por su parte, los parámetros *"lat"*, *"lon"* y *"radius"* filtran los resultados a partir de un área circular alrededor de un punto. Otra opción es el parámetro *"geometry"* que define un filtro geográfico en base a una geometría arbitraria. Finalmente, el parámetro *"name"* permite filtrar los resultados por el nombre de lugar. Podemos apreciar todos estos parámetros de filtrado espacial de forma gráfica en la Figura 30.

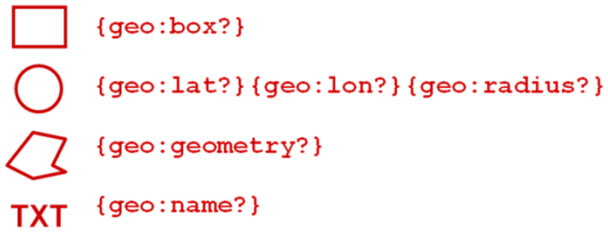

Figura 30: Parámetros de búsqueda de *OpenSearch-Geo*

A pesar de que ni la especificación *OpenSearch* ni su extensión *OpenSearch-Geo* especifican ningún formato obligatorio de respuesta (aunque ambos recomiendan soportar al menos el formato *Atom* [171]), los recursos devueltos como respuesta deben estar georreferenciados para que las aplicaciones cliente puedan manejarlos. Los formatos de respuesta comunes que ofrecen las implementaciones actuales de *OpenSearch-Geo* siguen dos aproximaciones. La primera de ellas consiste en utilizar formatos existentes no específicos, como HTML, *Java Script Object Notation*[142] (JSON) o *Atom*. En este caso, estos formatos son enriquecidos con anotaciones geográficas mediante, por ejemplo, el *Geo Microformat*[143] sobre HTML o la extensión *GeoRSS* sobre *Really Simple Sindication*[144] (RSS) y formatos *Atom*. La segunda aproximación consiste en utilizar formatos con soporte nativos de referencias geográficas como KML. Para proporcionar un enfoque equilibrado, nuestra decisión fue soportar *Atom* para consultas basadas en palabras clave y, para consultas espaciales *Atom* ampliado con

[142] http://www.json.org
[143] http://microformats.org/wiki/geo
[144] http://www.rssboard.org/rss-specification

GeoRSS, KML y su extensión MIMEXT (ver Sección 2.3), como se describe en la Sección 4.3.

4.2 Recopilación de Recursos

Cualquier sistema de información necesitará previamente descubrir los recursos que incorporará. La manera en la que el contenido es publicado para su descubrimiento a través de motores de búsqueda difiere notablemente de cómo se realiza en los catálogos. Desde el punto de vista del sistema, existen dos opciones para recopilar los recursos: o los proporciona el usuario directamente al sistema o se automatiza su recolección.

La primera opción es obvia, los usuarios son los encargados de proporcionar al sistema los recursos que quieren incorporar junto a su descripción a través de las interfaces o métodos específicos que ofrece el sistema para poder ser indexados y posteriormente recuperados. Este es el caso de los antiguos directorios web[145] o de los actuales catálogos de metadatos en el contexto de las IDE nacidos para facilitar la búsqueda de recursos. En estos sistemas, los creadores de contenidos pueden clasificar sus recursos de acuerdo a ciertos criterios como su temática o su localización, es decir, creando e introduciendo manualmente sus metadatos. Esto facilita la accesibilidad a la información al usuario, pues este puede encontrar más fácilmente recursos acordes a sus intereses. Pero este sistema, además de ser susceptible de engaño al clasificar los contenidos de acuerdo a los deseos de sus creadores, no resulta efectivo dado el trabajo de dar de alta los recursos en un gran número de directorios no es gratificante para los creadores de contenidos.

Para aliviar el problema de tener que dar de alta los recursos en diferentes directorios o catálogos iniciativas como OAI han desarrollado técnicas de *harvesting: Open Archives Initiative Protocol for Metadata Harvesting*[146] (OAI-PMH) [130]. Estas técnicas permiten compartir metadatos entre diferentes nodos, de forma que el usuario sólo tendrá que publicar su recurso manualmente en uno de los nodos del sistema. En el contexto de la IG el *harvesting* ha sido incluido en el

[145] http://en.wikipedia.org/wiki/Web_directory
[146] http://www.openarchives.org/pmh

estándar CSW de OGC a través de la operación opcional *Harvest* e implementado en algunos catálogos como *GeoNetwork*[147], de forma que periódicamente los metadatos contenidos en diferentes nodos son sincronizados permitiendo búsquedas centralizadas.

Por otra parte, para automatizar la recopilación de los recursos suelen usarse los llamados *crawlers*, también conocidos como *spiders* o *bots*, son programas que inspeccionan los recursos accesibles en la Web de forma metódica y automatizada [182] [125]. Uno de los usos más frecuentes que se les da consiste en recopilar todos los recursos visitados para su procesado posterior por un motor de búsqueda que los indexa proporcionando un sistema de búsquedas rápido. Los *crawlers* son usados actualmente por todos los buscadores de Internet, incluidos los más famosos como *Google*, *Yahoo!* o *Bing*, para recopilar sus recursos. Podemos encontrar un largo listado de las arquitecturas e implementaciones de *crawlers* más conocidas en su correspondiente entrada de la Wikipedia[148].

Los *crawlers* empiezan visitando una lista de URLs, identifican los enlaces presentes en dichos recursos y los añaden a la lista de URLs a visitar de manera recurrente de acuerdo a un determinado conjunto de reglas. Normalmente, se le proporciona al programa un grupo de direcciones iniciales, el *crawler* accede a estas direcciones, analiza los recursos y busca enlaces a nuevos recursos. Luego descarga estos nuevos recursos, analiza sus enlaces, y así sucesivamente. Sin embargo, algunos *crawlers* también pueden ser configurados para ejecutarse en otros contextos más controlados como la red interna de una organización, un servicio específico o un sistema de directorios completo ya sea local o remoto.

La principal ventaja de la recopilación automatizada de recursos es que no requiere intervención por parte del usuario que, simplemente, debe dejar sus recursos accesibles en el contexto en el que el *crawler* es ejecutado. Por el contrario, los recursos no son inmediatamente publicados como sucede con los catálogos, debido a que debemos esperar a que el *crawler* descubra nuestro recurso. El tiempo necesario dependerá del *crawler* específico y de la cantidad de recursos disponibles en el contexto de ejecución que deben ser

[147] http://www.geonetwork-opensource.org/stable/users/admin/harvesting
[148] http://en.wikipedia.org/wiki/Web_crawler

analizados. Además, dado que los usuarios no tienen el control siempre existe cierta incertidumbre sobre el éxito y la corrección del proceso, un pequeño precio a pagar por la facilidad de la recopilación automática de los recursos.

Por otra parte, como vimos en el capítulo anterior, los catálogos y su recopilación de recursos manual no resulta un método de publicación eficaz para gestionar la continua proliferación de nuevos recursos heterogéneos por parte de usuarios no expertos. Por ello, comprobada la validez de la recopilación automatizada de recursos mediante *crawlers* en el mundo Web, se pueden imaginar los beneficios que aportaría al contexto de la IG. Por lo tanto, se propone trabajar en la línea de evolución que ha seguido el mundo web, consiguiendo que los metadatos, a pesar de su papel fundamental, sean totalmente transparentes al usuario. Concluyendo que este método automático representa una solución eficaz para ordenar todos aquellos recursos creados masivamente por los usuarios y que no encajan en el proceso de catalogación tradicional.

4.3 Interfaz homogénea para búsquedas espaciales

Como se detalla en la Sección 4.1.2, en el contexto de la IG la única interfaz de búsqueda especificada formalmente es el estándar CSW. Sin embargo, Walsh [218] señaló la necesidad de prestar atención a las interfaces de búsqueda y descubrimiento ampliamente utilizadas en otras comunidades de información diferentes de los servicios de catálogo establecidos en el dominio geográfico [169]. Intentando poner en práctica esta idea, se presenta *OpenSearch* como un mecanismo de búsqueda que no es específico del dominio geoespacial, pero que permitirá que los usuarios realicen consultas espaciales sobre diferentes redes sociales y servicios de una manera homogénea.

La visión de la Web 2.0 ha calado en la sociedad y los usuarios se han convertido en proveedores de contenido [25]. Actualmente el contenido disponible en Internet es en su mayoría generado por los usuarios y, debido a la popularización de tecnologías como el GPS, la localización se ha convertido en el contexto predominante en anotar cualquier tipo de recurso, dando lugar a enormes cantidades de

información georreferenciada en prácticamente cualquier dominio [70]. Este crecimiento exponencial de recursos heterogéneos georreferenciados plantea nuevos retos para su descubrimiento [160].

A pesar de su popularidad, no existen muchas aproximaciones que permitan a los usuarios buscar recursos, independientemente de la naturaleza de las redes sociales o servicios donde se encuentran publicados [160]. De hecho, la mayoría de estos servicios han desarrollado su propia API funcional y utilizan formatos y esquemas de codificación específicos. Navegar y acceder a estos servicios para tener acceso a los diferentes recursos requiere un conocimiento exhaustivo de la interfaz de búsqueda concreta de cada servicio. Esto constituye un obstáculo técnico para el descubrimiento de recursos georreferenciados desde varias fuentes. Por lo que se requiere una interfaz de búsqueda uniforme y homogénea para aumentar la visibilidad de los contenidos generados por usuarios y publicados en diferentes servicios.

Para abordar este problema nuestro enfoque se basa en la creencia de que los usuarios esperan, y en realidad están acostumbrados a, interfaces de búsqueda basadas en una entrada de datos mínima. Esencialmente, las interfaces de búsqueda avanzadas pueden ser deseables e incluso estar disponibles, pero la mayoría de usuarios prefiere las interfaces de búsqueda sencillas para buscar el contenido adecuado de entre un gran número de posibilidades. Los beneficios de interfaces de búsqueda mínimas pueden ser ilustrados por la herramienta *"búsqueda personalizada"*[149], disponible en la mayoría de navegadores como un simple cuadro de texto en la parte superior derecha, y que resulta un sencillo e interesante servicio de descubrimiento que ejemplifica el pragmatismo y la sencillez de los servicios Web 2.0 [25]. Los usuarios pueden personalizar fácilmente esta herramienta eligiendo sobre una colección de repositorios de datos y servicios, como tiendas en línea, diccionarios o motores de búsqueda. Independientemente del repositorio de datos o servicio utilizado, la interfaz de búsqueda es siempre la misma: los usuarios escriben la palabra clave y obtienen la lista de resultados de *Google*, *Amazon*[150] o *Wikipedia*.

[149] http://www.mozilla.org/es-ES/firefox/features
[150] http://www.amazon.com

Este ejemplo representa el caso de uso más generalizado de la especificación *OpenSearch* [44]. *OpenSearch* se ha convertido rápidamente en un exitoso mecanismo de búsqueda sobre miles de sitios web, servicios y repositorios que están, cada vez más, adaptándose para exponer de una forma estándar y simple sus interfaces de búsqueda.

Una vez revisadas las principales características de la interfaz de búsqueda *OpenSearch* y su extensión *OpenSearch-Geo* (ver Sección 4.1.2), esta sección está dedicada a describir cómo los usuarios pueden realizar búsquedas sobre diferentes servicios y redes sociales de forma homogénea.

Tras analizar diferentes servicios de redes sociales con capacidades de georreferenciación, se han seleccionado aquellos que admiten la funcionalidad de filtrado geoespacial a través de sus APIs públicas [77]. Los servicios seleccionados son: *Twitter* (mensajes cortos de texto casi en tiempo real), *Flickr* (fotografías), *Geonames*[151] (nombres de lugares), *Wikipedia* (descripciones enciclopédicas de nombres de lugares), *YouTube* (video) y *OpenStreetMap*[152] (geometrías vectoriales etiquetadas).

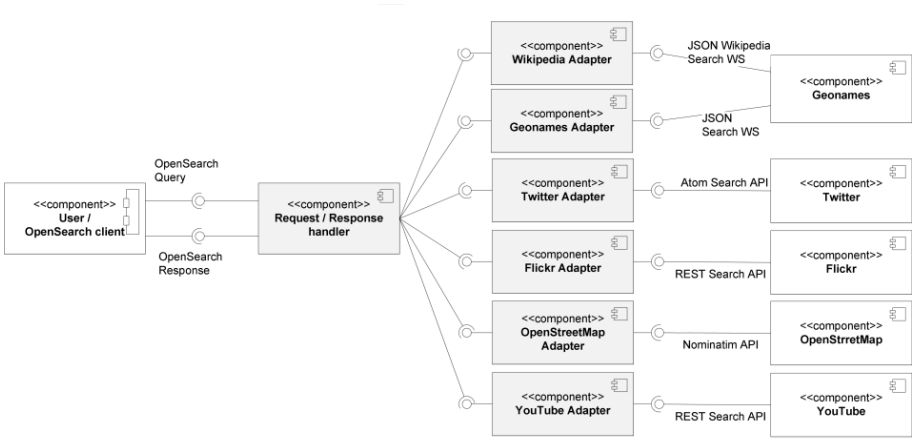

Figura 31: Adaptadores *OpenSearch*

Con el fin de proporcionar una interfaz de búsqueda común basada en *OpenSearch* sobre estos servicios, en [18] se han desarrollado un conjunto de adaptadores específicos para cada servicio, que mediarán

[151] http://www.geonames.org
[152] http://www.openstreetmap.org

entre la interfaz *OpenSearch* y las APIs particulares de los servicios. El conjunto de adaptadores junto con las redes sociales y servicios específicos se muestran en la Figura 31. Los componentes de desarrollo propio (Figura 31, sombreados) juegan un papel mediador entre las APIs específicas de los servicios (Figura 31, derecha) y potenciales clientes *OpenSearch* (Figura 31, izquierda). De modo que, una aplicación cliente podrá buscar en cualquiera de los servicios utilizando el mismo patrón de búsqueda simple.

Tanto los parámetros básicos de consulta como los opcionales pueden ser codificados en la propia URL, especificando las palabras clave, la paginación de resultados, la selección de idioma o la codificación de caracteres. Cada adaptador describe el conjunto de parámetros de consulta y los formatos de respuesta soportados a través de su documento de descripción del servicio [44]. Por ejemplo, una página web puede contener la etiqueta HTML especial *"discovery service"* que apunta al documento de descripción *OpenSearch* correspondiente. Como esta etiqueta es reconocida por la mayoría de navegadores modernos, estos activan la opción de *"añadir motor de búsqueda"* de la herramienta *"búsqueda personalizada"* mencionada anteriormente. Este sencillo mecanismo permite a las aplicaciones cliente entender las interfaces de búsqueda soportadas por los servicios y cómo construir consultas *OpenSearch* válidas.

El manejador de consultas y respuestas (*Request/Response handler* en Figura 31) actúa como un punto de entrada único para el conjunto de adaptadores por dos razones. Por un lado, permite a los usuarios controlar el procedimiento de búsqueda activando de forma selectiva el adaptador que se utilizará en función de los criterios de consulta. Los usuarios a través de un cliente *OpenSearch*, como el que incorpora la aplicación *VisioMIMEXT* (ver Sección 4.5), pueden seleccionar uno o varios adaptadores o incluso todo el conjunto. En este sentido, el cliente *OpenSearch* dentro de *VisioMIMEXT* juega el mismo papel que la herramienta *"búsqueda personalizada"* dentro de los navegadores.

Por otro lado, estas redes sociales y servicios no solo ofrecen interfaces de búsqueda específicas, sino que también proporcionan diferentes formatos de respuesta. Por ejemplo, *Flickr* devuelve los resultados de la consulta en formato *Atom* mientras que *Geonames* lo hace en formato JSON. El componente *Request/Response handler*,

por lo tanto, debe recoger y combinar todos los resultados de búsqueda en el mismo formato (*Atom*, KML o MIMEXT) para, finalmente, enviarlos a las aplicaciones cliente. Este tipo de configuración es flexible ya que es posible añadir nuevos adaptadores sin alterar la interfaz de descubrimiento desde la perspectiva del cliente. De este modo, los clientes y los adaptadores son componentes independientes, débilmente acoplados donde cada uno evoluciona por separado, mejorando la escalabilidad del sistema en su conjunto [172].

Aunque algunos servicios exponen interfaces de búsqueda a través de la especificación *OpenSearch* (*Flickr*, *Wikipedia*) estos no ofrecen la interfaz de búsqueda *OpenSearch-Geo*. El conjunto de adaptadores desarrollados soportan consultas espaciales sobre los servicios que soportan nativamente, en cierta medida, capacidades de búsqueda geográficas a través de su propia API. Por ejemplo, los usuarios pueden buscar recursos restringidos a un área determinada de interés representada como una geometría rectangular (*bounding box*).

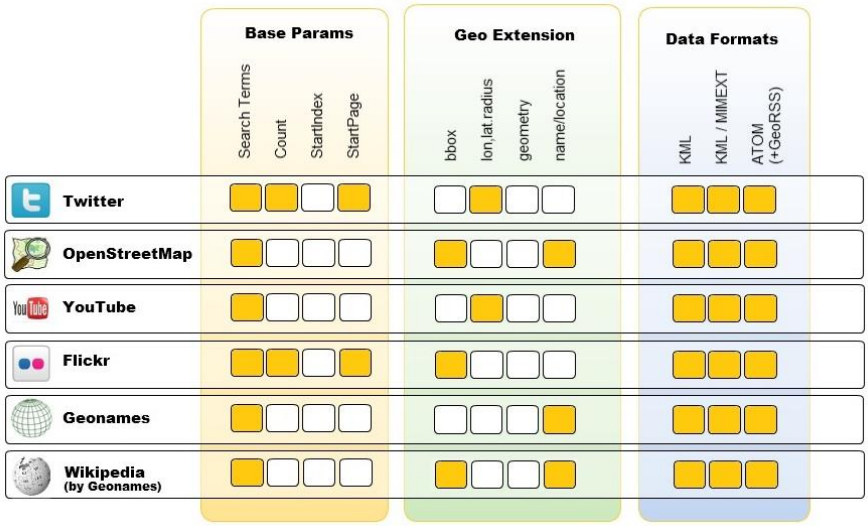

Figura 32: Filtros de búsqueda y formatos de respuesta soportados

La Figura 32 muestra las capacidades específicas de cada adaptador en términos de las características de *OpenSearch* y *OpenSearch-Geo*, junto con los formatos de respuesta soportados. Estas capacidades están limitadas por la funcionalidad ofrecida de forma nativa por cada API específica. Por ejemplo, algunos servicios

permiten el filtrado geográfico por *bounding box* y otros por un área circular determinada por su centro y su radio. En todos los casos se proporcionan como formatos geográficos estándar de respuesta KML y *Atom* ampliado con *GeoRSS*. Además, en nuestro caso, las respuestas KML incorporan la extensión MIMEXT, que gestiona colecciones de recursos georreferenciados de forma específica (ver Sección 2.3).

En cuanto a la precisión de búsqueda y rendimiento, dependerá totalmente de la precisión de las redes sociales y los servicios consultados. Por ejemplo, en realidad son pocos los *tweets*[153] georreferenciados y su localización es determinada por el perfil de usuario. En este sentido, si un usuario de *Twitter* no proporciona su ubicación, sus *tweets* no podrán ser buscados en base a los criterios espaciales propuestos. Además, la naturaleza de cada servicio conduce a limitaciones y requisitos diferentes en términos de descubrimiento. Por ejemplo, los recursos de *Flickr* pueden ser consultados en cualquier momento mientras que los recursos de *Twitter* son sólo descubribles durante una estrecha ventana de tiempo. De hecho, estas cuestiones plantean nuevos retos en el campo de la minería de datos de redes sociales [188].

4.4 Visualización

En general, la visualización se define como la generación de una imagen mental o real de algo abstracto o invisible.

De forma más específica, en el contexto científico, la visualización se entiende como la transformación de datos científicos y abstractos en imágenes. Por ejemplo, el dibujo de diagramas para visualizar funciones matemáticas o gráficos 3D para visualizar el interior de un hombre. Por lo tanto, la visualización es parte del proceso de representar la realidad. Donde los datos que se han obtenido de la realidad (aquello que se quiere examinar), generalmente datos abstractos, son transformados en imágenes permitiendo al espectador examinar y comprender la realidad de una forma más sencilla.

[153] Un *tweet* es cada una de las publicaciones de *Twitter,* consistente en un texto limitado a 140 caracteres.

La visualización se ha vuelto fundamental en el manejo y distribución actual de información, de forma que es casi imposible encontrar un artículo, libro o escrito que no incluya algún tipo de gráfico para representar sus resultados. La razón principal es que el sentido más desarrollado de los humanos es la vista y por eso la visualización es la forma más fácil de comunicar la información, especialmente cuando es compleja o viene en grandes cantidades.

Las técnicas de visualización han sido aplicadas para el análisis de datos procedentes de cualquier área científica desde tiempos inmemoriales. Sin embargo, con la aparición de la informática las técnicas de visualización se han sofisticado, permitiendo el uso de los gráficos por computadora, que permiten el manejo de datos cada vez más complejos.

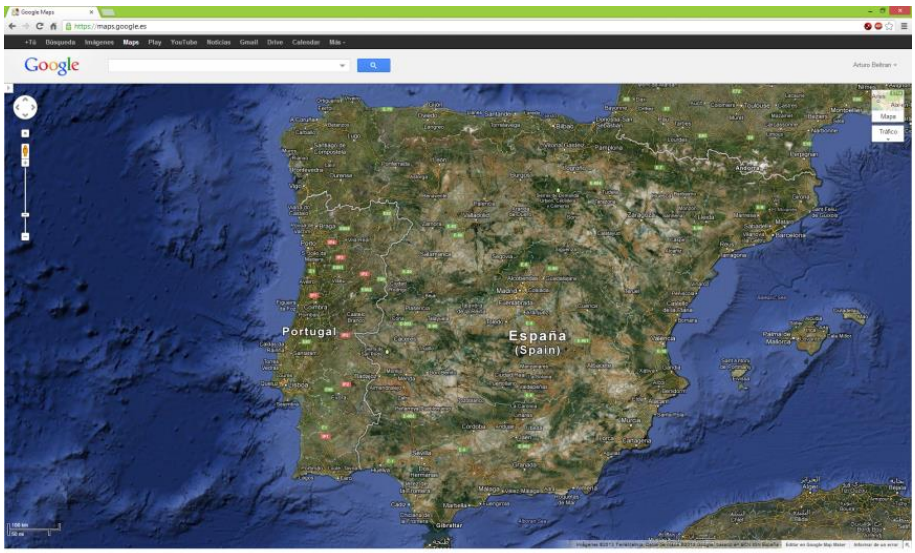

Figura 33: Ejemplo de servicio de mapas (*Google Maps*)

En el contexto de la IG, tradicionalmente la información ha sido representada sobre mapas en papel. Pero tras la aparición de la informática y de los primeros SIG basados en la web han aparecido multitud de herramientas y servicios. En los últimos años el área de la IG ha sufrido un gran auge y han aparecido numerosos servicios probablemente como consecuencia de la entrada de grandes empresas en el negocio de la IG (*Google*, *Yahoo!* o *Microsoft*), el incremento y mejora de las conexiones a Internet y el amplio uso de dispositivos de localización como receptores GPS. La herramienta

más común en este campo son los servicios de mapas disponibles en la Web [154]. Actualmente existen multitud de implementaciones de servicios de mapas disponibles, tanto de empresas privadas como *Google Maps*[154] (ver Figura 33), *Yahoo! Maps*[155], *Bing Maps*[156], *Map Quest*[157], etc., como integrados en geoportales [20] presentes normalmente en nodos IDE.

Estas herramientas de visualización, permiten a los usuarios acceder de forma sencilla a información geográfica y realizar búsquedas de forma integrada. Sin embargo, no sólo los servicios de mapas han llegado a ser populares, dada la naturaleza de la IG, otro tipo de aplicaciones han surgido recientemente para visualizar contenido geográfico en 3D[158].

La visualización de datos geoespaciales siempre ha sido un aspecto importante en el desarrollo de aplicaciones geoespaciales. Huang et al. [110] introdujeron el *Virtual Reality Modelling Language*[159] (VRML) como una forma de integrar las tecnologías SIG con técnicas de realidad virtual para la visualización el análisis y la exploración de datos espaciales. Zhu et al. [237] introdujeron 3D GIS en un entorno urbano que permitía a los usuarios explorar la información ambiental y cultural sobre la ciudad. Con la popularización de las tecnologías web, Hobona et al. [106] exploraron el uso de las tecnologías *Java 3D*[160] para soportar la visualización geoespacial en 3D a través de la Web de datos vectoriales y ráster. Zhang et al. [234] desarrollaron un entorno virtual 3D en línea como una plataforma de colaboración para la publicación, intercambio y análisis de información geoespacial.

Tal y como la tecnología ha ido avanzando, los globos virtuales basados en tecnología 3D se han convertido recientemente en parte del entorno de los SIG, ofreciéndonos técnicas y entornos de visualización más intuitivas [32] [49]. En comparación con las anteriores aplicaciones de visualización de IG en 3D [106] [110] [234] [237], la característica más importante de los globos virtuales es la

---

154 https://maps.google.es
155 http://maps.yahoo.com
156 http://www.bing.com/maps
157 http://www.mapquest.es
158 http://es.wikipedia.org/wiki/Tridimensional
159 http://www.w3.org/MarkUp/VRML
160 http://java3d.java.net

visualización perfecta de diferentes dimensiones espaciales y el ajuste del nivel de zoom. En primer lugar, las diversas dimensiones espaciales (accidentes geográficos, carreteras, edificios, etc.) permiten a los usuarios visualizar la Tierra en su conjunto y saltar desde un punto de vista a otro. En segundo lugar, las mejoras en las técnicas de zoom permiten a los usuarios acercarse y alejarse sin problemas y con una resolución continua [230].

Figura 34: Ejemplo de globo virtual (*Google Earth*)

Recientemente Sheppard y Cizek [198] han remarcado los potenciales beneficios de los sistemas basados en globos virtuales. Poniendo de manifiesto que los globos virtuales no solo proporcionan un modelo 3D de la Tierra con imágenes de satélite, sino que también ofrecen a los usuarios un acceso rápido a grandes cantidades de información geoespacial, obteniendo altos niveles de satisfacción debido a su capacidad para visualizar y navegar a través de sus propios recursos georreferenciados. Esto enlaza con la creciente cantidad y variedad de contenidos generados por usuarios disponibles en las redes sociales, y el hecho de que dichos recursos pueden ser potencialmente georreferenciados [135] y por lo tanto explorados a través de aplicaciones basadas en globos virtuales. En la Figura 34 podemos apreciar un ejemplo de estas herramientas de visualización.

Como también se señala en [198], los globos virtuales son aplicaciones fáciles de usar dirigidas tanto a usuarios expertos como a

usuarios no expertos. Esto se demuestra por el incremento de los desarrollos realizados sobre globos virtuales (por ejemplo, *Google Earth*[161] o *NASA WorldWind*[162]) en una gran variedad de escenarios tanto a nivel científico, por ejemplo para el modelado del medio ambiente [43] [208] [231], como en aplicaciones geográficas de tipo *mashup* dirigidas a usuarios no expertos [183].

Como hemos visto, la visualización de los recursos nos permite examinarlos y comprenderlos de una forma más sencilla. Por ello, la visualización juega un papel importante en el descubrimiento de recursos. Centrándonos en el contexto de la IG, dada la naturaleza de este tipo de recursos es conveniente representarlos y visualizarlos en entornos 3D. En este sentido, los globos virtuales resultan una herramienta muy útil permitiendo la visualización e integración de datos tanto para usuarios expertos como para usuarios no expertos. Además, utilizados como complemento de interfaces de búsqueda, los globos virtuales, permiten la navegación entre recursos de una forma más fácil y más intuitiva para los usuarios, ya que los mecanismos de búsqueda y visualización pueden estar integrados sin problemas en el mismo globo virtual. Por ejemplo, los usuarios pueden efectuar búsquedas más precisas considerando la zona actual que se muestra en el globo virtual o determinar su ubicación exacta pinchando en él.

4.5 VisioMIMEXT

La facilidad de producción de contenidos y su publicación en redes sociales supone una gran cantidad de recursos disponibles. Además, cada vez más los usuarios están equipados con dispositivos que incorporan localización GPS como *smartphones* o cámaras que permiten capturar la posición actual y anotarla en el recurso [135]. Como resultado, las redes sociales se están transformando en inmensos repositorios de recursos georreferenciados listos para ser compartidos, consultados y buscados por la gente. Cuando los usuarios realizan una búsqueda, en primer lugar intentan descubrir recursos de su interés escribiendo algunas palabras clave. Cuando resultados devueltos son recursos georreferenciados, como hemos

[161] http://www.google.com/earth
[162] http://www.worldwindcentral.com

visto en la sección anterior, es deseable que los usuarios puedan visualizarlos en servicios de mapas, *mashups* multimedia o aplicaciones de globo virtual.

Algunas herramientas como reproductores multimedia o servicios de compartición de fotos ya permiten a los usuarios visualizar sus contenidos teniendo en cuenta su contexto espacial. Sin embargo, los usuarios demandan constantemente nuevas capacidades para georreferenciar, descubrir y visualizar cualquier recurso multimedia. Por ello, las nuevas aplicaciones requieren un enfoque integrado que combina tecnologías y técnicas de diferentes campos como la recuperación de información, la anotación de los recursos y los SIG.

En este trabajo, se quiere destacar el papel de los globos virtuales que, gracias a la visualización, resultan herramientas de fusión de datos tanto para los expertos como para el público en general. Con este objetivo en mente, deseamos integrar en el globo virtual un mecanismo de búsqueda espacial que nos permita buscar y recuperar recursos georreferenciados de diversas fuentes de forma homogénea, funcionalidad que como hemos visto ofrece *OpenSearch* (ver Sección 4.3). Proporcionando a los usuarios un mecanismo para buscar fácilmente recursos no sólo basándose en palabras clave o etiquetas, sino también teniendo en cuenta restricciones espaciales como un área geográfica determinada. El conjunto de resultados, visto en términos de colecciones de recursos georreferenciados, podrán ser anotados de forma homogénea usando MIMEXT, la extensión de KML propuesta en la Sección 2.3 como solución para georreferenciar cualquier tipo de recurso.

En este contexto se desarrolló la aplicación VisioMIMEXT que extiende una herramienta de globo virtual para integrar sobre él la búsqueda, visualización y explotación conjunta de estas colecciones de recursos georreferenciados. VisioMIMEXT tiene como objetivo mejorar la búsqueda y visualización de recursos georreferenciados centrándose en el punto de vista del usuario. VisioMIMEXT está desarrollado en Java sobre la plataforma *Eclipse Rich Client Platform*[163] (RCP) de forma que puede ser usada en diferentes sistemas operativos. Para implementar el globo virtual dentro de la aplicación y para visualizar los diferentes recursos convenientemente

[163] http://www.eclipse.org/home/categories/rcp.php

en base a su localización geográfica se ha usado el *SDK World Wind Java*[164] desarrollado por la NASA. Actualmente soporta la visualización de recursos georreferenciados de diferentes tipos como texto, audio, imágenes y video, aunque puede ser extendido para soportar otros tipos de recursos. La Figura 35 representa la estructura de esta aplicación.

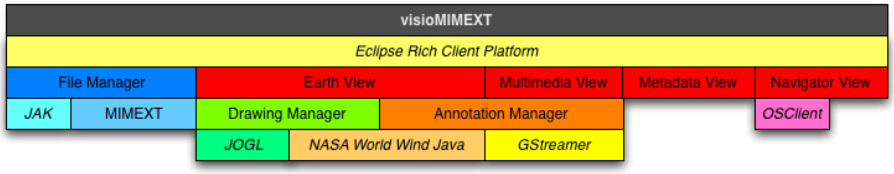

Figura 35: VisioMIMEXT - Estructura de la aplicación

La aplicación VisioMIMEXT se ha desarrollado sobre la plataforma Eclipse RCP de forma que gran parte de las tareas relacionadas con la interfaz gráfica de la aplicación ya están implementadas, además permite su ejecución en distintas plataformas. Por debajo de esta capa se pueden observar los distintos componentes que se han desarrollado para la aplicación. Los principales componentes de cualquier aplicación RCP son las vistas (*view*) que representan la parte visible de la aplicación, en la Figura 35 apreciamos *Earth View*, *Multimedia View*, *Metadata View* y *Navigator View* (representadas en rojo). La explicación de la estructura y funcionalidad de la aplicación se organizará en base a ellas.

Navigator View

El objetivo de *Navigator View* es facilitar la navegación a través del globo virtual implementado por la *Earth View* y permitir realizar consultas con contexto espacial. Dado que *Navigator View* se ocupa de los aspectos de la interfaz de usuario, integrará todos los componentes funcionales que implementan la *geocodificación*[165], la gestión de capas y el cliente *OpenSearch*.

La funcionalidad de *geocodificación* permite a los usuarios *"volar a"* (*fly to*) cierta posición en el globo calculando las coordenadas

[164] http://worldwind.arc.nasa.gov/java

[165] La *geocodificación* es el proceso que permite obtener unas coordenadas geográficas a partir del nombre o dirección de un lugar, y viceversa.

geográficas dado el nombre del lugar o su dirección (marco *Fly to* en la Figura 36). En este sentido, actúa como cliente, ya que el proceso de *geocodificación* es operado a través de la API del servicio *Yahoo! BOSS PlaceFinder*[166].

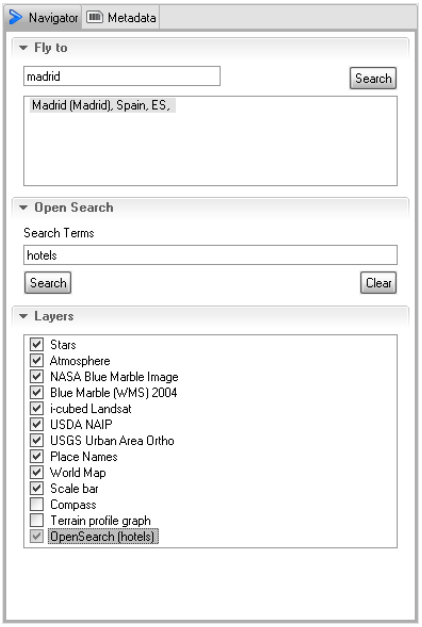

Figura 36: VisioMIMEXT - *Navigator View*

Por otra parte, *Navigator View* también gestiona el concepto de capas superpuestas sobre el globo virtual (marco *Layers* en la Figura 36), que permite activar y desactivar cualquier capa visible en la lista. Vale la pena señalar que los resultados de búsqueda se almacenan en caché y se trata como una capa. Es posible desactivar una capa con los resultados de una búsqueda y, posteriormente, recuperarla sin repetir la consulta. Por ejemplo, la capa *OpenSearch (hotels)* que aparece en la Figura 36 representa los recursos georreferenciados obtenidos como resultado de la consulta con la palabra clave *"hotels"* en la zona de Madrid.

El cliente *OpenSearch* (*OSClient* en la Figura 35) permite a los usuarios buscar recursos a través de la interfaz de búsqueda estándar de *OpenSearch* y su extensión *OpenSearch-Geo* (ver Sección 4.3). Como se ha comentado anteriormente, las consultas pueden basarse en palabras clave y/o en un área geográfica de interés (*bounding box*).

[166] http://developer.yahoo.com/boss/geo

Las palabras clave son introducidas directamente por el usuario en el campo *"Search Terms"* (marco *OpenSearch* Figura 36), mientras que el área de interés se fija automáticamente en la consulta en base a la vista activa del globo virtual (Figura 37, derecha). Por ejemplo, cuando se acerca al nivel de la ciudad de Madrid gracias a la funcionalidad de *geocodificación*, esta área geográfica se inyecta en el parámetro *"box"* de la consulta, de modo que los recursos recuperados estarán restringidos a dicha zona. De este modo, mientas las palabras clave son explícitamente tecleadas por los usuarios, los criterios espaciales de búsqueda son obtenidos automáticamente de acuerdo con el nivel de zoom actual (área geográfica visible) en la *Earth View* (Figura 37).

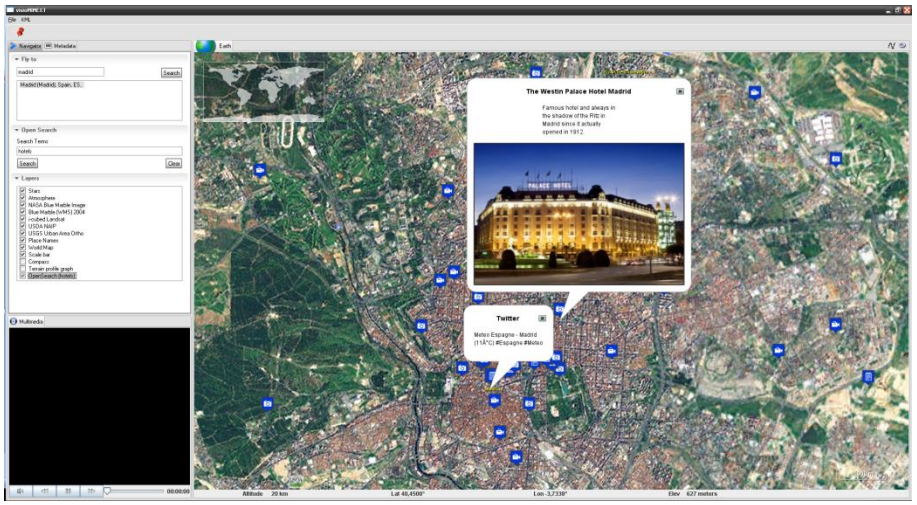

Figura 37: VisioMIMEXT - Interfaz principal

Como en el caso del servicio de *geocodificación*, el componente *OpenSearch* también actúa como una aplicación cliente que realiza la consulta del usuario y obtiene la colección de recursos resultante ya codificada en el formato MIMEXT. En lugar de depender de servicios externos o de terceros, la consulta se realizará sobre los adaptadores descritos en la Sección 4.3. Gracias a ellos, la consulta del usuario se lanza a varias redes sociales y, posteriormente, integra los resultados codificándolos en un archivo en formato MIMEXT que los describe. Por lo tanto, el cliente *OpenSearch* (Figura 36 y Figura 37 izquierda) sólo captura el resultado MIMEXT de la consulta del usuario y lo envía al componente *File Manager* (Figura 35) para que lo interprete y lo visualice en *Earth View* y *Metadata View*, como se detalla a continuación. La interfaz completa y las diferentes vistas de la

aplicación VisioMIMEXT se muestran en la Figura 37, el ejemplo muestra sobre el globo virtual los recursos georreferenciados recuperados usando la palabra clave *"hotels"* en el área de Madrid.

Metadata View

El principal objetivo de *Metadata View* es permitir que los usuarios inspeccionen la colección de recursos desde la perspectiva de MIMEXT. En lugar de mostrar todas las etiquetas de MIMEXT, por razones de claridad, esta vista permite inspeccionar en una estructura de árbol las etiquetas más relevantes como <*Name*>, <*Description*> o <*ExtendedData*>, así como las anotaciones relacionadas con la localización y los tipos MIME de los recursos. Es posible acceder a todos los recursos listados en la Figura 38 sobre el globo virtual dado que ambas vistas (*Earth* y *Metadata*) están sincronizadas. La Figura 38 es una captura de pantalla de *Metadata View*, continuando con el ejemplo anterior, se muestran los recursos recuperados en formato MIMEXT y en el recurso expandido podemos ver todos los detalles de una imagen recuperada desde *Flickr*.

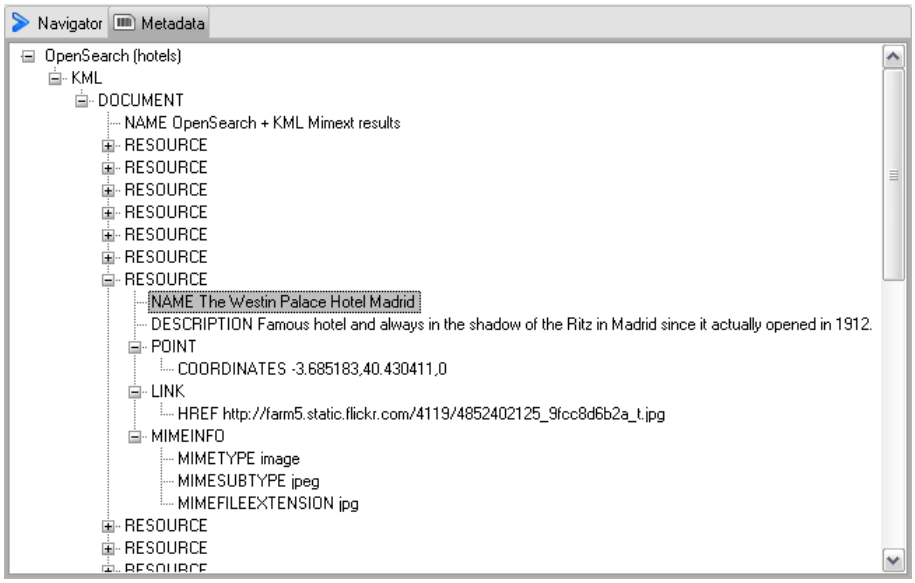

Figura 38: VisioMIMEXT - *Metadata View*

La funcionalidad de *Metadata View* se delega en el componente *File Manager* (Figura 35), que se encarga de procesar e interpretar los archivos KML y MIMEXT.

Los métodos de *parseo*[167] permiten leer los resultados de búsqueda entrantes e interpretar la estructura y la semántica de las etiquetas MIMEXT. Después de estudiar diferentes opciones para *parsear* KML, se eligió la librería *Java API for KML*[168] (JAK). A diferencia de la popular librería *libKML*[169], JAK ofrece una implementación completa del estándar KML y también soporta extensibilidad, lo que la hace adecuada para la implementación de *parsers* para extensiones de KML como MIMEXT.

El componente *File Manager* también puede descomprimir archivos KMZ y cargar su contenido en el globo virtual. Por el contrario, un archivo MIMEXT puede ser comprimido junto con los recursos que describe en un archivo KMZ permitiendo compartir e intercambiar una colección de recursos georreferenciados en un solo archivo.

Earth View

Para abordar los asuntos de visualización de los recursos multimedia, la aplicación VisioMIMEXT proporciona las vistas *Earth View* y *Multimedia View*. La primera de ellas proporciona un modelo 3D de la Tierra y permite visualizar la mayoría de los recursos. La segunda es una vista especializada para la reproducción de determinados tipos de contenido que no se pueden ser reproducidos directamente sobre el globo virtual.

La Figura 39 es una captura de pantalla de *Earth View*, en ella se están visualizando diferentes tipos de recursos georreferenciados sobre el globo virtual. *Earth View*, para las tareas de visualización, se basa principalmente en el componente *Drawing Manger* (Figura 35). En lo referente a los aspectos gráficos, como hemos dicho anteriormente, se utiliza el globo virtual *NASA World Wind Java*[170] (WWJ). Esta herramienta gestiona el globo virtual en sí mismo y las capas que se superponen sobre él. Todos los elementos de KML pueden ser potencialmente visualizados sobre el globo virtual, sin embargo, sólo aquellos que se refieren a los recursos (etiquetas KML

[167] *Parsear* implica la utilización de un analizador sintáctico que convierte el texto de entrada en otras estructuras (comúnmente árboles), que son más útiles para el posterior análisis de los datos.
[168] http://code.google.com/p/javaapiforkml
[169] http://code.google.com/p/libkml
[170] http://worldwind.arc.nasa.gov/java

<*Placemark*> y MIMEXT <*Resource*>) han sido implementados y en consecuencia visualizados, así como también aquellos recursos incluidos dentro de las etiquetas KML <*Document*> y <*Folder*> que actúan como contenedores. Por ejemplo, en la Figura 38 se muestra la lista de etiquetas <*Resource*> anidadas dentro de la etiqueta <*Document*> que contiene toda la colección de recursos recuperados como resultado a la consulta.

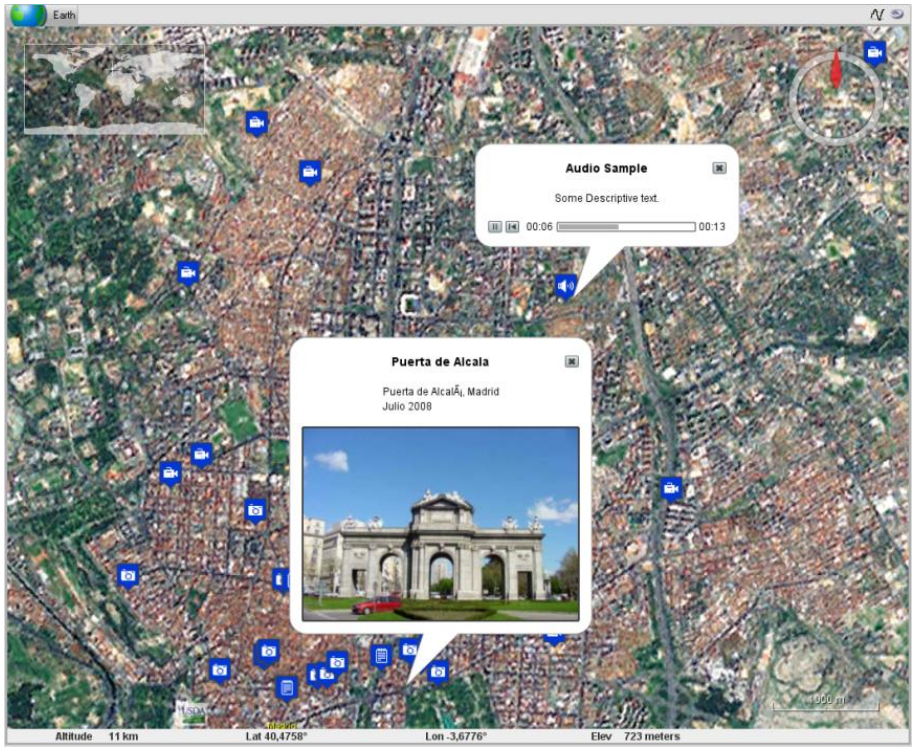

Figura 39: VisioMIMEXT - *Earth View*

La representación geográfica de estas etiquetas KML se basa en la anotación sobre la localización de los recursos, que contiene su geometría asociada (puntos, líneas, polígonos, etc.). Por otra parte, en base al tipo de recurso se utilizan diferentes tipos de iconos para identificar visualmente los diferentes tipos. Por ejemplo, en la Figura 39, se pueden apreciar diferentes iconos azules asociados a los recursos de audio, video, fotos y texto.

Nos encontramos con algunas limitaciones en la forma en que los métodos geométricos de WWJ transformaban las geometrías de KML (<*Point*>, <*Line*>, <*Polygon*> y <*LinearRing*>) en sus homólogas

formas gráficas. WWJ se basa en las primitivas de *OpenGL* [229] para *renderizar*[171] la representación geométrica en el globo. El principal problema era que ni WWJ ni *OpenGL* soportaban el formato KML de forma nativa, lo que significa que no hay equivalencia directa entre geometrías KML y las primitivas geométricas *OpenGL*.

Esta cuestión fue abordada extendiendo la herramienta WWJ para soportar la visualización de los elementos de KML (y MIMEXT). De este modo, es posible interpretar la geometría de KML, transformarla en primitivas de *OpenGL* y, finalmente, visualizar los resultados en el globo virtual. En particular, nos basamos en *Java bindings for OpenGL*[172] (JOGL) que nos permite adaptar cualquier geometría de origen a las primitivas geométricas de *OpenGL*. Por ejemplo, la etiqueta <*LinearRing*> de KML puede ser representada en *OpenGL* como una línea con el mismo punto de inicio y fin.

Por otra parte, el componente *Annotation Manager* (Figura 35) gestiona la representación gráfica de los recursos dependiendo de su tipo. Por ejemplo, un recurso de audio es visualizado de forma diferente a una foto (Figura 39). Estos elementos son conocidos generalmente como *"annotations"* en WWJ o *"balloons"* en *Google Earth*[173], y se utilizan normalmente para representar, adaptar o categorizar un recurso sobre el globo virtual. Por esta razón, el componente *Annotation Manager* incluye un conjunto de clases, cuyo objetivo es ampliar los *balloons* que WWJ define por defecto y permitir la representación de los diferentes recursos georreferenciados integrada en el globo virtual, dependiendo de su naturaleza o tipo. Por ejemplo, se implementan diferentes *balloons* para recursos de tipo audio e imagen, de forma que mientras los *balloons* para audio integran un sencillo reproductor para reproducir el recurso, los *balloons* para imágenes pueden mostrar la imagen directamente (Figura 39).

Al hacer clic en un icono sobre el globo virtual, un *balloon* se abre y el recurso multimedia seleccionado se muestra (o se reproduce en el caso de un archivo de audio). Sin embargo, algunos tipos de recursos implican una complejidad adicional que hace que sea difícil visualizar

---

[171] *Renderizar* es el proceso por el que se genera una imagen desde un modelo.
[172] http://java.net/projects/jogl
[173] De aquí en adelante utilizaremos el término *balloon* en vez de *WWJ annotation* para evitar confusiones.

el recurso dentro del globo. Para estos casos, hemos diseñado la *Multimedia View*.

Multimedia View

La *Multimedia View* permite reproducir todo aquel contenido que, por razones técnicas, no puede ser reproducido directamente sobre el globo virtual. En este caso nos referimos de forma particular a los recursos de vídeo. El problema proviene de las limitaciones inherentes de las escenas de *OpenGL* y, dado que WWJ está basado en esta libraría gráfica, cualquier elemento visualizado en el globo virtual debe ajustarse a dichas limitaciones, lo que inevitablemente impide la correcta reproducción de vídeos [224] [229].

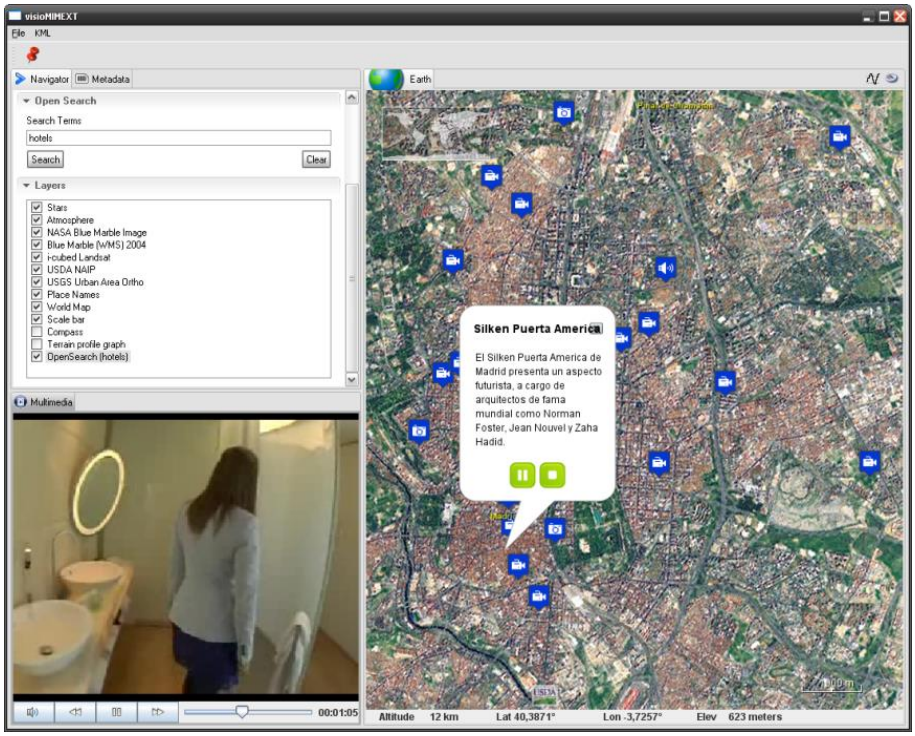

Figura 40: VisioMIMEXT - *Multimedia View*

Por ello, se sugiere como solución la *Multimedia View*. Los *balloons* para recursos de vídeo incluyen botones que permiten a los usuarios controlar la reproducción, sin embargo, la reproducción del contenido de vídeo se realiza en una vista separada distinta de *Earth View*. Continuando con el ejemplo anterior de los hoteles en Madrid, la Figura 40 muestra la *Multimedia View* en la esquina inferior izquierda

(nombrada como *"Multimedia"*) mientras se reproduce un video, los botones de control aparecen ligados al *balloon* georreferenciado correspondiente.

De igual forma que en la *Earth View*, el componente *Annotation Manager* gestionará los *balloons* para el contenido de vídeo en la *Multimedia View*. Este componente contiene la librería multimedia necesaria para la gestión y reproducción de video. En concreto, se utilizó la librería de código abierto *GStreamer*[174] que soporta multitud de formatos de vídeo.

Consideraciones finales sobre VisioMIMEXT

En esta sección, se ha presentado la aplicación VisioMIMEXT como una prueba de concepto que integra interfaces de búsqueda *OpenSearch*, la extensión de KML MIMEXT para la georreferenciación de los recursos, y un conjunto de componentes que permiten la búsqueda y la visualización de diferentes tipos de recursos georreferenciados sobre un globo virtual.

La aplicación propuesta demuestra cómo las interfaces de búsqueda *OpenSearch* permiten mejorar el descubrimiento de recursos, es decir, los usuarios son capaces de encontrar recursos relacionados formulando una consulta uniforme sobre diferentes redes sociales y servicios, restringiendo los resultados a un contexto geográfico. Esta interfaz de búsqueda, complementada con la extensión de KML MIMEXT, permite enriquecer los resultados de búsqueda con aspectos como el tipo MIME de los recursos o su posición geográfica. Además, el descubrimiento de recursos a través de una interfaz de usuario basada en un globo virtual resulta más fácil e intuitivo desde el punto de vista de los usuarios, dado que los mecanismos de búsqueda y visualización se integran de forma natural en el mismo globo virtual. De este modo, los usuarios pueden efectuar búsquedas más precisas considerando el área visualizada en el momento o determinar su ubicación exacta.

[174] http://gstreamer.freedesktop.org

5. Solución Integrada

La cantidad y variedad de recursos georreferenciados disponibles en la web crece día a día [137] [128]. Este hecho demuestra el interés de los usuarios y el papel fundamental que la información con contexto geográfico juega en la sociedad. Por ello, resultan necesarios mecanismos que permitan descubrir recursos georreferenciados de acuerdo a su contenido y sus características.

Actualmente existen numerosas redes sociales y servicios Web especializados donde los usuarios pueden compartir sus recursos (imágenes, video, texto, etc.) y que permiten realizar búsquedas en base a una localización. Por otra parte, de manera más formal, en el contexto de los SIG se han realizado grandes esfuerzos en generar catálogos de metadatos [113] para mejorar el descubrimiento de la IG que contienen. Sin embargo, la publicación de recursos georreferenciados continúa siendo un proceso manual, tedioso y realizado normalmente por expertos, lo que dificulta su compartición y, en consecuencia, aún resulta complicado encontrar contenidos georreferenciados relevantes.

Podemos tomar como referencia el mundo web, en su inicio el *World Wide Web*[175] (WWW) comenzó a ser poblado con recursos de forma masiva, haciendo cada vez más difícil encontrar contenidos que fueran relevantes. En consecuencia, aparecieron los primeros directorios para clasificar y facilitar el descubrimiento de los sitios web

[175] http://es.wikipedia.org/wiki/World_Wide_Web

y sus recursos. En cierto sentido, estos directorios eran similares a los catálogos actuales, donde los creadores de contenido pueden clasificar sus recursos en función de ciertos criterios como la calidad o áreas temáticas, es decir, creando y publicando de forma manual los metadatos. Los directorios incrementaron la accesibilidad a los recursos y facilitaron su descubrimiento en base a los intereses de los usuarios. Sin embargo, este sistema no resultó eficaz, la cantidad de trabajo para registrar un recurso en un gran número de directorios no era gratificante para los creadores de contenido. Además, el sistema era susceptible al engaño debido a que el contenido era clasificado de acuerdo a los deseos de sus creadores.

La evolución de este sistema dio lugar a los buscadores, *Lycos*[176], *Yahoo!* o *Google* se dieron cuenta de las deficiencias y empezaron a recopilar ellos mismos información de cada recurso cuyos creadores dejaban accesible en la WWW. Esta labor es realizada por los conocidos robots o *crawlers*, que se dedican a recorrer sistemáticamente los recursos disponibles con el fin de obtener de ellos el máximo de información (metadatos) que su tecnología les permite. En base a estos metadatos se podrán indexar los recursos de una forma más exacta y eficiente proporcionando resultados más relevantes y exactos a las búsquedas realizadas por los usuarios. Tal ha sido el éxito de estos buscadores que hoy es inimaginable la búsqueda de información y la navegación por la red sin acceder a alguno de estos servicios.

Con el objetivo de facilitar el descubrimiento de recursos georreferenciados esta tesis, como contribución, propone un *workflow* (ver Figura 41) basado en el funcionamiento común observado en estos sistemas. Este *workflow* introducido brevemente en el Capítulo 1 ha servido como hilo conductor para estructurar esta tesis.

Figura 41: *Workflow* propuesto

En primer lugar, como hemos visto en el Capítulo 4, cualquier sistema necesitará previamente descubrir los recursos que

[176] http://www.lycos.com

incorporará, por ello se ha incluido la fase de *Recopilación*. Seguidamente, como hemos visto en el Capítulo 2, en este escenario heterogéneo las descripciones de los recursos parecen ser la pieza clave de cualquier SI. En la fase de *Descripción*, los metadatos permiten describir los recursos en base a sus propiedades, características y contexto (tipo, contenido, origen, calidad, fecha de creación, localización, etc.). Por otra parte, como hemos visto en el Capítulo 3, la *Publicación* de los recursos de acuerdo a sus descripciones, permite organizarlos y facilita su descubrimiento. Finalmente, como hemos visto en el Capítulo 4, en la fase de *Descubrimiento*, intervienen las interfaces de búsqueda homogéneas que, junto con las herramientas que permiten la *Visualización* y/o explotación de los recursos, ayudan al usuario a encontrar contenidos georreferenciados relevantes de una forma integrada y sencilla. Como se muestra en la Figura 41, los *Metadatos* son el elemento clave del *workflow*. Los metadatos, obtenidos en el proceso de descripción, posibilitan la publicación de los recursos en base a sus características, permitiendo su descubrimiento y la correcta explotación por parte de los usuarios.

Comprobada la validez de esta metodología en la WWW, es fácil imaginar los posibles beneficios en el contexto de la IG. Por lo tanto, se ha trabajado en la línea de evolución que siguió la WWW, intentando que los metadatos, a pesar de su papel fundamental en todo el proceso, sean totalmente transparentes para los usuarios. Pero que, a su vez, permitan la recopilación, descripción, publicación, descubrimiento y visualización o explotación de los recursos disponibles de forma eficiente.

Por ello, en este capítulo se presenta una primera aproximación para desarrollar un sistema de indexación y búsqueda de recursos georreferenciados, llamado *GeoCrawler*. El principal objetivo de *GeoCrawler* es mejorar el descubrimiento de estos recursos y, consecuentemente, su accesibilidad. Para ello, de forma sinérgica, maneja diferentes tipos de recursos georreferenciados, diferentes métodos de generación de metadatos y diferentes estrategias de publicación y descubrimiento. Para permitir y facilitar el descubrimiento de los recursos *GeoCrawler* recopilará, describirá y publicará de forma automática todos los recursos que encuentre disponibles.

5.1 Arquitectura del Sistema

De acuerdo con la visión general que ofrece el *workflow* propuesto (ver Figura 41) y siguiendo la línea de evolución de la WWW, visualizamos un sistema autónomo que recopile automáticamente todos los recursos georreferenciados disponibles, los describa en función de sus características (incluyendo características espaciales) y luego ofrezca una plataforma de búsqueda eficaz para permitir su descubrimiento. De esta forma, como contribución, hemos diseñado una arquitectura del sistema que cubra los requisitos antes mencionados. El sistema está formado por varios componentes, la Figura 42 representa cómo estos componentes están conectados entre sí componiendo el sistema y cómo encajan dentro del *workflow* propuesto.

El diagrama UML[177] de Componentes (Figura 42) permite visualizar con más facilidad la estructura general del sistema que se detalla a continuación.

En primer lugar observamos el componente *Crawler* que será el responsable de la *Recopilación* de los recursos disponibles. Adicionalmente este componente dirigirá el proceso de generación de metadatos e indexación.

En segundo lugar, cubriendo la funcionalidad exigida por la fase de *Descripción* del *workflow* propuesto, observamos el componente *MDGenerator*. Este componente será invocado por el *Crawler* cada vez que encuentre un recurso y se encargará de describirlo a través de la generación automática de metadatos. Siguiendo la metodología propuesta para la generación automática de metadatos (ver Sección 2.2.1) el componente *MDGenerator* obtendrá metadatos de varios componentes y los compilará para proporcionar una descripción en un formato homogéneo. El componente *MDExtractor* extraerá toda la información relevante que puede ser obtenida del propio recurso y su contenido. Por otra parte, el componente *MDContext* será el encargado de obtener toda la información relativa al contexto de creación y de explotación del recurso. Por su parte, el componente *MDMeasures* permitirá adquirir toda la información que provenga de posibles sensores u otros dispositivos de medición que estén

[177] http://en.wikipedia.org/wiki/Unified_Modeling_Language

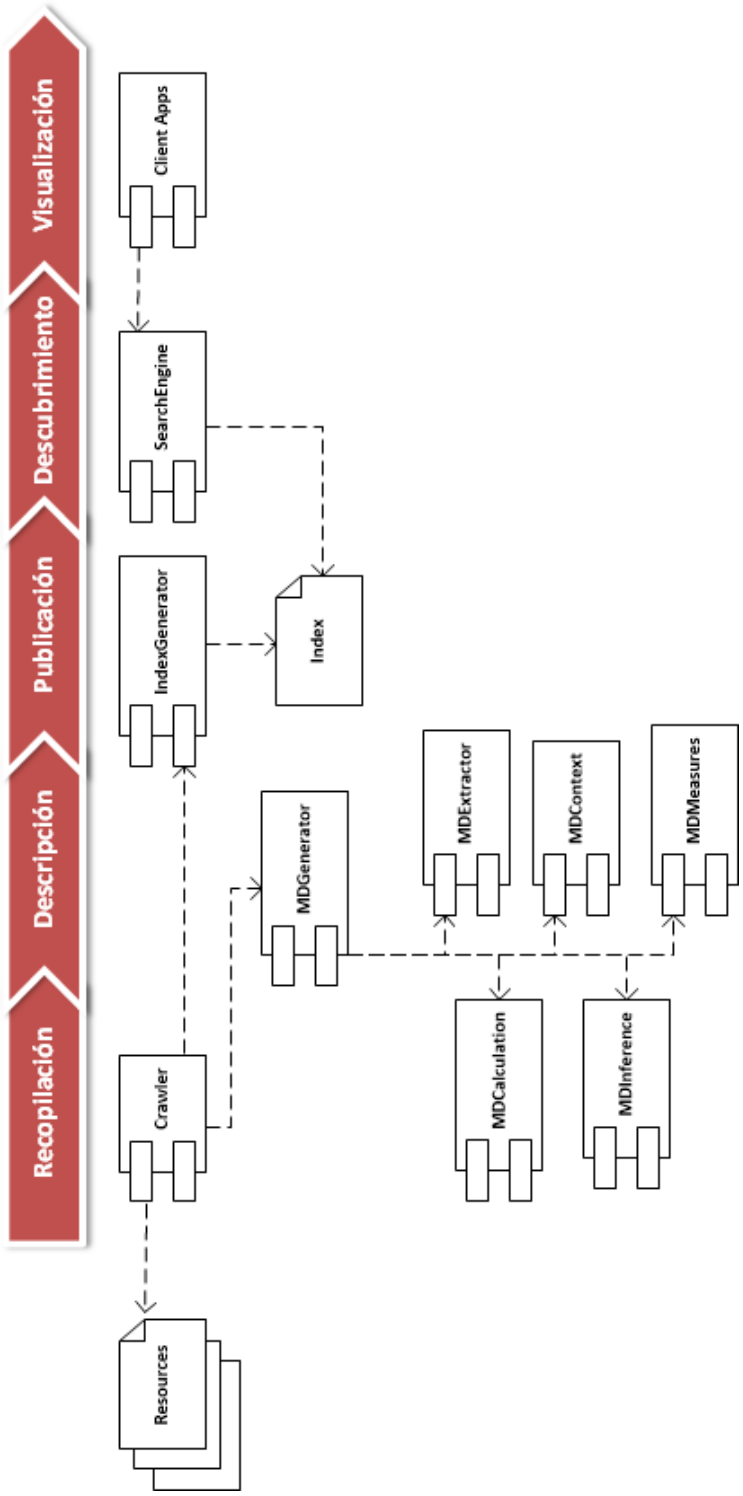

Figura 42: Arquitectura - Diagrama UML de Componentes

disponibles en el sistema. Sobre la base de metadatos obtenidos por estos componentes, *MDGenerator* podrá utilizar componentes que aplicarán métodos deductivos de generación de metadatos. Estos métodos incluyen el cálculo de nuevos elementos de metadatos a través de procesos computacionales (*MDCalculation*) y la inferencia de metadatos (*MDInference*) que, a su vez, incluye técnicas de minería de datos y de recuperación de información.

A continuación, el *Crawler* invocará al componente que ofrece la funcionalidad de la fase de *Publicación*. El responsable de la creación y mantenimiento del índice (*Index*) sobre las descripciones de los recursos es el componente *IndexGenerator*. De forma que, una vez haya obtenido la descripción del recurso, el *Crawler* lo añadirá al índice a través del componente *IndexGenerator*. El trabajo realizado por este componente, junto con el hecho de que el índice generado será utilizado por el motor de búsqueda, supone la publicación de los recursos.

Seguidamente observamos el componente *SearchEngine* que será el encargado de ofrecer la funcionalidad de *Descubrimiento*. Este componente interactúa con el sistema simplemente al utilizar el índice generado para procesar las consultas que recibe. De forma que ofrecerá interfaces de búsqueda efectivas que permitan el descubrimiento de los recursos en base a sus descripciones indexadas.

Finalmente, el componente *Client Apps* representa las potenciales aplicaciones que ofrecerán al usuario una interfaz gráfica para formular sus consultas y las herramientas necesarias para la *Visualización* de los recursos.

Gracias a la Figura 42 hemos podido entender mejor los componentes que forman el sistema y sus relaciones. Por otra parte, para entender mejor el funcionamiento del sistema y sus componentes, la Figura 43 presenta un diagrama UML de secuencia que muestra cómo interactúan los procesos entre ellos y en qué orden. El diagrama representa los participantes y los componentes que intervienen en el sistema y la secuencia de mensajes intercambiados entre ellos para llevar a cabo la funcionalidad del escenario.

Inicialmente, desde el punto de vista del sistema, la única intervención humana que requiere el sistema es ser lanzado por el administrador (*Administrator*), que deberá proporcionar unas URLs iniciales (*seeds*) para indicar al componente *Crawler* dónde debe empezar a recopilar los recursos. Cuando este componente es iniciado creará un índice (o cargará uno existente) a través del componente *IndexGenerator*. Además, iniciará el componente *SearchEngine* configurándolo adecuadamente para que use dicho índice. A partir de ese momento, el componente *SearchEngine* queda a la espera de recibir consultas por parte de los usuarios y, por su parte, el *Crawler* empieza a iterar en su rutina. Esta rutina empieza accediendo a las URLs proporcionadas, analizando los recursos obtenidos para conocer si son soportados por el sistema o no y actualizando la lista de URLs incluyendo las nuevas URLs obtenidas al analizar los recursos. Si un recurso es soportado por el sistema, será descrito a través del componente *MDGenerator*. La descripción estará formada por los metadatos obtenidos de *MDExtractor* y de los otros componentes de generación de metadatos como se mostró en la Figura 42. Entonces, *MDGenerator* compilará todos los metadatos recibidos en un registro con un formato común y los devolverá como la descripción del recurso (*generatedMD*). Finalmente, cuando el *Crawler* recibe la descripción del recurso lo añadirá al índice a través del componente *IndexGenerator*, quedando disponible para ser consultado por los usuarios a través del motor de búsqueda (*SearchEngine*) inmediatamente.

Por otra parte, como se aprecia en la Figura 43, desde el punto de vista del usuario (*User*) será posible interactuar con el sistema a través de aplicaciones cliente (*Client Apps*). El usuario podrá expresar a través de la interfaz gráfica del cliente sus necesidades de información en forma de consulta (*query*) a partir de palabras clave (*KWords*) y/o el contexto geográfico (*pos*) en el que está interesado. El cliente dará el formato adecuado a la consulta de acuerdo a una de las interfaces de búsqueda proporcionadas por el componente *SearchEngine* y realizará la petición. Tras procesar la consulta (*query*) el componente *SearchEngine* devolverá los resultados (*results*) obtenidos a la aplicación cliente que realizó la petición. Entonces, el cliente procesará estos resultados para que el usuario pueda visualizarlos y le ofrecerá una vista de ellos (*resultsView*).

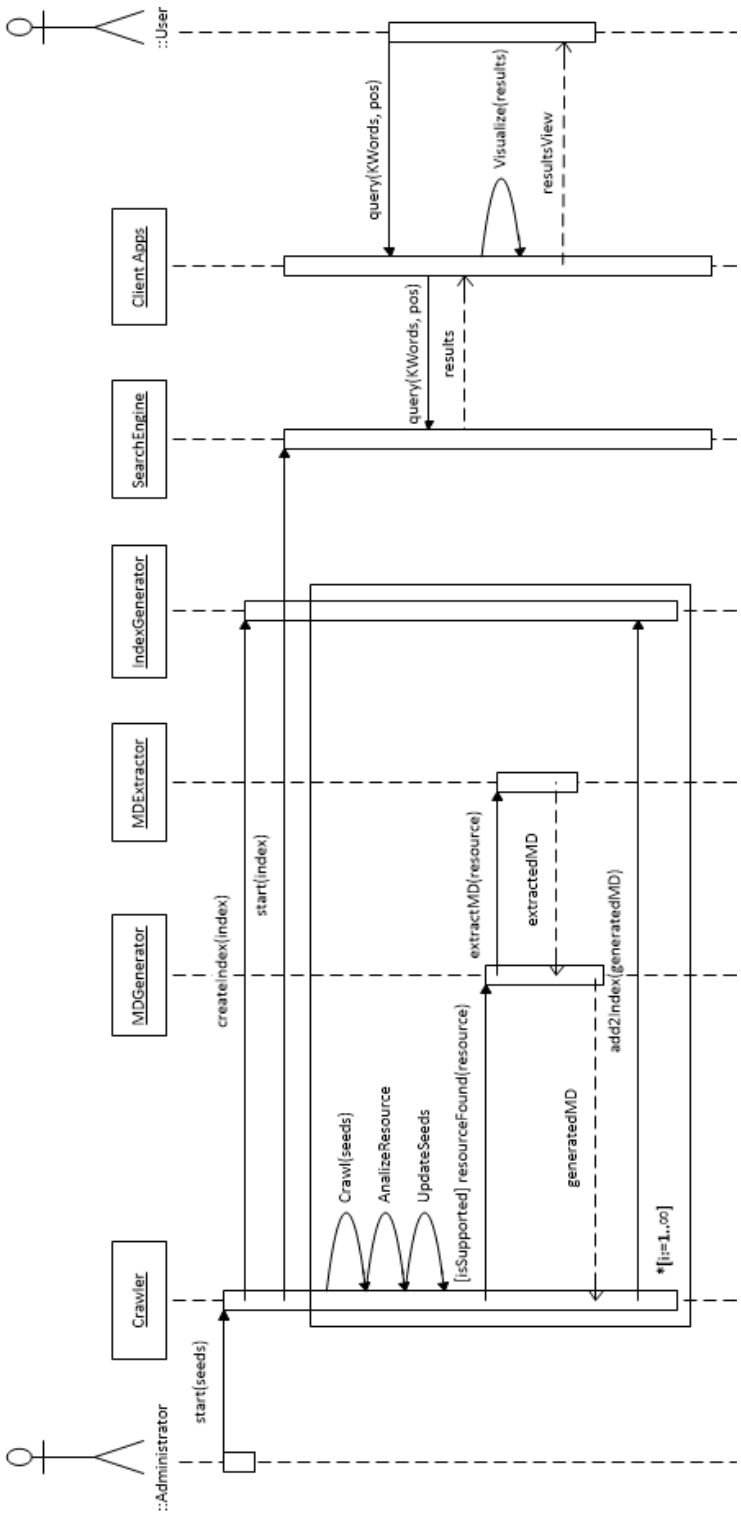

Figura 43: Arquitectura - Diagrama UML de Secuencia

Como hemos visto, combinando estos componentes podemos obtener un sistema que proporcione la funcionalidad necesaria para cubrir nuestros requerimientos. La siguiente sección explica en más detalle cómo encajan estas piezas así como otros detalles de implementación.

5.2 *GeoCrawler*

El principal objetivo de este capítulo es conseguir un sistema que permita indexar y buscar recursos georreferenciados siguiendo la metodología que ha triunfado en la WWW. En la sección anterior hemos descrito la arquitectura y la funcionalidad que debería tener el sistema. En esta sección se detalla cómo se ha desarrollado, implementando la arquitectura propuesta, un sistema al que hemos llamado *GeoCrawler*. A continuación, en primer lugar, veremos qué elementos se utilizarán para implementar el sistema. Posteriormente se detallará cómo encajan estos elementos para componer el sistema y, finalmente, se explicarán de forma muy resumida los principales detalles de implementación.

5.2.1 Elementos Utilizados

En esta sección se describen varios elementos software existentes que serán reutilizados y combinados para componer el sistema que llamamos *GeoCrawler*. Además se explicará cómo estos elementos están relacionados con los componentes presentados en la Figura 42 y, por lo tanto, cómo encajan dentro del *workflow* propuesto.

El *Crawler*

Respecto al *crawler*, teniendo en cuenta los requerimientos del sistema y tras analizar varias herramientas de este tipo [47], incluyendo una de desarrollo propio para explorar el contenido de una máquina local [46]. Se considera que la solución que más se ajusta a las necesidades es el proyecto *Nutch*[178] de la comunidad Apache.

[178] http://nutch.apache.org

Nutch es una implementación de código abierto en Java de un *crawler*, proporcionando todas las herramientas que necesita para ejecutarlo. Entre sus bondades pueden destacarse la transparencia y el entendimiento del sistema, dado que al ser de código abierto cualquiera puede ver cómo funcionan sus algoritmos. Otro de sus puntos fuertes es la extensibilidad, *Nutch* es muy flexible y permite que los desarrolladores puedan añadir funciones de filtrado de recursos, de indexación o de procesamiento para nuevos tipos de recursos.

El *crawler* obtiene los recursos y construye un índice invertido, que posteriormente un motor de búsqueda puede utilizar para responder las consultas de los usuarios. La principal ventaja de *Nutch* es que en lo referente a la generación de índices está basado en las librerías del proyecto *Lucene* que se ha comentado anteriormente. Por lo que, con la configuración adecuada, puede generar índices compatibles con este motor de búsqueda.

Nutch habitualmente puede funcionar a una de estas tres escalas: sistema de archivos local, intranet, o en la web entera. Las tres tienen características diferentes. Por ejemplo, rastrear un sistema de ficheros local es fiable en comparación con los otros dos, ya que los errores de red no se producen y las copias caché del contenido de la página no son necesarias. En contraste, el rastreo de la web entera se encuentra en el otro extremo.

De esta forma, el proyecto *Nutch* cubrirá la funcionalidad de los componentes *Crawler* e *IndexGenerator* (ver Figura 42), aunque será necesario configurarlo adecuadamente y extenderlo con nuevas funciones de indexación, así como para el filtrado y procesamiento de los recursos georreferenciados. De esta forma, tras obtener las descripciones, *Nutch* también será capaz de generar los índices textuales y espaciales de acuerdo con las consideraciones hechas en la Sección 3.3.

Generación de descripciones

Consideramos que, actualmente, es muy difícil conseguir un sistema totalmente autónomo. Dado que siempre será necesaria la intervención de los usuarios para introducir o por lo menos validar los elementos de metadatos menos intuitivos. Hay que tener en cuenta que algunos elementos de metadatos son mucho más difíciles de

averiguar o inferir que otros, por ejemplo el título, el resumen, las palabras clave, etc. La idea es empezar rellenando los campos básicos para el descubrimiento, permitiendo que el sistema ejecute búsquedas básicas de forma correcta. Más adelante se podrá evaluar la necesidad y los beneficios de completar de forma estricta los metadatos de acuerdo a un estándar (como ISO 19115 o *Dublin Core*). En este punto, bajo nuestro punto de vista, es preferible obtener los metadatos básicos para describir los recursos e indexarlos en base a ellos que quedarnos atascados intentando generar un registro de metadatos de forma rigurosa.

Por ello, a partir del conjunto de recursos disponibles obtenidos por el *crawler*, sus descripciones serán generadas a través de la plataforma común para la generación de metadatos obtenida tras la integración de los proyectos *Apache Tika* y *OSGeo FDO* [19] que hemos presentado en la Sección 2.2.3. Inicialmente el sistema sólo incorporará la funcionalidad de los componentes *MDExtractor* y *MDContext* (ver Figura 42) implementados por la plataforma propuesta.

Para incorporarlo en el sistema, se desarrollará un componente que se integrará dentro de *Nutch* y que será invocado cada vez que este encuentre un nuevo recurso. Este componente, gracias a la funcionalidad de la plataforma común para la generación de metadatos, devolverá descripciones homogéneas de los recursos para ser indexadas. De esta forma queda cubierta la funcionalidad del componente *MDGenerator* presentado en la Figura 42.

Plataforma de búsqueda

Teniendo en cuenta los requerimientos del sistema descritos anteriormente, la solución que más se ajusta a las necesidades es el proyecto de la comunidad Apache *Solr*[179] basado en el motor de búsqueda e indexación *Lucene*.

Solr es la plataforma de búsqueda de código abierto del proyecto *Lucene*. Sus características principales incluyen potentes búsquedas textuales, marcado de coincidencias, búsqueda por facetas, *clustering* dinámico, integración de bases de datos y manejo de documentos avanzados (por ejemplo, Word, PDF). Además, *Solr* es altamente

[179] http://lucene.apache.org/solr

escalable, proporcionando búsquedas distribuidas y replicación de índices. Por ello, proporciona las funciones de navegación y búsqueda de muchos de los sitios más grandes del mundo de Internet.

Solr está escrito en Java y se ejecuta como un servidor de búsqueda independiente dentro de un contenedor de *servlets*[180] como *Tomcat*[181]. *Solr* utiliza la librería de *Lucene* y sus índices como núcleo para procesar las consultas, por lo que con la configuración adecuada será totalmente compatible con los índices generados a través de *Nutch*. *Solr* ofrece interfaces de búsqueda basadas en REST[182] como HTTP/XML o JSON que hacen que sea fácil de usar desde virtualmente cualquier lenguaje de programación. La potente configuración externa de *Solr* permite que pueda adaptarse a casi cualquier tipo de aplicación, además cuenta con una arquitectura de *plugins* que nos permite extender su funcionalidad cuando se requiere una personalización más avanzada.

De esta forma, la plataforma de búsqueda *Solr* y su núcleo basado en las librerías de *Lucene* cubrirá la funcionalidad del componente *SearchEngine* presentado en la Figura 42. Aunque deberá configurarse de forma apropiada y será necesario extender la funcionalidad que ofrece *Solr* mediante un componente que de soporte a la interfaz de búsqueda *OpenSearch* y su extensión *OpenSearch-Geo* (ver Sección 4.3) que requiere nuestro sistema.

Aplicaciones cliente

Solr ofrece diferentes interfaces de búsqueda (además se ha incorporado *OpenSearch*) que hacen que sea fácil de usar desde virtualmente cualquier lenguaje de programación. Por ello, el tipo y la cantidad de aplicaciones cliente es potencialmente ilimitado. Sin embargo, para nuestro sistema, como hemos discutido en la Sección 4.4, consideramos que el cliente debe estar basado en un globo virtual que permita la búsqueda y la visualización de los recursos de forma integrada y sencilla. Por ello, usaremos la aplicación VisioMIMEXT presentada en la Sección 4.5.

[180] Los *servlets* son programas Java que se ejecutan en el lado del servidor.
[181] http://tomcat.apache.org
[182] http://es.wikipedia.org/wiki/Representational_State_Transfer

5.2.2 Composición del Sistema

Con los elementos descritos en la sección anterior es posible implementar un sistema de indexación y búsqueda, a continuación se detalla como encajan estas piezas. En la Figura 44 se puede ver el planteamiento de este sistema y su relación con el *workflow* propuesto.

En primer lugar, *Nutch* cubrirá la funcionalidad requerida por las tres primeras fases del *workflow* propuesto (*Recopilación*, *Descripción* y *Publicación*). Por una parte, *Nutch* implementa toda la funcionalidad necesaria para recopilar los recursos disponibles en el contexto en el que es ejecutado. Por otra parte, delegando la funcionalidad a la plataforma de generación de metadatos formada por *Apache Tika* y *OSGeo FDO*, obtendrá las descripciones de los recursos en un formato homogéneo. Finalmente, *Nutch* publicará los recursos en base a sus descripciones generando un índice textual y espacial (*Textual & Spatial Index*) gracias a las librerías de *Lucene*. Por otra parte, *Solr* cubrirá la funcionalidad de la fase de *Descubrimiento*. En base al índice generado por *Nutch*, *Solr* ofrece diferentes interfaces de búsqueda a través de las cuales el usuario podrá realizar sus consultas. Además, se le ha añadido soporte para la interfaz de búsqueda *OpenSearch* y su extensión *OpenSearch-Geo*. A partir de este punto, cualquier tipo de usuario desde cualquier tipo de dispositivo debería ser capaz de realizar búsquedas sobre los recursos indexados a través de aplicaciones cliente como VisioMIMEXT.

A través de VisioMIMEXT los usuarios podrán expresar sus necesidades de información a través de palabras clave y restringiendo los resultados a un contexto geográfico. En base a estas necesidades, la aplicación construirá una consulta en formato *OpenSearch-Geo* y la enviará sobre el componente que implementa dicha interfaz de búsqueda sobre *Solr*. Este componente juega un papel mediador entre las APIs específicas de *Solr* y la aplicación cliente VisioMIMEXT, de forma que convertirá la consulta *OpenSearch-Geo* para que pueda ser ejecutada sobre *Solr* de acuerdo al algoritmo que combina un índice espacial sobre un índice textual presentado en la Sección 3.3. Una vez *Solr* haya ejecutado la consulta el componente codificará los resultados obtenidos en formato MIMEXT (ver Sección 2.3) y los devolverá al cliente. VisioMIMEXT, tras recibir como resultado el

conjunto de recursos codificados en MIMEXT, los visualizará sobre el globo para que el usuario pueda examinarlos y en el caso de estar interesado reproducirlos.

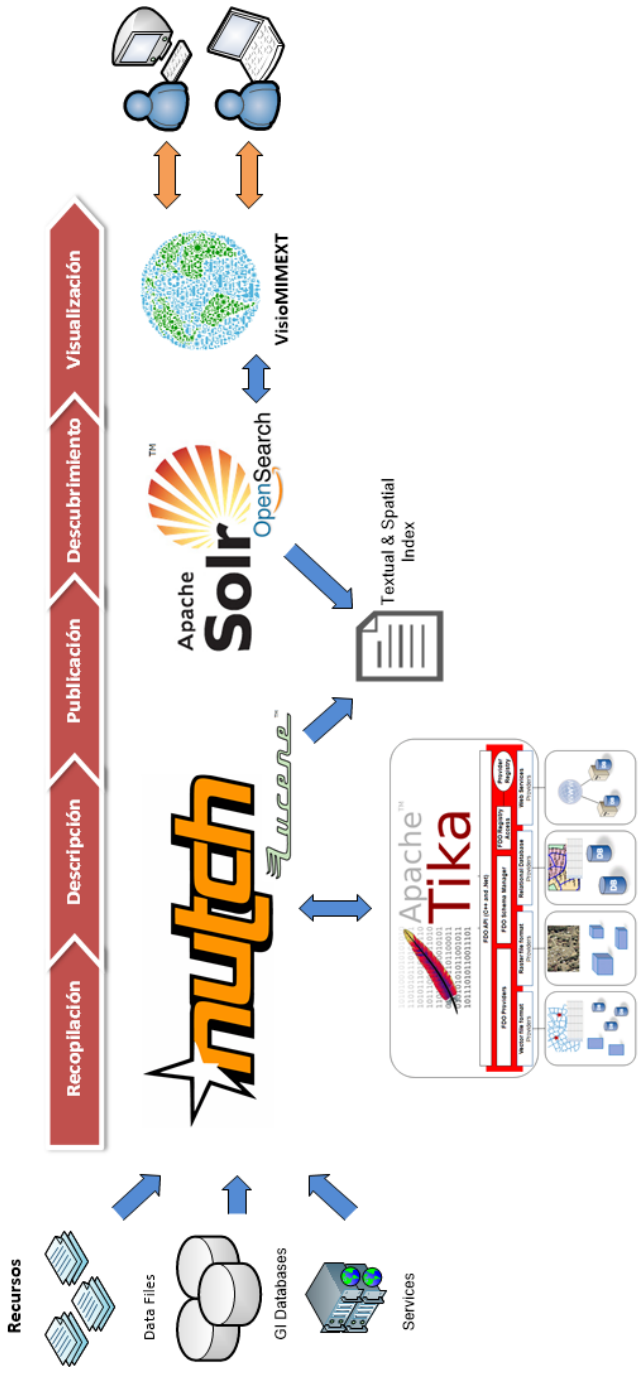

Figura 44: ***GeoCrawler***

De esta forma el sistema, siguiendo el *workflow* propuesto, cubre todo el ciclo de vida de los recursos. Empezando cuando estos son puestos disponibles por los creadores y finalizando cuando son descubiertos y explotados por los consumidores.

5.2.3 Implementación

Esta sección explica los principales detalles de implementación realizados para adaptar los elementos (ver Figura 44) que se utilizarán para componer el sistema y cómo están relacionados con los componentes presentados en la Figura 42 y, por lo tanto, cómo encajan dentro del *workflow* propuesto. Siguiendo el orden de este *workflow* empezaremos con el componente *Crawler*, como se ha explicado en la sección se usará el *crawler* del proyecto *Nutch*.

Nutch no dispone de soporte para formatos geoespaciales. Viendo la arquitectura de este proyecto, la primera idea sería acoplar la solución que integra *Apache Tika* y *OSGeo FDO* en la parte *crawler* para generar las descripciones de los recursos con formatos geoespaciales y posteriormente poder indexarlos.

Para ello, se implementó un nuevo *plugin*, que, aparte de los ficheros de configuración y compilación correspondientes, podemos resumir en dos elementos. En primer lugar se implementó un nuevo analizador o *parser* que será invocado cuando el *crawler* detecte que un recurso es de uno de los tipos soportados por la plataforma de extracción de metadatos en la que se va a basar. De este modo, de forma resumida, el *parser* a través de la funcionalidad de la solución que integra *Apache Tika* y *OSGeo FDO* [19] (ver Sección 2.2.3) generará y asociará al nuevo recurso encontrado todos los metadatos que es capaz de generar. Este *parser*, es la implementación del componente *MDGenerator* (ver Figura 42) y la plataforma de generación de metadatos cubre la funcionalidad de los componentes *MDExtractor* y *MDContext* (ver Figura 42). En segundo lugar se implementó la parte de indexación (ver *IndexGenerator* en Figura 42) que, aparte de gestionar el índice, permite generar y filtrar los contenidos que se van a indexar.

Para entender mejor este elemento, será necesario conocer un poco más acerca del funcionamiento de *Nutch* y acerca de los índices

de *Lucene*. Las librerías de *Lucene* basan su funcionamiento en objetos llamados *Documentos* (*Document*), que están compuestos por uno o varios *Campos* (*Fields*), que siguen la filosofía de propiedades en pares (nombre, valor). Un *Documento* puede ser visto como un registro del índice y los *Campos* son los elementos básicos de indexación. Aunque estos *Campos* no pueden ser indexados por si solos (deben ser agrupados en *Documentos*), las consultas se realizarán en referencia a ellos. Por lo tanto, a alto nivel, un *Documento* se correspondería con el concepto del registro de metadatos y los *Campos* con cada uno de los elementos que componen ese registro.

Nutch, para generar los índices, asocia un *Documento* de *Lucene* a cada uno de los recursos encontrados. El nuevo *plugin*, en base a los metadatos que se han asociado al recurso mediante el *parser* recopilará, filtrará y formateará de forma adecuada la información que se va a incluir en el índice, es decir añadirá los *Campos* de forma apropiada al *Documento* asociado al recurso.

Llegados a este punto, dada la naturaleza espacial de los recursos, como hemos visto en la Sección 3.3, se desean combinar índices textuales y espaciales, aportando estos últimos un gran valor añadido a las búsquedas al poder restringir los resultados en un contexto geográfico. Para ello, se va a seguir una estrategia de indexación espacial integrada sobre los índices textuales de *Lucene*. Básicamente, en la fase de indexación deben obtenerse los valores X_{max}, X_{min}, Y_{max} e Y_{min} correspondientes al MBR que contiene al recurso y sobre los que posteriormente se realizarán las consultas, e incluirlos en el índice. Es decir que para cada recurso el *parser* deberá extraer esa información y la parte de indexación incluir esos cuatro *Campos* en los *Documentos* de los recursos. Además, se deberá tener en cuenta que los valores de las coordenadas estén representados mediante un sistema de referencia común y aplicable a todo el globo terráqueo, como WGS84.

A parte de la implementación del *plugin*, hay algunos archivos de configuración de *Nutch* que deben tenerse en cuenta:

- *nutch-default.xml* y *nutch-site.xml*: en estos dos archivos en formato XML se especifican todos los parámetros de configuración de *Nutch* y sus *plugins*.

- *regex-urlfilter.txt*: este archivo contiene una serie de expresiones regulares que permiten configurar y filtrar las URLs que el *crawler* va a inspeccionar.
- *solrindex-mapping.xml*: en este archivo en formato XML se especifican las relaciones entre los campos creados por *Nutch* y sus *plugins* con los campos definidos y esperados por *Solr*.

Tras desarrollar el *plugin* y configurar adecuadamente *Nutch*, solo faltará configurar de forma apropiada *Solr*, que ha sido utilizado como la implementación del componente *SearchEngine* (ver Figura 42). Para poner a funcionar la plataforma de búsqueda la mayor parte de la configuración se va a realizar en su esquema (*schema.xml*).

En este archivo, en primer lugar, se realiza una extensa definición de tipos de datos. Para cada definición de un nuevo tipo de datos se especifica un nombre único que será usado en las definiciones de los campos y, posteriormente, varios atributos que determinarán el comportamiento real del tipo de datos. Entre estos atributos, debemos destacar el atributo *class* que indicará en cuál de los tipos de datos implementados y soportados por *Solr* se va a basar el nuevo tipo de datos que se está definiendo. Otros atributos permitirán configurar el comportamiento del tipo de datos en aspectos como cuál será el criterio de la ordenación en los resultados si el recurso no tiene valor en el campo de este tipo de datos u otros aspectos específicos de ciertos tipos de datos como la precisión de los datos numéricos. Además, para los tipos de datos basados en texto, es posible especificar de forma personalizada los analizadores que se van a utilizar, la forma en la que el texto va a ser dividido y los filtros que se van a aplicar. Todo esto permite, en base a los tipos de datos básicos (texto, numéricos, fecha…), especificar de forma muy concreta cómo van a ser analizados e indexados los valores de los campos y permite ajustar de una forma muy precisa el comportamiento y las prestaciones, tanto en rendimiento como en precisión de búsqueda posterior.

Tras la definición de tipos de datos se especifican los campos que se van a incluir. Cada campo es identificado por un nombre único, y a través de sus atributos se especifica su tipo de datos y otros aspectos como si va a ser almacenado, indexado o es un campo requerido.

Por último, se especifica cual es el campo identificador y por lo tanto único de los documentos. Por otra parte, se indicarán otros aspectos como el campo de búsqueda por defecto, para cuando en las búsquedas no se especifica un campo concreto, la operación por defecto y además se puede indicar cómo se forman algunos campos que se componen de otros. Esto último resulta muy útil para, por ejemplo, acumular en un solo campo todos los valores de los campos textuales más interesantes y definir este último como campo de búsqueda por defecto.

Campo	Tipo	Descripción
id	String	Identificador único del recurso (coincide con su URL)
host	URL	URL del host del recurso
site	String	Site del recurso
tstamp	Date	Fecha de indexación
anchor	String	Enlaces que contiene el recurso
type	String	*MIME Type* del recurso
contentLength	Long	Tamaño del archivo (si lo es)
date	Date	Fecha de creación del recurso
gc_url	URL	URL del recurso
gc_name	String	Nombre del recurso
gc_path	URL	Ruta de acceso del recurso
gc_type	String	Tipo de recurso
gc_lastModified	Date	Fecha de modificación
gc_length	Long	Tamaño
gc_description	Text	Descripción del recurso
gc_title	Text	Título del recurso
gc_keywords	Text	Palabras clave
gc_publisher	Text	Información acerca del autor del recurso
gc_schema	URL	Esquema según el tipo del recurso
gc_source	URL	Origen del recurso
gc_resourceName	String	Nombre del principal esquema interno del recurso
gc_resourceDescription	Text	Descripción del esquema interno del recurso
gc_resourceTitle	Text	Título del esquema interno del recurso
gc_resourceMinX	Double	Coordenada X_{min} del BBOX
gc_resourceMinY	Double	Coordenada Y_{min} del BBOX
gc_resourceMaxX	Double	Coordenada X_{max} del BBOX
gc_resourceMaxY	Double	Coordenada Y_{max} del BBOX
gc_resourceCRS	String	Nombre del sistema de coordenadas del recurso
gc_resourceGeometry	String	Tipo de geometría del recurso
gc_resourceNumRecords	Int	Número de registros del recurso
gc_resourceRestrictions	String	Restricciones del recurso
gc_resourceHints	String	Consejos del recurso
gc_atributeNames	String	Recopilación de los nombres de las *features*
gc_completeXML	String	XML completo generado por la plataforma de extracción de metadatos

Figura 45: Principales campos indexados

Con todo esto, el sistema de indexación y búsqueda ya es funcional. Solo queda por conocer cuáles son los campos que se han definido en el índice. La Figura 45 lista dichos campos junto con su tipo de datos (básico) y su descripción. Los campos cuyo nombre

empieza por *"gc_"* son los que proporciona el nuevo *plugin* implementado.

Finalmente, como hemos comentado anteriormente, se desea extender la funcionalidad que ofrece *Solr* mediante un componente que de soporte a la interfaz de búsqueda *OpenSearch* y su extensión *OpenSearch-Geo* (ver Sección 4.3) que requiere nuestro sistema. Para ello se ha desarrollado un nuevo componente que implementa dicha funcionalidad. Este componente juega un papel mediador entre las interfaces de búsqueda específicas de *Solr* y las consultas recibidas en formato *OpenSearch*, de forma que convertirá estas consultas para que puedan ser ejecutadas sobre *Solr*.

De forma resumida, para cada consulta recibida este componente obtendrá las palabras clave y las coordenadas de la caja envolvente (*bounding box*) especificadas por el usuario (o aplicación cliente). Con estos elementos construirá una nueva consulta válida para la interfaz de búsqueda de *Solr*. En esta nueva consulta, por una parte, para realizar el filtrado textual, las palabras clave serán buscadas sobre el campo de búsqueda por defecto que, como hemos dicho anteriormente, acumula los valores de los campos textuales más interesantes. Por otra parte, para realizar el filtrado por contexto geográfico, las coordenadas del *bounding box* especificado por el usuario serán comparadas con los campos (ver *gc_resourceMinX*, *gc_resourceMinY*, *gc_resourceMaxX* y *gc_resourceMaxY* en la Figura 45) que contienen las respectivas coordenadas de los MBRs de los recursos indexados, de acuerdo con el algoritmo planteado en la Sección 3.3. Recuperaremos aquellos recursos que **no** cumplan esta condición ya que serán los recursos cuyo MBR (MBR_1) intersecta con el *bounding box* especificado (MBR_2):

$$[MBR_1(X_{min}) > MBR_2(X_{max})] \text{ OR } [MBR_1(X_{max}) < MBR_2(X_{min})] \text{ OR}$$
$$[MBR_1(Y_{min}) > MBR_2(Y_{max})] \text{ OR } [MBR_1(Y_{max}) < MBR_2(Y_{min})]$$

Tras ejecutar la consulta, obtendremos de *Solr* los resultados que hayan pasado tanto el filtro textual como el filtro espacial, es decir, aquellos recursos cuya descripción se ajuste a las palabras clave y que se encuentren dentro del área geográfica especificada. Una vez obtenidos los resultados de *Solr* el componente codificará los resultados obtenidos en formato MIMEXT (ver Sección 2.3) y los devolverá al cliente.

5.3 Resultados

El resultado de este capítulo es un sistema de indexación y búsqueda de recursos geoespaciales en el cual, de forma paralela a la evolución del mundo web, los metadatos, a pesar de su papel fundamental, serán totalmente transparentes al usuario. Estos metadatos permiten recopilar, describir, publicar, descubrir y visualizar de forma eficiente los recursos disponibles. De modo que los recursos quedan publicados y disponibles globalmente para que puedan ser encontrados proporcionando a los usuarios la capacidad de descubrir y acceder a los recursos que alberga.

En primer lugar se abordó el problema de la recopilación de los recursos, que fué automatizada mediante aplicaciones de tipo *crawler* (ver Sección 4.2). Como se ha visto en la sección anterior, esta contribución ha sido puesta en práctica en la implementación de *GeoCrawler*, más concretamente para este componente nos hemos apoyado en el *crawler* del proyecto *Nutch* dado que ofrece toda la funcionalidad necesaria para recopilar los recursos disponibles en el contexto en el que es ejecutado.

En segundo lugar, el sistema afronta el problema de la descripción de recursos georreferenciados en base a la metodología propuesta para la generación de metadatos (ver Sección 2.2.1), de esta forma se pretende facilitar la descripción de los recursos al automatizar la producción de metadatos evitando el tedioso trabajo manual más propenso a errores. Esta metodología ha sido demostrada mediante la implementación de una plataforma de extracción de metadatos obtenida tras la integración de los proyectos *OSGeo FDO* y *Apache Tika* [19] (ver Sección 2.2.3) que permite generar descripciones homogéneas de recursos con formatos geoespaciales. Esta contribución ha sido incorporada en el sistema implementando un nuevo *plugin* para el *crawler* del proyecto *Nutch* que, delegando la funcionalidad a la plataforma de generación de metadatos, obtendrá las descripciones de los recursos. Por otra parte, también sería posible obtener las descripciones de los recursos a través de servicios DESCaaS (ver Sección 2.4), aunque estos no han sido incorporados en la primera versión del sistema.

En tercer lugar, se aborda el problema de la publicación de recursos georreferenciados siguiendo la idea de publicar los recursos

generados de forma integrada en el flujo de trabajo (ver Sección 3.2.1). Para ello, se ha usado el método propuesto para publicar los recursos mediante la indexación de sus descripciones combinando índices textuales y espaciales, de forma que los índices espaciales se integran sobre los índices textuales (ver Sección 3.3). Esta contribución ha sido comprobada y demostrada mediante la integración de un índice espacial sobre los índices textuales del proyecto *Lucene*. Estos índices, de forma integrada en el flujo de trabajo, son creados desde el *crawler Nutch* y plenamente accesibles desde la plataforma de búsquedas *Solr* dado que ambas soluciones se basan en las librerías del proyecto *Lucene* y gracias a la configuración efectuada.

A continuación, el sistema soluciona el problema del descubrimiento de recursos georreferenciados proporcionando la interfaz de consulta común y homogénea propuesta (ver Sección 4.3), esta interfaz es aplicable a un amplio espectro de servicios de recursos georreferenciados en la red permitiendo el descubrimiento integrado de diferentes recursos georreferenciados provenientes de diferentes fuentes. En la implementación de *GeoCrawler*, *Solr* cubrirá la funcionalidad de la fase de descubrimiento. En base al índice generado por *Nutch*, *Solr* ofrece diferentes interfaces de búsqueda a través de las cuales el usuario podrá realizar sus consultas. Además, en base a nuestra contribución, se le ha añadido soporte para la interfaz de búsqueda *OpenSearch* y su extensión *OpenSearch-Geo*. Por lo que, a partir de este punto, cualquier tipo de usuario desde cualquier tipo de dispositivo debería ser capaz de realizar búsquedas homogéneas sobre los recursos indexados a través de aplicaciones cliente como VisioMIMEXT (ver Sección 4.5).

Además de ofrecer una interfaz de búsqueda común que permita a los usuarios expresar sus necesidades de información en forma de consultas, es necesario anotar los recursos resultantes de forma homogénea con la información pertinente. Este problema ha sido solucionado mediante la adopción de MIMEXT (ver Sección 2.3) como formato de respuesta a las consultas, gracias este mecanismo es posible anotar y georreferenciar los recursos independientemente de su naturaleza.

Finalmente, el problema de la visualización e integración de datos heterogéneos ha sido solucionado gracias a herramientas basadas en

globos virtuales (ver Sección 4.4). Tanto esta contribución como la interfaz de consulta común y homogénea propuesta (ver Sección 2.2.3), junto con el mecanismo para la anotación de recursos MIMEXT (ver Sección 2.3), han sido comprobadas y demostradas mediante el desarrollo de la aplicación VisioMIMEXT (ver Sección 4.5) basada en un globo virtual, que integra mecanismos de búsqueda y visualización de forma natural sobre el globo virtual y que servirá como un primer cliente del sistema *GeoCrawler*.

De forma resumida, el sistema es capaz de recopilar todos los recursos interesantes existentes en el ámbito en el que se ejecute (máquina local, intranet, red abierta). Y, tras obtener las descripciones a través de la plataforma de generación de metadatos, es capaz de generar los índices textuales y espaciales en base a ellas. Gracias a estos índices, mediante las interfaces de usuario y los algoritmos de búsqueda del proyecto *Solr*, las consultas ejecutadas por los usuarios podrán filtrar los resultados de acuerdo a las características específicas que tiene la información geográfica y especialmente por su localización. Además, se ha añadido soporte para la interfaz de búsqueda *OpenSearch* y su extensión *OpenSearch-Geo* de forma que los usuarios podrán realizar sus consultas a través de un mecanismo de búsqueda común y homogéneo.

Este sistema, dada la naturaleza espacial de los recursos, al combinar índices textuales y espaciales aporta un gran valor añadido a las búsquedas de recursos georreferenciados al poder restringir los resultados en un contexto geográfico.

Gracias a los conceptos y las contribuciones que se han incorporado, el sistema proporciona una plataforma de búsqueda potente y eficiente con capacidades de búsqueda avanzadas. Además, gracias a la arquitectura propuesta y a la implementación, se trata de un sistema escalable, flexible y adaptable mediante la configuración en archivos XML y extensible gracias a su arquitectura basada en *plugins*, por lo que puede ser adaptado a cualquier caso de uso de forma específica. Otro de los puntos fuertes es que sus interfaces se basan en estándares abiertos y podemos realizar consultas mediante interfaces tipo REST o HTTP y recibir las respuestas en formatos como XML, JSON o MIMEXT gracias al componente implementado, que hacen que sea fácil de usar desde virtualmente cualquier lenguaje de programación. Por todo ello resulta

relativamente sencillo ofrecer aplicaciones cliente con interfaces de usuario más amigables y visuales para realizar consultas sobre la plataforma de búsqueda del sistema. Como vimos en la Sección 4.4, para consultas de carácter geoespacial es especialmente interesante poder obtener el contexto geográfico en el que se va a restringir la consulta desde un visualizador de mapas como el que ofrece el proyecto *OpenLayers*[183] para clientes Web o desde un globo virtual como el que integra la aplicación cliente VisioMIMEXT.

[183] http://openlayers.org

6. Conclusiones

Este capítulo resume las ideas y aportaciones más importantes alcanzadas en el trabajo de investigación que abarca esta tesis. En primer lugar se presentan las conclusiones, indicando las principales aportaciones del trabajo realizado y, en segundo lugar, se discutirán las limitaciones y las posibles líneas de trabajo futuro.

6.1 Aportaciones

Actualmente, a pesar del gran número y variedad de recursos georreferenciados disponibles y a pesar de los esfuerzos realizados para crear grandes catálogos de metadatos aún resulta difícil encontrar contenidos georreferenciados relevantes de forma integrada. Tratando de abordar el problema, se ha propuesto un *workflow* adaptado a las necesidades de la IG que cubre todo su ciclo de vida, empezando cuando los recursos son puestos disponibles por los creadores y finalizando cuando son descubiertos y explotados por los consumidores. Este *workflow* consigue que los metadatos, pese a su papel fundamental, sean transparentes al usuario ya que automatiza su proceso de creación y publicación. En este sentido, el trabajo de investigación presentado en esta tesis, tiene como objetivo la recopilación de los recursos disponibles, su descripción a través de la generación automática de metadatos, la organización y publicación de los recursos en base a sus descripciones y, finalmente, ofrecer

estrategias de descubrimiento integradas y sencillas desde el punto de vista del usuario.

En esta sección, se resumen las conclusiones y las principales aportaciones de esta tesis. De igual modo que la tesis, esta sección seguirá la estructura de las fases básicas del *workflow* propuesto.

Descripción

Las descripciones de los recursos son críticas para permitir y mejorar su descubrimiento, acceso y explotación. Por lo tanto, son el elemento clave para conseguir una buena integración de los datos y el buen funcionamiento de cualquier SI, visto como la arquitectura básica para compartir, descubrir y usar recursos heterogéneos. Con el fin de describir los recursos, se han presentado los metadatos y se han analizado diferentes aspectos respecto a ellos, como sus objetivos, su funcionalidad o su relevancia en el contexto de la IG.

Dada la necesidad de facilitar la descripción de los recursos, se ha propuesto una metodología para la generación de metadatos que pretende automatizar su producción, con una mínima intervención de usuario. Gracias a ella, se evitan errores y se libera a los creadores del arduo y monótono trabajo de crear los metadatos que suele derivar en una baja calidad y/o falta de ellos. Esta metodología, permitirá mejorar progresivamente la generación automática de metadatos y la calidad resultante de éstos según se vayan aplicando y mejorando los diferentes métodos que la componen. Además, la metodología propuesta tiene en cuenta e intenta recopilar información, como la que proviene del proceso de creación, que actualmente pasa desapercibida pero puede ser muy relevante. En consecuencia, el resultado de aplicar esta metodología será obtener más metadatos, de mejor calidad y corrección, más completos y con reducida participación por parte del usuario.

Esta metodología ha sido aplicada con éxito en dos casos de estudio reales, donde se ha implementado parte de ella. En primer lugar el prototipo del gestor de metadatos de gvSIG es capaz de manejar metadatos tanto para uso interno de la aplicación, aumentado de la eficiencia en el flujo de trabajo del usuario, como para su publicación permitiendo compartir los recursos en una IDE. Esta implementación ha servido como una primera prueba de concepto de la metodología para la creación de metadatos propuesta ofreciendo

nuevas técnicas de extracción automática de metadatos que permiten disminuir el esfuerzo requerido por parte del usuario para describir y documentar los recursos para, posteriormente, ser compartidos y reutilizados. Además, el prototipo permite la recolección de información durante el proceso de creación de los datos, esta es una fuente de información volátil dado que sólo estará disponible en el momento en el que el recurso es creado y por esa razón es importante obtener y almacenar toda la información posible en ese momento. En el segundo caso de uso, la aplicación de la metodología para la generación de metadatos propuesta, ha permitido desarrollar una plataforma que permite obtener información y acceder de forma homogénea a recursos heterogéneos en base a la integración de los proyectos *OSGeo FDO* y *Apache Tika*. Esta plataforma puede ser un componente básico para todas las aplicaciones que requieran generar metadatos, ya que da soporte para describir recursos multimedia de una gran variedad de formatos, especialmente de formatos de IG.

Por otra parte, se requiere un mecanismo genérico que permita anotar cualquier tipo de recurso y se ha propuesto como solución una extensión de KML llamada MIMEXT. Esta extensión permite georreferenciar recursos heterogéneos sin depender del tipo de datos o formato. MIMEXT se basa en técnicas de anotación externa, por lo que no modifica la estructura interna de los formatos de los recursos. Esto permite que los recursos representados tanto en formatos actuales como futuros puedan ser descritos usando MIMEXT. No obstante, la extensión de KML propuesta representa una solución simple y no implica la creación de un nuevo lenguaje para describir los diferentes tipos de recursos. De hecho, esta solución está impulsada por el principio de reutilización, y tiene en cuenta determinados aspectos tales como la interoperabilidad con otras tecnologías y estándares, facilidad de uso, o el grado de asimilación por parte del mundo académico, la industria y los usuarios en general. Además, gracias al uso de archivos KMZ, que permiten encapsular en un mismo archivo el recurso, su información de geolocalización junto con otros metadatos en formato MIMEXT y otros recursos relacionados, se consigue componer una única unidad lógica y física que facilita la correcta administración, actualización y sincronización de datos y metadatos.

Finalmente, para facilitar la creación y distribución de metadatos se ha propuesto un nuevo paradigma llamado *Description as a Service*

(DESCaaS). El objetivo de este paradigma es proporcionar descripciones de recursos homogéneas *on-line*, con el fin de mejorar la accesibilidad, interoperabilidad y descubrimiento de los recursos. A través de estas descripciones, se puede obtener información detallada acerca del recurso y de su contenido. Ofreciendo descripciones, DESCaaS está cubriendo la falta de mecanismos que permitan a los usuarios publicar, encontrar y acceder a recursos distribuidos de forma eficiente. La principal ventaja de DESCaaS es que mejora el descubrimiento de los recursos pero, además, permite la publicación y reutilización de procesos de generación de metadatos, permite interactuar con otros servicios (encadenamiento), promociona la interoperabilidad a través de descripciones estándar y permite añadir anotaciones semánticas a los recursos. Dada la definición abstracta y general de DESCaaS, muchos casos de uso diferentes pueden ajustarse al paradigma. Su validez y beneficios han sido comprobados a través de un caso de uso real dentro del proyecto ENVISION, donde se ha implementado un servicio DESCaaS que proporciona descripciones NcML para recursos en formato NetCDF.

Publicación

La publicación de los recursos de acuerdo a sus descripciones permite compartirlos y facilita su posterior descubrimiento, de forma que estos pueden actualizarse, corregirse, modificarse y reutilizarse por parte de usuarios con intereses similares, evitando duplicar esfuerzos tanto físicos como económicos. Por ello, persiguiendo la interoperabilidad técnica, sintáctica y semántica entre sistemas, se han revisado diferentes estrategias de publicación de recursos georreferenciados y diferentes estándares que especifican la información que deben contener los metadatos.

Los catálogos, a través del estándar CSW, son el método de publicación más utilizado actualmente en el contexto de la IG. Por lo general, los recursos son registrados en los catálogos por los usuarios que introducen sus metadatos manualmente con el fin de ponerlos a disposición de los demás. Tras ello, los metadatos son procesados y almacenados en la aplicación a la espera de ser recuperados. Este proceso implica la inmediata publicación y accesibilidad de los recursos, pero también implica interacción humana y control durante el proceso.

En base a esta estrategia, se propone una metodología para publicar los recursos en servicios de catálogo de forma integrada en el flujo de trabajo y como demostración se han presentado dos casos de estudio reales en los que se han implementado sendas soluciones para la publicación de metadatos en catálogos a través del estándar CSW y su perfil transaccional CSW-T. En primer lugar el prototipo de la extensión de metadatos de gvSIG pretende vincular, de forma integrada en el flujo de trabajo, la producción de recursos con la de metadatos y con la publicación de estos últimos en un servicio de catálogo. Consiguiendo aumentar la cantidad de metadatos publicados, ya que se reduce drásticamente el esfuerzo necesario para describir correctamente los recursos y publicarlos. De esta forma, utilizando esta solución integrada, el usuario puede cerrar el ciclo de vida de los metadatos dentro de la misma aplicación. Puede crear, modificar y publicar los metadatos utilizando el prototipo, y más tarde descubrir y recuperar los metadatos utilizando el cliente de catálogo integrado en gvSIG que, además, permite recuperar los recursos vinculados. En segundo lugar, en el contexto del proyecto ENVISION, se ha presentado un mecanismo para la publicación en servicios de catálogo de descripciones anotadas semánticamente. Al añadir semántica a las descripciones de los recursos, estos pueden ser descubiertos en base a consultas sobre el significado real del contenido de los recursos, mejorando su descubrimiento. Además el enfoque basado en estándares OGC, garantiza la interoperabilidad y la integración con las soluciones ya existentes.

En la actualidad, gracias a las redes sociales y a la proliferación de dispositivos de posicionamiento, usuarios no expertos crean y comparten grandes cantidades de recursos georreferenciados, una tarea reservada anteriormente a los profesionales. Los catálogos son herramientas muy útiles y potentes para el descubrimiento de contenido geográfico. Sin embargo, su uso está más centrado en los usuarios profesionales ya que requieren un cierto grado de conocimiento y/o experiencia para la creación de los metadatos y su publicación. Estos factores limitan el uso de catálogos como soluciones eficaces para gestionar la enorme cantidad de nuevos recursos. Con el fin de proponer soluciones más automatizadas y con la vista puesta en el descubrimiento de los recursos, se propone el uso de técnicas de indexación, abstrayendo esta funcionalidad de los catálogos, de forma que sea posible descubrir los recursos usando

conjuntos de metadatos simples, en lugar de crear formatos estándar complejos como los que se almacenan en los catálogos. Por ello, se ha realizado un análisis de distintos tipos de indexación sobre metadatos, así como la integración de estos índices dentro de un sistema de recuperación de información. La mayor parte de los metadatos que describen los recursos pueden ser considerados como texto, los índices textuales permiten realizar consultas basadas en términos o palabras clave de forma eficiente sobre este contenido. Por otra parte, en este contexto resulta muy interesante poder recuperar los recursos teniendo en cuenta su localización, los índices espaciales permiten acelerar la búsqueda y devolución de los recursos respecto a este tipo de consultas. Finalmente, para satisfacer las necesidades de información de los usuarios se propone combinar ambos tipos de índices integrando los índices espaciales sobre los índices textuales, permitiendo de este modo la resolución conjunta y eficiente de operaciones espaciales y textuales.

Descubrimiento

La finalidad de describir y publicar los recursos no es otra que permitir y facilitar su descubrimiento, nuestro objetivo final. El descubrimiento implica encontrar los recursos relevantes de acuerdo a nuestras necesidades de información. Por ello, se han revisado diferentes estrategias actuales para el descubrimiento de recursos georreferenciados y otros factores que influyen en el proceso de descubrimiento como la calidad de los metadatos que describen los recursos.

Cualquier sistema de información necesitará previamente descubrir los recursos que incorporará. Desde el punto de vista del sistema, existen dos opciones para recopilar los recursos: o los proporciona el usuario directamente al sistema (catálogos) o se automatiza su recolección (*crawlers*). La principal ventaja de la recopilación automatizada de recursos es que no requiere intervención por parte del usuario que, simplemente, debe dejarlos accesibles. Sin embargo, los recursos no son inmediatamente publicados como sucede con los catálogos, hay que esperar a que el *crawler* los descubra. Además, dado que no se tiene el control del proceso existe cierta incertidumbre sobre el éxito y la corrección del mismo, pero de este modo el sistema es menos susceptible a engaño ya que los recursos no son clasificados de acuerdo a los deseos de sus creadores. Comprobada

la validez de la recopilación automatizada de recursos mediante *crawlers* en el mundo Web se propone seguir su línea de evolución. Este método representa una solución eficaz para organizar todos aquellos recursos creados masivamente por los usuarios y que no encajan en el proceso de catalogación tradicional.

Los usuarios expresan sus necesidades de información en forma de consultas a través de las interfaces de búsqueda que ofrecen los servicios de descubrimiento. En el contexto de la IG la única interfaz de búsqueda especificada formalmente es el estándar CSW que permite realizar consultas basadas en palabras clave y/o filtros espaciales sobre los metadatos publicados en catálogos. Sin embargo, se requiere una interfaz de búsqueda uniforme y homogénea para aumentar la visibilidad de los contenidos generados por usuarios y publicados en diferentes servicios y redes sociales. *OpenSearch* es un método simple y estándar para buscar recursos disponibles en la red. Además, gracias a su extensión *OpenSearch-Geo* podemos realizar consultas filtrando los resultados en base a un contexto espacial. El principal beneficio de *OpenSearch* es que proporciona un mecanismo de búsqueda común y homogéneo para realizar consultas a través de HTTP, mientras que es lo suficientemente flexible como para adaptarse a recursos de diversa naturaleza. Al mantener una interfaz de búsqueda básica pero bien definida fomenta el desacoplamiento entre los clientes y los servicios lo que mejora enormemente la escalabilidad de los sistemas. Así, del mismo modo que se propone MIMEXT como una forma común para georreferenciar recursos independientemente de su naturaleza, se ha optado por *OpenSearch* como interfaz de consulta común, aplicable a un amplio espectro de servicios de recursos georreferenciados en la red.

La visualización se ha vuelto fundamental en el manejo y distribución actual de información. La visualización de los recursos permite examinarlos y comprenderlos de una forma más sencilla. Por ello, la visualización juega un papel importante en el descubrimiento de recursos. Centrándonos en el contexto de la IG, dada la naturaleza de este tipo de recursos es conveniente representarlos y visualizarlos en entornos 3D. En este sentido, los globos virtuales resultan una herramienta muy útil permitiendo la visualización e integración de datos tanto para usuarios expertos como para usuarios no expertos.

VisioMIMEXT es una prueba de concepto que integra la interfaz de búsqueda *OpenSearch*, complementada con el formato MIMEXT para georreferenciar y posteriormente visualizar los resultados, sobre un globo virtual. VisioMIMEXT demuestra cómo a través de *OpenSearch* los usuarios son capaces de descubrir recursos formulando consultas uniformes sobre diferentes fuentes y restringiendo los resultados a un contexto geográfico, de forma que mejora el descubrimiento. El descubrimiento de recursos a través de una interfaz de usuario basada en un globo virtual resulta más fácil e intuitivo desde el punto de vista de los usuarios, dado que los mecanismos de búsqueda y visualización se integran de forma natural en el mismo globo virtual.

Solución Integrada

Las contribuciones de esta tesis no estarían completas sin el desarrollo de un prototipo de un sistema completo de indexación y búsqueda de recursos georreferenciados basado en el *workflow* y las soluciones propuestas. *GeoCrawler* ha sido desarrollado utilizando software libre, adaptando los componentes disponibles siempre que ha sido posible, y desarrollando nuevas alternativas cuando no lo fue o cuando los componentes existentes no cumplían los requisitos exigidos. El principal objetivo de *GeoCrawler* es mejorar el descubrimiento de recursos georreferenciados y, consecuentemente, su accesibilidad. Para ello, de forma sinérgica, maneja diferentes tipos de recursos georreferenciados, diferentes métodos de generación de metadatos y diferentes estrategias de publicación y descubrimiento. De forma que recopilará, describirá y publicará automáticamente todos los recursos que encuentre disponibles, facilitando su posterior descubrimiento desde aplicaciones cliente.

6.2 Limitaciones y Líneas de Trabajo Futuro

El trabajo desarrollado en esta tesis representa tan solo una primera aproximación para el desarrollo de un sistema de indexación y búsqueda de recursos georreferenciados y algunos aspectos relacionados. Existe una gran variedad de puntos y direcciones que todavía quedan por explorar y resolver en el ámbito del descubrimiento de recursos georreferenciados. Por otra parte, para conocer el verdadero alcance de una tesis, resulta interesante conocer

tanto sus contribuciones como sus limitaciones. Por ello, en esta sección se discuten algunas limitaciones del trabajo presentado y posibles extensiones del mismo.

Del mismo modo que la sección anterior, vamos a organizar las limitaciones y las líneas de trabajo futuro siguiendo las fases básicas del *workflow* propuesto.

Descripción

Se ha propuesto una completa metodología para la generación de metadatos que pretende automatizar su producción para facilitar la descripción de los recursos. Sin embargo, sólo se ha implementado parte de ella. Con la implementación actual se generan un buen conjunto de metadatos, y en base a estos es posible aplicar el resto de métodos deductivos para mejorar y completar las descripciones. Para inferir nuevos metadatos, tenemos pensado explorar el uso de diferentes técnicas de razonamiento y algoritmos de minería de datos para determinar su aplicabilidad en este contexto. Además, también tenemos planeado explorar cómo generar de forma automática el texto que contienen los elementos de metadatos menos intuitivos como el título o el resumen. Asimismo, otra de las posibles mejoras es utilizar ontologías para la generación de metadatos, algunos autores [2] han conseguido resultados prometedores aplicando este tipo de técnicas sobre conjuntos de recursos homogéneos.

Por otra parte, DESCaaS supone una gran mejora para obtener descripciones de recursos al permitir la publicación y reutilización de los procesos de generación de metadatos. Sin embargo, los procesos de descripción servidos por DESCaaS todavía tienen que abordar los principales problemas de generación de metadatos para obtener una descripción adecuada, es decir la falta de información sobre algunos recursos y manejar la enorme variedad de formatos de datos y estándares.

Finalmente, creemos que resulta esencial impulsar la investigación en todos los campos relacionados con la generación y gestión de metadatos dado el papel fundamental que estos juegan en cualquier SI. Conseguir más metadatos y de mejor calidad permite indexar o catalogar los recursos de una forma más exacta y, en consecuencia, es posible proporcionar resultados de búsqueda más relevantes y precisos.

Publicación

En esta tesis se han presentado dos aproximaciones para publicar recursos georreferenciados en servicios de catálogo. Las actuales implementaciones de catálogo proporcionan una infraestructura sólida para publicar y descubrir recursos georreferenciados basándose en los metadatos que los describen. Pero, en algunos casos, la información contenida en los metadatos es demasiado genérica, no suficientemente flexible, dependiente del idioma y en su mayoría basada en los aspectos sintácticos del servicio. Añadiendo semántica a las descripciones de los recursos mediante anotaciones, los recursos pueden ser descubiertos a través de consultas sobre el significado real del recurso y su contenido. La aproximación presentada en la Sección 3.2.3 permite publicar estas anotaciones semánticas, sin embargo esta aproximación se encuentra limitada ya que las anotaciones semánticas de los recursos tienen que ser realizadas de forma manual por parte de usuarios expertos lo que resulta tedioso y costoso. Un objetivo futuro debería ser anotar semánticamente de forma automática los recursos. Como ejemplo, en [192] se propone automatizar la anotación semántica de servicios usando vocabularios como *DBpedia*[184] o *GeoNames*[185].

Por otra parte, queremos explorar otras estrategias para publicar metadatos y recursos intentando mejorar su capacidad para ser descubiertos. Una primera estrategia puede ser publicar los recursos directamente en servicios o redes sociales específicas para el tipo del recurso, usando sus metadatos para documentarlo adecuadamente. Por ejemplo, podemos publicar mapas en *MapServer*, imágenes en *Flickr* o videos en *YouTube*. Otra opción es poner los recursos disponibles para que los buscadores generales de Internet como *Google*, *Yahoo!* o *Bing* los encuentren e indexen. En este caso debemos tener en cuenta los formatos que soporta cada uno de ellos, dejando simplemente el recurso disponible o construyendo una descripción (textual, KML, etc.) para que sea indexada. Otro método que podemos explorar es el uso de redes *peer-to-peer*[186] (P2P) [195]

[184] http://dbpedia.org
[185] http://www.geonames.org
[186] Una red *peer-to-peer* es una red de computadoras que funciona sin clientes ni servidores fijos, sino mediante una serie de nodos que se comportan como iguales entre sí.

[7] para compartir recursos dentro de un organización o de forma global en redes abiertas. La idea es explorar diferentes alternativas de publicación en combinación con varios niveles de generación de metadatos para evaluar la capacidad para ser descubiertos de los recursos publicados.

Descubrimiento

En este aspecto, nuestros planes futuros incluyen aumentar el espectro de criterios de búsqueda para aprovechar el poder de *OpenSearch*. Las dimensiones temporales y semánticas deben ser consideradas en el momento de formular la consulta, con el fin de restringir los resultados a un cierto período de tiempo y a términos específicos de una taxonomía.

Por otra parte, es importante para la próxima versión de VisioMIMEXT gestionar y mejorar la precisión de los resultados de búsqueda, de modo que el sistema sea capaz de identificar, en cierta medida, falsos positivos.

Finalmente, queremos explorar cómo involucrar a los usuarios en la descripción y publicación de recursos georreferenciados. La participación de los usuarios, con sus aportaciones y valoraciones, es de gran valor para cualquier SI. De hecho esta es la filosofía de la Web 2.0 que tanto éxito ha tenido. Por ello, tendría sentido en ciertos escenarios permitir a los usuarios catalogar o clasificar los recursos asignándoles etiquetas o valorándolos según su opinión. En relación con esto, también sería posible anotar los recursos obtenidos como resultado de una consulta y que han resultado más interesantes para el usuario con las palabras clave de la consulta que formuló. En este sentido existen algunos trabajos relacionados, como el proyecto *GeoNetwork* que recoge estadísticas analizando las consultas que se realizan en el catálogo, o [122] que propone el enriquecimiento automático de metadatos espaciales. Por último, la aplicación de técnicas de *gamification*[187] constituye posiblemente la línea de trabajo futuro más prometedora. Creemos que es interesante aplicar este tipo de técnicas [84] para conseguir que los usuarios compartan sus recursos y los describan de forma apropiada.

[187] Las técnicas de *gamification* se basan en el uso de la mecánica de jugabilidad en contextos ajenos a los juegos, con el fin de que las personas adopten cierto comportamiento.

Bibliografía

[1] Abargues C, Granell C, Diaz L, Huerta J, Beltran A. 2009. Discovery of User-Generated Geographic Data Using Web Search Engines, Advances in Geoscience and Remote Sensing, Gary Jedlovec (Ed.), ISBN: 978-953-307-005-6, InTech.

[2] Abugessaisa I. 2010. Geospatial metadata extraction from product description document applying methods from ontology engineering. Int. J. Metadata, Semantics and Ontologies, Vol. 5, No. 4, pp.321–332.

[3] Adobe. 2007. Extensible metadata platform (XMP). http://www.adobe.com/products/xmp/overview.html

[4] Al-Masri E, Mahmoud Q. 2008. Discovering Web Services in search engines. IEEE Internet Computing. 12, 3, May-Jun 2008, 74-77

[5] Anderson J, Pérez J. 2001. The nature of indexing: how humans and machines analyze messages and texts for retrieval: part I. Information Processing and Management: An International Journal, Vol.37 n.2, pp. 231–254

[6] Anguix A, Díaz L. 2008. gvSIG: A GIS desktop solution for an open SDI. Journal of Geography and Regional Planning Vol. 1(3), May, 2008, pp. 041-048.

[7] Antoniadis P, LeGrand B. 2007. Incentives for resource sharing in self-organized communities: From economics to social psychology. In Digital Information Management, 2007. ICDIM '07

[8] ANZLIC. 1996. ANZLIC Guidelines: Core Metadata Elements Version 1 Report: ANZLIC Working Group on Metadata, July 1996.

[9] Asociación Española de Normalización y Certificación (AENOR). 1998. Mecanismo de Intercambio de Información Geográfica Relacional formado por Agregación (MIGRA): UNE 148001 EXP 1998, versión 1. Comité Técnico de Normalización 148 de AENOR (AEN/CTN 148)

[10] Baca M. 2008. Introduction to Metadata: Pathways to Digital Information (version 3.0). In Getty Research Institute.

[11] Baeza R, Ribeiro B. 1999. Modern information retrieval (Vol. 463). New York.: ACM press.

[12] Balfanz D. 2002. Automated Geodata Analysis and Metadata Generation. Society of Photo-Optical Instrumentation Engineers -SPIE-, Bellingham/Wash. Visualization and Data Analysis

[13] Ballagh L, Raup B, Duerr R, Khalsa S, Helm C, Fowler D, Gupte A. 2011. Representing scientific data sets in KML: methods and challenges. Comput Geosci 37(1):57–64

[14] Barton J, Currier S, Hey J. 2003. Building quality assurance into metadata creation: An analysis based on the learning objects and e-prints communities of practice. Proceedings of Dublin Core Conference 2003. Metadata Research and Applications, 2003, (pp. 39-48).

[15] Batcheller J. 2008. Automating geospatial metadata generation – An integrated data management and documentation approach, Computers & Geosciences, 34: 287-398.

[16] Beard K. 1996. A Structure for Organizing Metadata Collection. Third International Conference/ Workshop on Integrating GIS and Environmental Modelling, January 21-25

[17] Belimpasakis P, Saaranen A. 2010. Sharing with people: a system for user-centric content sharing. Multimed Syst 16:399–421

[18] Beltran A, Abargues C, Granell C, Núñez M, Díaz L, Huerta J. 2012. A virtual globe tool for searching and visualizing geo-referenced media resources in social networks. International Journal on Multimedia Tools and Applications. Springer Netherlands, 2012. ISSN: 1380-7501. DOI: 10.1007/s11042-012-1025-0

[19] Beltran A, Granell C, Huerta J. 2011. Descripción de recursos multimedia georreferenciados. Proceedings of V Jornadas de SIG Libre (SIG Libre 2011). Girona, Spain, Mar 2011. ISBN: 978-84-694-1624-2.

[20] Bernard L, Kanellopoulos I, Annoni A, Smits P. 2005. The European Geoportal – one step towards the establishment of a European Spatial Data Infrastructure. Computers, Environment and Urban Systems, 29, 1, 15-31

[21] Berners-Lee T, Hendler J, Lassila O. 2001. The Semantic Web. Scientific American.

[22] Bishop I, Escobar F, Karuppannan S, Suwarnarat K, Williamson I, Yates P, Yaqub H. 2000. Spatial data infrastructures for cities in developing countries: Lessons from the Bangkok experience. Cities 17(2) 85-96.

[23] Bodoff D. 2006. Relevance for browsing, relevance for searching. J. Am. Soc. Inf. Sci., 57: 69–86. doi: 10.1002/asi.20254

[24] Boldi P, Vigna S. 2005. MG4J at TREC 2005. In Ellen M. Voorhees y Lori P. Buckland, editors, The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, number SP 500-266 in Special Publications. NIST, 2005

[25] Boll S. 2007. MultiTube–where multimedia and Web 2.0 could meet. IEEE Multimed 14(1):9–13

[26] Boll S, Klas W, Sheth A. 1998. Overview on Using Metadata to Manage Multimedia Data, in Sheth, A., Klas, W., editors (1998) Multimedia Data

Management, Using Metadata to Integrate and Apply Digital Media, McGraw-Hill, New York, USA.

[27] Borenstein N, Freed N. 1993. MIME (Multipurpose internet mail extensions) Part One: mechanisms forspecifying and describing the format of internet message bodies. RFC 1521, http://tools.ietf.org/html/rfc1521

[28] Boutell M, Luo J. 2005. Beyond pixels: Exploiting camera metadata for photo classification. In Pattern Recognition, v. 38, n. 6, pp. 935-946.

[29] Bray T, Paoli J, Sperberg-McQueen C, Maler E. 2000. Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation 6 October 2000. http://www.w3.org/TR/2000/REC-xml-20001006.

[30] Bruce T, Hillmann D. 2004. The continuum of metadata quality: Defining, expressing, exploiting. In Hillmann D., Westbrooks E. (Eds.), Metadata in practice, (pp. 238-256). Chicago: ALA.

[31] Bui Y, Jung P. 2006. An assessment of metadata quality: A case study of the national science digital library metadata repository. In Haidar Moukdad (ed.), CAIS/ACSI 2006 Information Science Revisited: Approaches to Innovation

[32] Butler D. 2006. Virtual globes: the web-wide world. Nature 439:776–778

[33] Bulterman D, Rutledge L. 2008. SMIL 3.0: Interactive Multimedia for the Web, Mobile Devices and Daisy Talking Books. Hardcover ISBN: 978-3-540-78546-0, November, 2008

[34] Bultermann D. 2004. Is It Time for a Moratorium on Metadata? IEEE Multimedia, 11(4):10-17, IEEE Computer Society Press, Los Alamitos, Ca, USA

[35] Campbell T. 2008. Fostering a Culture of Metadata Production. GSDI10: Tenth International Conference for Spatial Data Infrastructure, St. Augustine, Trinidad February 25-29, 2008. http://www.gsdi.org/gsdiconf/gsdi10/papers/TS8.2paper.pdf

[36] Caplan P. 2003. Metadata Fundamentals for All Librarians. Chicago: American Library Association.

[37] Caplan P. 1995. You call it corn, we call it syntax-independent metadata for document-like objects. The Public Access Computer Systems Review, v. 4, n. 6, 1995.

[38] Carney D, Smith J, Place P. 2005. Topics in Interoperability: Infrastructure Replacement in a System of Systems (CMU/SEI-2005-TN-031). Pittsburgh, Pa: Software Engineering Institute, Carnegie Mellon University, November 2005. Web document. http://www.sei.cmu.edu/reports/05tn031.pdf

[39] Cerda D. 2005. El mundo según Google: Google Earth y la creación del dispositivo GeoSemántico global. http://geosemantica.gearth.googlepages.com

[40] CGIAR-CSI. 2004. Metadata Tips: Why Metadata?. http://www.csi.cgiar.org/metadata/Metadata_Why.asp

[41] Chellappa R, Wilson C, Sirohey S. 1995. Human and machine recognition of faces: a survey. Proceedings of IEEE, vol, 83, num 5, 1995, 705-741.

[42] Chen Y, Suel T, Markowetz A. 2006. Efficient query processing in geographic web search engines. En SIGMOD'06: Proc. of the ACM SIGMOD Conference, pp. 277–288, 2006.

[43] Chien N, Tan S. 2011. Google Earth as a tool in 2-D hydrodynamic modelling. Comput Geosci 37 (1):38–46

[44] Clinton D. 2012. OpenSearch 1.1 Draft 5 specification. http://www.opensearch.org

[45] Conrad C. 2006. The Global Earth Observation System of Systems: Science Serving Society, Space Policy, Volume 22, Issue 1, February 2006, Pages 8-11, ISSN 0265-9646, 10.1016/j.spacepol.2005.12.004.

[46] Consorcio España Virtual. 2009. E.1.2.3 Informe de avance en la investigación en Metadatos. Entregable del Proyecto CENIT España Virtual, 2009. GeoSpatiumLab, CNIG e Indra Espacio.

[47] Consorcio España Virtual. 2010. E.1.2.5 Informe de avance en la investigación en Metadatos. Entregable del Proyecto CENIT España Virtual, 2009. GeoSpatiumLab, CNIG e Indra Espacio.

[48] Consorcio España Virtual. 2009. E.3.4.3 Informe de avance en la investigación en Tecnologías de Indexación y Búsqueda. Entregable Proyecto CENIT España Virtual, 2009.

[49] Craglia M, Goodchild M. 2008. Next-generation digital Earth: a position paper from the Vespucci initiative for the advancement of geographic information science. Intern J Spatial Data Infra Res 3:146–167

[50] Craglia M, Kanellopoulos I, Smits P. 2007. Metadata: where we are now, and where we should be going. Proceedings of 10th AGILE International Conference on Geographic Information Science 2007. Aalborg University, Denmark

[51] Craven T. 2001. Description meta tags in public home and linked pages. Library and Information Science Research Electronic Journal vol.11 (2)

[52] Croft W, Metzler D, Strohman T. 2010. Search engines: Information retrieval in practice (p. 283). Addison-Wesley.

[53] Currier S, Barton J, O'Beirne R, Ryan B. 2004. Quality assurance for digital learning object repositories: Issues for the metadata creation process. ALT-J, Research in Learning Technology, 12(1), 5-20.

[54] Custard M. 2005. Using Machine Learning to Support Quality Judgements. vol. 11 (10) ISSN 1082-9873, 2005

[55] Danko D. 2002. ISO/TC 211 Geographic information/Geomatics The Standards in Action Workshop in Gyeongju, Korea: Implementation. Web document.

http://www.isotc211.org/WorkshopGyeongju/Presentations/Metadata.ppt

[56] Davis C, Fonseca F, Câmara G. 2009. Beyond SDI: Integrating Science and Communities to Create Environmental Policies for the Amazon. International Journal of Spatial Data Infrastructures Research, 4, 156-174.

[57] Day M, Tzong-Han T, Sung C, Hsieh C, Lee C, Wu S, Wu K, Ong C, Hsu W. 2007. Reference metadata extraction using a hierarchical knowledge representation framework. In Decision Support Systems, v. 43, pp. 152–167.

[58] DePaor D, Whitmeyer J. 2011. Geological and geophysical modeling on virtual globes using KML, COLLADA, and Javascript. Comput Geosci 37(1):100–110

[59] Devillers R, Gervais M, Bédard Y, Jeansoulin R. 2002. Spatial data quality: from metadata to quality indicators and contextual end-user manual. In OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management (pp. 21-22)

[60] Díaz L, Gould M, Beltran A, Llaves A, Granell C. 2008. Multipurpose Metadata Management in gvSIG. Proceedings of the academic track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, 29 September – 3 October 2008, Cape Town, South Africa. ISBN: 978-0-620-42117-1, pp 90-99

[61] Díaz L, Granell C, Beltran A, Llaves A, Gould M. 2008a. Extracción Semiautomática de Metadatos: Hacia los metadatos implícitos. II Jornada de SIG Libre. Universidad de Girona.

[62] Díaz L, Martín C, Gould M, Granell C, Manso M. 2007. Semi-automatic Metadata Extraction from Imagery and Cartographic data, International Geoscience and Remote Sensing Symposium (IGARSS 2007). Barcelona, Julio 2007. IEEE CS Press, pp. 3051-3052.

[63] Diehl S, Neuvel J, Zlatanova S, Scholten H. 2006. Investigation of user requirements in emergency response sector: the Dutch case. Second Symposium on Gi4DM, 25-26 September, Goa, India.

[64] Dong Z. 2010. Automated Extraction and Retrieval of Metadata by Data Mining - A Case Study of Mining Engine for National Land Survey Sweden. Master Thesis in Geomatics, Department of Technology and Built Environment, Universtity of GÄVLE.

[65] Downey D. 2007. What Do Geologists Need to Know about Metadata? http://www.searchanddiscovery.net/documents/2007/07030downey/images/downey.pdf

[66] Durrell W. 1985. Data Administration. A Practical Guide to Data Administration. McGraw-Hill, 1985

[67] Dushay N, Hillmann D. 2003. Analyzing metadata for effective use and re-use. DCMI Metadata Conference and Workshop, Seattle, USA.

[68] Duval E, Hodgins W, Sutton S, Weibel S. 2002. Metadata Principles and Practicalities. D-Lib Magazine. Web document. http://www.dlib.org/dlib/april02/weibel/04weibel.html

[69] ECNBII. 2003. FGDC Biological Data Profile As it maps Dublin Core. GeoConnection. Web document.http://www.geoconnections.org/developersCorner/devCorner_devNetwork/meetings/2003.06.10/Present/ECNBII.ppt

[70] Editorial Nature. 2008. A place for everything. Nature 453(2):2

[71] Ercegovac Z. 1999. Introduction. Journal of the American Society for Information Science, v. 50, n. 13, p. 1165-1168, 1999

[72] European Commission. 2008a. Draft Guidelines – INSPIRE metadata implementing rules based on ISO 19115 and ISO 19119. European Commission.

[73] European Commission. 2008b. Draft implementing measure: REGULATION../…/EC implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata V2. European Commission Commitology Register.

[74] Federal Geographic Data Committee (FGDC). 2011. Geospatial Metadata Factsheet. Web document. http://www.fgdc.gov/library/factsheets/documents/GeospatialMetadata-July2011.pdf

[75] Federal Geographic Data Committee (FGDC). 2000. Content Standard for Digital Geospatial Metadata Workbook, version 2.0. Metadata Ad Hoc Working Group.

[76] Federal Geographic Data Committee (FGDC). 1998. Content Standard for Digital Geospatial Metadata, version 2.0: Document FGDC-STD-001-1998. Federal Geographic Data Committee, Metadata Ad Hoc Working Group.

[77] Fonts O, Huerta J, Díaz L, Granell C. 2009. OpenSearch-geo: the simple standard for geographic search engines. Proceedings IV Jornadas SIG Libre

[78] Forsyth D, Wilensky R. 2003. Research issues for digital libraries. NSF Post-DL Futures Workshop, Chatham, MA, Junio, 2003.

[79] Franklin C, Hane P. 1992. An introduction to GIS: Linking maps to databases. Database, 15(2), 17–22.

[80] Freed N, Borenstein N. 1996a. Multipurpose internet mail extensions (mime) part one: format of internet message bodies, RFC 2045. http://tools.ietf.org/html/rfc2045

[81] Freed N, Borenstein N. 1996b. Multipurpose internet mail extensions (MIME) part two: media types, RFC 2046. http://tools.ietf.org/html/rfc2046

[82] Friesen N. 2004. International LOM Survey: Report (Draft).

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

[84] Garcia I, Rodríguez L, Benedito M, Trilles S, Beltran A, Díaz L, Huerta J. 2012. Mobile Application for Noise Pollution Monitoring through Gamification Techniques. Lecture Notes in Computer Science, 2012, Volume 7522/2012, Entertainment Computing – ICEC 2012, pp. 562-571, DOI: 10.1007/978-3-642-33542-6_74.

[85] Gasser L, Stvilia B. 2001. A new framework for information quality. Technical report, ISRN UIUCLIS-2001/1+AMAS, 2001.

[86] Gayatri, Ramachandran S. 2007. Understanding Metadata. The Icfai Journal of Information Technology, March 2007.

[87] Geller G, Nativi S, Nemani R. 2008. The Model Web: Enhancing model interoperability for ecological forecasting and other disciplines. Geophysical Research Abstracts, Vol. 10, EGU2008-A-01439, 2008, SRef-ID: 1607-7962/gra/EGU2008-A-01439, EGU General Assembly 2008.

[88] Georgiadou Y, Puri S, Sahay S. 2005. Towards a potential research agenda to guide the implementation of Spatial Data Infrastructures - A case study from India. International Journal of Geographical Information Science 19(10): 1113—1130

[89] Global Spatial Data Infrastructure Association (GSDI). 2013. Developing spatial data infrastructures: The SDI cookbook wiki version. http://www.gsdidocs.org/GSDIWiki/index.php/Main_Page

[90] Goodchild M. 2008. Assertion and authority: the science of user-generated geographic content. http://www.geog.ucsb.edu/%7Egood/papers/454.pdf

[91] Goodchild M. 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4): 10. 0343-2521.

[92] Gore A. 1999. The Digital Earth: Understanding our planet in the 21st Century. Photogrammetric Engineering and Remote Sensing 65 (5) 528.

[93] Greenberg J, Spurgin K, Crystal A. 2006. Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. Int. J. Metadata, Semantics and Ontologies, Vol.1, n.1

[94] Greenberg J, Spurgin K, Crystal A. 2005. Final Report for the AMeGA (Automatic Metadata Generation Applications) Project. UNC, School of Information and Library Science University of North Carolina.

[95] Greenberg J. 2004. Metadata extraction and harvesting: a comparison of two automatic metadata generation applications. Journal of Internet Cataloguing. Vol.6 n.4, pp. 59–82.

[96] Greenberg J, Pattuelli M, Parsia B, Davenport W. 2001. Author-generated Dublin Core metadata for web resources: A baseline study in an organization. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2001, (pp. 38–46). National Institute of Informatics.

[97] Grupo de Sistemas de Información Avanzados (IAAA) de la Universidad de Zaragoza. 2010. CatMDEdit User Manual v4.5. http://iaaa.cps.unizar.es/software/index.php/CatMDEdit_English_user_manual

[98] Guttman A. 1984. R-Trees: A Dynamic Index Structure for Spatial Searching. Proc. of SIGMOD'84, ACM Press (1984) 47-57

[99] Guy M, Powell A, Day M. 2004. Improving the Quality of Metadata in Eprint Archives. Ariadne Magazine, (38), 2004.

[100] Hand D, Mannila H, Smyth P. 2001. Principles of Data Mining, Cambridge. The MIT Press.

[101] Harris R, Browning R. 2003. Global Monitoring for Environment and Security: data policy considerations, Space Policy, Volume 19, Issue 4, November 2003, Pages 265-276, ISSN 0265-9646, 10.1016/j.spacepol.2003.08.004.

[102] Hedorfer M, Bianchin A. 1999. The Venice Lagoon Experimental GIS at the IUAV. Interop99: The 2nd International Conference on Interoperating Geographic Information Systems. Web document. http://www.hedorfer.it/docs/rsalv/rsalv1io-ENG.pdf

[103] Herring C. 1994. An Architecture of Cyberspace: Spatialization of the Internet. Champaign, Il: The US Army Construction Engingeering Research Laboratory.

[104] Hill L. 2006. Georeferencing. The MIT Press. ISBN 0-262-08354-6|0-262-08354-6.

[105] Hillmann D, Phipps J. 2007. Application profiles: Exposing and enforcing metadata quality. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2007, (pp. 52-62).

[106] Hobona G, James P, Fairbairn D. 2006. Web-based visualization of 3D geospatial data using Java3D. IEEE Comput Graph Appl 26(4):28–33

[107] Holden C. 2003. From Local Challenges to a Global Community: Learning Repositories and the Global Learning Repositories Summit. The Academic ADL Co-Lab, November 2003.

[108] Howe D. 1993. Free on-line dictionary of computing. http://foldoc.org/index.cgi?Metadata.

[109] Hrebicek J, Pillmann W. 2009. Shared Environmental Information System and Single Information Space in Europe for the Environment: Antipodes or Associates? In Hrebicek (ed.) Proceedings of European conference of the Czech Presidency of the Council of the EU: Towards eEnvironment - Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe. Brno, Czech Republic: Masaryk University, 1-8, 2009

[110] Huang B, Jiang B, Li H. 2001. An integration of GIS, virtual reality and the internet for visualization, analysis and exploration of spatial data. Int J Geogr Inf Sci 15(5):439–456

[111] Hughes B. 2004. Metadata quality evaluation: Experience from the open language archives community. Digital Libraries: International Collaboration and Cross-Fertilization, (320–329).

[112] IEEE. 1990. Standard Computer Dictionary—A Compilation of IEEE Standard Computer Glossaries. New York, NY: 1990

[113] INSPIRE EU Directive. 2007. Directive 2007/2/EC of the European Parliament for establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union, L 108/1, Vol 50, 25 April 2007.

[114] International Organization for Standarization (ISO). 2007. "ISO 19139:2007 Geographic information - Metadata - XML schema implementation". International Organization for Standardization (ISO).

[115] International Organization for Standarization (ISO). 2003. ISO19115:2003. Geographic Information – Metadata

[116] International Organization for Standarization (ISO). 2003b. "ISO 15836:2003 Information and documentation - The Dublin Core metadata element set". International Organization for Standardization (ISO).

[117] International Standard Organization (ISO). 2002. ISO 19101:2002 Geographic Information – Reference Model. International Standard Organization

[118] JEITA. 2002. Exchangeable image file format for digital still cameras: Exif Version 2.2. Technical Standardization Committee on AV & IT Storage Systems and Equipment. Specification by JEITA, April 2002.

[119] Johnson P. 2005. Good Practice Guide for Developers of Cultural Heritage Web Services. Research Officer, UKOLN. Enlace: http://www.ukoln.ac.uk/interop-focus/gpg/Metadata

[120] Jokela S. 2001: "Metadata enhanced content management in media companies", Acta Polytechnica Scandinavica, Ma 114. Finnish Academies of Technology, Helsinki. Available at: http://lib.tkk.fi/Diss/2001/isbn9512256932/isbn9512256932.pdf

[121] Jones M, Taylor G. 2003. Metadata: Spatial Data Handling and Integration Issues. School of Computing Technical Report. Issued: Febrary 2003.

[122] Kalantari M, Olfat H, Rajabifard A. 2010. Automatic Spatial Metadata Enrichment: Reducing Metadata Creation Burden through Spatial Folksonomies. GSDI12: Realising Spatially Enabled Societies, Singapore 19-22 October 2010.

[123] Kawtrakul A, Yingsaeree C. 2005. Unified Framework for Automatic Metadata Extraction from Electronic Document. In Proceedings of IADLC2005 (The International Advanced Digital Library Conference), pp. 71-77, Nagoya, Japan.

[124] Kildow M. 1996. The value of Metadata (An NSDI report). US Fisheries and Wildlife Services. Web document. http://www.r1.fws.gov/metadata/meta.html

[125] Kobayashi M, Takeda K. 2000. Information retrieval on the web. ACM Computing Surveys (ACM Press) 32 (2): 144–173. doi:10.1145/358923.358934

[126] Kolodney U. 2004. Metadata requirements for a digital repository to accompany the American Anthropological Society's AnthroSource portal Project. University of Texas at Austin School of Information

[127] Kothuri R, Ravada S, Abugov D. 2002. Quadtree and R-tree indexes in oracle spatial: a comparison using GIS data. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data (pp. 546-557). ACM.

[128] Kralidis A. 2007. Geospatial Web Services: The Evolution of Geospatial Data Infrastructure. Scharl, K. Tochtermann (Eds.) The Geospatial Web, How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society. Springer London.

[129] Kresse W, Fadaie K. 2004. ISO Standards for Geographic Information. Heidelberg, pp. 322. Springer.

[130] Lagoze C, Van de Sompel H. 2003. The making of the Open Archives Initiative protocol for metadata harvesting. Library hi tech, 21(2), 118-128.

[131] Leiden K. 2001. A Review of Human Performance Models for the Prediction of Human Error. National Aeronautics and Space Administration, USA, 125pp

[132] Liddy E, Sutton S, Paik W, Allen E, Harwell S, Monsour M, Turner A, Liddy J. 2001. Breaking the metadata generation bottleneck: preliminary findings. In: Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia, pp. 464

[133] Lieberman M, Samet H, Sankaranarayanan J, Sperling J. 2007. STEWARD: Architecture of a Spatio-Textual Search Engine. ACMGIS'07: Proc. of the 15th ACM Int. Symp. on Advances in GIS, pp. 186 – 193, 2007.

[134] Löffler H, Baranger W, Steidl M. 2007. Photo Metadata White Paper 2007. IPTC, the International Press Telecommunications Council

[135] Luo J, Joshi D, Yu J, Gallangher A. 2011. Geotaging in multimedia and computer visión - a survey. Multimed Tool Appl 51(1):187–211

[136] MacEachren A, Kraak M. 2001. Research challenges in geovisualization. Cartography and Geographic Information Science, 28, 3–12. doi:10.1559/152304001782173970

[137] Maguire D. 2006. Geographic Earth Explorers: A New Software Paradigm for Visualizing And Analyzing Geography? 2006 Annual Meeting of the Association of American Geographers. Chicago, IL.

[138] Manber U, Myers G. 1990. Suffix arrays: a new method for on-line string searches. En SODA'90: Proc. of the Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 319–327, Philadelphia, USA, 1990.

[139] Manola F, Miller E. 2004. "RDF Primer: W3C Recommendation" 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210

[140] Manolopoulos Y, Nanopoulos A, Papadopoulos A, Theodoridis Y. 2005. R-Trees: Theory and Applications. Springer-Verlag New York, Inc.

[141] Manso M, Beltran A. 2012. Automatic Metadata Generation for Geospatial Resource Discovery. Book chapter in "Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications." IGI Global, 2012. Web. 7 Mar. 2012. doi:10.4018/978-1-4666-0945-7

[142] Manso M, Wachowicz M, Bernabe M. 2010. The design of an automated workflow for metadata generation. The 4th Metadata & Semantic research conference. S. Sánchez-Alonso and I.N. Athanasiadis (Eds.): MTSR 2010, CCIS 108, pp. 275-287. Springer, Heidelberg.

[143] Manso M. 2009. El uso de los metadatos para el desarrollo de un modelo de interoperabilidad para las Infraestructuras de Datos. Tesis doctoral dirigida por Mónica Wachowicz y Miguel Ángel Bernabé Poveda. Universidad Politécnica de Madrid, 2009.

[144] Manso M, Bernabé M. 2009. Geographic Information Implicit Metadata: Characterization of Temporal Cost and Error Types and Rates in Manual Compilation. GeoFocus, vol. 9, pp. 317-336.

[145] Manso M, Wachowicz M, Bernabé M. 2009. Automatic Metadata Creation for Supporting Interoperability Levels of Spatial Data Infrastructures. GSDI 11 World Conference: Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges. Rotterdan – Holanda, June 11-19.

[146] Manso M, Wachowicz M, Bernabé M, Sánchez A, Rodríguez A. 2008. Modelo de Interoperabilidad Basado en Metadatos (MIBM), Proceedings JIDEE 2008, Adeje (Tenerife), pp. 14-15, 2008

[147] Manso M, Nogueras J, Bernabé M, Zarazaga F. 2004. Automatic Metadata Extraction from Geographic Information, in: papers presented at the AGILE 2004 Conference, May 1st, 2004, Heraklion, Greece.

[148] Masser I. 2005. GIS Worlds: Creating Spatial Data Infrastructures, ESRI Press, ISBN: 1-58948-122-4, Redlands, California.

[149] Masser I, Rajabifard A, Williamson I. 2008. Spatially enabling governments through SDI implementation. International Journal of Geographical Information Science. Vol. 22, No. 1, (2008) 5–20

[150] Maué P, Roman D. 2011. The ENVISION Environmental Portal and Services Infrastructure. Proceedings of International Symposium on

Environmental Software Systems (ISESS), June 2011, Brno, Czech Republic.

[151] Michels H, Roth M, Beltran A. 2012. Semantic DESCaaS – Extending the Description as a Service Concept to Enable Semantic Annotations. Proceedings of the 15th AGILE International Conference on Geographic Information Science (AGILE 2012). Avignon, France. April 2012. ISBN: 978-90-816960-0-5.

[152] Middleton C, Baeza R. 2007. A comparison of open source search engines.

[153] Milstead J, Feldman S. 1999. Metadata: Cataloging by any other name. Online 25-31. Web document. http://www.onlineinc.com/onlinemag/OL1999/milstead1.html

[154] Mitchell T. 2005. Web Mapping Illustrated: Using Open Source GIS Toolkits, O'Reilly Media, ISBN 9780596008659

[155] Moellering H, Brodeur J. 2006. Towards a North American Profile of the ISO 19115 World Spatial Metadata Standard. GSDI-9 Conference Proceedings, 6-10 November 2006, Santiago, Chile. Web document. http://gsdidocs.org/gsdiconf/GSDI-9/papers/TS12.4paper.pdf

[156] Moen W, Stewart E, McClure C. 1997. Assessing metadata quality: Findings and methodological considerations from an evaluation of the US Government information locator service (GILS). Proceedings of the Advances in Digital Libraries Conference, 1998, (p. 246). IEEE Computer Society

[157] Morgenstern M. 1998. Integrating Web and Database Information for Collaboration through Explicit Metadata, Proceedings of the Seventh International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, 1998, IEEE

[158] Morris S, Nagy Z, Tuttle J. 2007. North Carolina Geospatial Data Archiving Project. NCSU Libraries and North Carolina Center for Geographic Information & Analysis. Web document.http://www.digitalpreservation.gov/partners/ncgdap/high/NCGDAP_InterimReport_June2008_final.pdf

[159] Moura E, Navarro G, Ziviani N, Baeza-Yates R. 1998. Fast searching on compressed text allowing errors. En SIGIR'98: Proc. of the 21th ACM SIGIR Conference, pp. 298–306, 1998.

[160] Naaman M. 2012. Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. Multimed Tool Appl 56(1):9–34

[161] Najjar J, Ternier S, Duval E. 2004. User behavior in learning object repositories: An empirical analysis. Proceedings of the ED-MEDIA 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications, AACE, 2004, (pp. 4373–4379).

[162] National Research Council (NRC). 2007. Successful response starts with a map, Improving geospatial support for disaster management. The National Academy Press, Washington DC, USA, 184p.

[163] Nativi S, Caron J, Davis E, Domenico B. 2005. Design and implementation of NetCDF markup language (NcML) and its GML-based extension (NcML-GML). Computers & Geosciences 31 (2005) 1104–1118.

[164] Nebert D, Whiteside A, Vretanos P. 2007. Open Geospatial Consortium Inc. OpenGIS®Catalogue Services Specification. 2007.

[165] Nebert D. 2004. Developing Spatial Data Infrastructures: The SDI Cookbook. Web document. http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf

[166] National Information Standards Organization (NISO). 2004. Understanding Metadata. Web document. http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

[167] Nievergelt J, Hinterberger H, Sevcik K. 1981. The Grid File: An Adaptable, Symmetric Multi-Key File Structure. En Proc. of the ECI Conference, pp. 236–251, 1981.

[168] Nogueras J, Zarazaga F, Muro P. 2005b. Geographic Information Metadata for Spatial Data Infrastructures: Resources, Interoperability and Information Retrieval. Springer 2005, XXII, 264 p. ISBN 978-3-540-24464-6

[169] Nogueras J, Zarazaga F, Béjar R, Álvarez P, Muro P. 2005. OGC Catalog services: a key element for the development of spatial data infrastructures. Comput Geosci 31 (2):199–209

[170] Nottingham M, Sayre R. 2008. Geonetwork opensource: The complete manual. http://www.fao.org/geonetwork/docs/Manual.pdf

[171] Nottingham M, Syare R. 2005. The atom syndication format. RFC 4287, http://tools.ietf.org/html/rfc4287

[172] Nuñez M, Díaz L, Granell C, Huerta J. 2011. Web 2.0 Broker: a tool for massive collection of user information. European Geosciences Union (EGU) General Assembly 2011 (EGU 2011), Vienna, Austria

[173] O'Reilly T. 2005. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. http://oreilly.com/web2/archive/what-is-web-20.html

[174] Ochoa X, Duval E. 2006. Quality Metrics for Learning Object Metadata. In Pearson E., Bohman P. (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2006, (pp. 1004-1011). Chesapeake, VA: AACE.

[175] Olfat H, Rajabifard A, Kalantari M. 2010. Automatic Spatial Metadata Update: a New Approach. FIG Congress 2010: Facing the Challenges – Building the Capacity. Sydney, Australia, 11-16 April 2010.

[176] Oosterom P. 2004. Geo-information Standards in Action. ISO TC211/Metadata (Danko, D.). Web document. http://www.ncg.knaw.nl/Publicaties/Groen/pdf/42Standards.pdf

[177] Open Geospatial Consortium (OGC). 2012. OGC® Geography Markup Language (GML) — Extended schemas and encoding rules v.3.3. Open Geospatial Consortium

[178] Open Geospatial Consortium (OGC). 2008. OpenGIS Keyhole Markup Language (KML) Implementation Specification, Version 2.2.0. Open Geospatial Consortium Inc (Open GIS Consortium Inc).

[179] Ortiz L, Zabala A, Casanovas P. 2008. Generación de metadatos según las Reglas de Implementación de Metadatos de la directiva INSPIRE en el marco del Departament de Medi Ambient i Habitage (DMAH) de la Generalitat de Catalunya. V Jornadas Técnicas de la IDE de España (JIDEE2008), Tenerife.

[180] Ostensen O, Danko D. 2005. Global Spatial Metadata Activities in the ISO/TC211 Geographic Information Domain. World Spatial Metadata Standards: Scientific and Technical Descriptions, and Full Descriptions with Crosstable. H. Moellering, H.J.G.L. Aalders & A. Crane (Editors). Elsevier Ltd.

[181] Parker D, Buchanan H, Hoult C, Taylor G, Coombes M. 1996. Guidelines for geographic information content and quality, Association for Geographic Information, London.

[182] Pinkerton B. 1994. Finding what people want: Experiences with the WebCrawler. In Proceedings of the Second International World Wide Web Conference (Vol. 94, pp. 17-20)

[183] Programmable Web. 2013. http://www.programmableweb.com

[184] Purves R, Jones C. 2011. Geographic Information Retrieval. SIGSPATIAL Special 3, 2 (July 2011), 2-4. DOI 10.1145/2047296.2047297

[185] Rajabifard A, Feeney M, Williamson I. 2002. Future directions for SDI development. International Journal of Applied Earth Observation and Geoinformation 4 (2002) 11–22

[186] Rew R, Davis G. 1990. NetCDF: an interface for scientific data Access. Computer Graphics and Applications, IEEE. On page(s): 76 – 82, Volume: 10, Issue: 4, DOI: 10.1109/38.56302. 1990.

[187] Roman D, Schade S, Berre A, Rune N, Langlois J. 2009. Model as a Service (MaaS), Proceedings of AGILE Workshop: Grid Technologies for Geospatial Applications. June 2009.

[188] Russell M. 2011. Mining the social web. O'Reilly Media, Sebastapol

[189] Samet H. 2006. Multidimensional and Metric Data Structures. M. Kaufmann (2006)

[190] Samet H, Webber R. 1985. Storing a Collection of Polygons Using Quadtrees. ACM Transactions on Graphics. July 1985: 182-222. InfoLAB.

[191] Santoro M, Mazzetti P, Nativi S, Fugazza C, Granell C, Díaz L. 2012. Methodologies for Augmented Discovery of Geospatial Resources. In Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications, 172-203. doi:10.4018/978-1-4666-0945-7.ch009

[192] Saquicela V, Vilches L, Corcho O. 2011. Lightweight semantic annotation of geospatial rest ful services. Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II, ESWC'11, pages 330–344, Berlin, Heidelberg, 2011. Springer-Verlag.

[193] Scharl A, Tochtermann K. 2007. The geospatial web: how geobrowsers, social software and the Web 2.0 are zapping the network society. Springer Verlag, ISBN: 978-1-84628-826-5

[194] Schelkens P, Skodras A, Ebrahimi T. 2009. The JPEG 2000 Suite. Wiley, Wiley-IS&T Series in Imaging Science and Technology, 2009.

[195] Schollmeier R. 2002. A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications. In Proceedings of the First International Conference on Peer-to-Peer Computing, IEEE.

[196] Scholten H, Fruijter S, Dilo A, VanBorkulo E. 2008. Spatial Data Infrastructure for Emergency Response in Netherlands. Remote sensing and GIS technologies for monitoring and prediction of disasters. Nayak and Zlatanova Eds, 179-197.

[197] Sheldon T. 2001. Linktionary. Entrada «Metadata». http://www.linktionary.com/m/metadata.html

[198] Sheppard S, Cizek P. 2009. The ethics of Google Earth: crossing thresholds from spatial data to landscape visualisation. J Environ Manag 90(6):2102–2117

[199] Smits P, Friiss-Christensen A. 2007. Resource Discovery in a European Spatial Data Infrastructure. IEEE Transactions on Knowledge and Data Engineering 19(1) pp. 85-95.

[200] Steinacker A, Ghavam A, Steinmetz R. 2001. Metadata standards for Web-based resources. MultiMedia, IEEE , vol.8, no.1, pp.70-76, Jan-Mar 2001 doi: 10.1109/93.923956

[201] Suh B, Bederson B. 2007. Semi-Automatic Photo Annotation Strategies Using Event Based Clustering and Clothing Based Person Recognition. In Interacting With Computers, v. 19, n. 4, pp. 524-544. Elsevier.

[202] Taubman D, Marcellin M. 2002. JPEG 2000: Image Compression Fundamentals, Standards and Practice. Kluwer International Series in Engineering and Computer Science, Secs 642.

[203] Taussi M. 2007. Automatic production of metadata out of geographic datasets (master's thesis). University of Technology, Department of Surveying. Helsinki, Espoo.

http://www.tkk.fi/Units/Cartography/theses/master/2007/Diplomityo_Taussi_M.pdf

[204] Taylor. 2004. The Organization of Information. 2nd ed. Westport, CN: Libraries Unlimited

[205] Tolk A. 2003. Beyond Technical Interoperability—Introducing a Reference Model for Measures of Merit for Coalition Interoperability, Proceedings of the 8th ICCRTS, Washington, D.C., June 17-19, 2003

[206] Tolosana R, Alvarez J, Lacasta J, Nogueras J, Muro P, Zarazaga F. 2006. On The Problem Of Identifying The Quality Of Geographic Metadata. Lecture Notes in Computer Science, Research and Advanced Technology for Digital Libraries, ECDL 2006, Volumen: 4172, pp 232 – 243.

[207] Toma I, Maué P. 2010. D5.1 Deployment of the open-source OGC Catalogue. ENVISION Deliverable D5.1, 2010

[208] Tomaszewski B. 2011. Situation awareness and virtual globes: applications for disaster management. Computers & Geoscience 37(1):86–92

[209] Tsai F. 2011. Web-based geographic search engine for location-aware search in Singapore. Expert Syst Appl 38:1011–1016

[210] Turner A. 2012. The OpenSearch Geo extension (Draft 2). http://www.opensearch.org/Specifications/OpenSearch/Extensions/Geo/1.0/Draft_2

[211] Turner A. 2006. Introduction to Neogeography, O'Reilly Media, ISBN: 9780596529956

[212] Turnitsa C, Tolk A. 2006 Battle Management Language: A Triangle with Five Sides Proceedings of the Simulation Interoperability Standards Organization (SISO) Spring Simulation Interoperability Workshop (SIW), Huntsville, AL, April 2-7, 2006

[213] Turpin A, Moffat A. 1997. Fast file search using text compression. En Proc. of the 20th Australasian Computer Science Conference, pp. 1–8, 1997.

[214] Turton I. 2008. GeoTools. En: G. B. Hall, M. G. Leahy (eds.), Open Source Approaches in Spatial Data Handling. Springer, pp. 153-169.

[215] Vaid S, Jones C, Joho H, Sanderson M. 2005. Spatio-Textual Indexing for Geographical Search on the Web. En SSTD'05: Proc. of the 9th Int. Symp. on Spatial and Temporal Databases, volume 3633 of LNCS, pp. 218 – 235, 2005.

[216] Vandenbroucke D, Crompvoets J, Vancauwenberghe G, Dessers E, Van Orshoven J. 2009. A Network Perspective on Spatial Data Infrastructures: Application to the Sub-national SDI of Flanders (Belgium). Transactions in GIS, 13(1) 105-122.

[217] Vossen G, Hagemann S. 2007. Unleashing Web 2.0: From Concepts to Creativity. Morgan Kaufmann, Burlington, MA.

[218] Walsh J. 2007. On Spatial Data Search. Terradue White Paper. http://www.terradue.com/images/T2-Research-07-003-OnSearch.pdf

[219] Warmerdam F. 2008. The Geospatial Data Abstraction Library. In: G. B. Hall, M. G. Leahy (eds.), Open Source Approaches in Spatial Data Handling. Springer, pp. 87-104.

[220] West J, Hess T. 2002. Metadata as a knowledge management tool: supporting intelligent agent and end user access to spatial datall. Decision Support Systems 32, pp. 247–264

[221] Williamson I. 2004. Building SDIs—the challenges ahead. In Proceedings of the 7th International Conference: Global Spatial Data Infrastructure, 2–6 February, Bangalore, India.

[222] Wilson E. 1998. Manuals Go Click. The Age. Melbourne, Australia

[223] Witten I, Moffat A, Bell T. 1999. Managing gigabytes: compressing and indexing documents and images. Morgan Kaufmann, 2nd edición, 1999.

[224] Woo M, Neider J, Davis T, Shreiner D. 1999. OpenGL programming guide: the official guide to learning OpenGL. Addison-Wesley Professional

[225] Wood J, Dykes J, Slingsby A, Clarke K. 2007. Interactive visual exploration of a large spatiotemporal dataset: reflections on a geovisualization mashup. IEEE Trans Vis Comput Graph 13 (6):1176–1183

[226] Woodley M, Clement G, Winn P. 2003. DCMI Glossary. Web document. http://dublincore.org/documents/2003/08/26/usageguide/glossary.shtml

[227] World Wide Web Consortium (W3C). 2004. The Semantic Web Activity. http://www.w3.org/2001/sw

[228] World Wide Web Consortium (W3C). 2002. Metadata Activity Statement. http://www.w3.org/Metadata/Activity.html

[229] Wright R, Haemel N, Sellers G, Lipchak B. 2004. OpenGL SuperBible. Addison-Wesley Professional

[230] Wu H, He Z, Gong J. 2010. A virtual globe-based 3D visualization and interactive framework for public participation in urban planning processes. Comput Environ Urban Syst 34(4):291–298

[231] Yamagishi Y, Yanaka H, Suzuki K, Tsuboi S, Isse T, Obayashi M, Tamura H, Nagao H. 2010. Visualization of geosciences data on Google Earth: development of a data converter system for seismic tomographic. Computers & Geosciences 36(3):373–382

[232] Zarazaga F, Lacasta J, Nogueras J, Torres M, Muro P. 2003. A Java Tool for Creating ISO/FGDC Geographic Metadata. In Geodaten- und Geodienste-Infrastrukturen - von der Forschung zur praktischen Anwendung. Beiträge zu den Münsteraner GI-Tagen. IfGI prints. 2003, vol. 18, pp. 17-30.

[233] Zeigler B, Murzy A, Yilmaz L. 2006. Artificial Intelligence in Modelling and Simulation. Encyclopedia of Complexity and System Science. Springer-Verlag, Germany (2006)

[234] Zhang J, Gong J, Lin H, Wang G, Huang J, Zhu J, Xu B, Teng J. 2007. Design and development of distributed virtual geographic environment system based on web services. Inf Sci 177:2968-3980

[235] Zhang B, Gonçalves M, Fox E. 2003. An OAI-Based Filtering Service for CITIDEL from NDLTD. Proceedings of the 6th International Conference on Asian Digital Libraries (IACDL 2003), Lecture Notes on Computer Science. Springer Verlag, Number 2911, ISBN 3-540-20608-6, pp 590-601, 2003.

[236] Zhou Y, Xie X, Wang C, Gong Y, Ma W. 2005. Hybrid index structures for location-based web search. En CIKM'05: Proc. of the 13th ACM CIKM Conference, pp. 155–162, New York, USA, 2005.

[237] Zhu Q, Li D, Zhang Y, Zhong Z, Huang D. 2002. CyberCity GIS (CCGIS): integration of DEMs, images, and 3D models. Photogramm Eng Remote Sens 68(4):361–367

[238] Zlatanova S, Fabbri A. 2009. Geo-ICT for Risk and Disaster Mangement. En Scholten, v/d Velde & van Manen (eds.): Geospatial Technology and the Role of Locations in science. Springer Dordrecht, 239-266.

[239] Zlatanova S, Dilo A. 2010. A Data Model for Operational and Situational Information in Emergency Response: The Dutch case. Proceedings of Gi4DM 2010 Conference on Geomatics for Crisis Management. Torino, Italy, Feb. 2010.

[240] Zobel J, Moffat A. 2006. Inverted files for text search engines. In: ACM Comput. Surv. , 38, 6, ACM, 2006.

Anexo A

Publicaciones

Los resultados de esta tesis han sido publicados (o se encuentran en proceso de revisión) en revistas, libros y conferencias nacionales e internacionales. A continuación se enumeran algunas de las publicaciones más relevantes.

A.1 Revistas

- Beltran A, Diaz L. 2013. GeoCrawler: An Integrated System for Publishing, Indexing and Searching Georeferenced Resources. Enviado a Computers & Geosciences (CAGEO).
- Beltran A, Michels H. 2013. Description as a Service: Improving data discovery. Enviado a Iternational Journal of Spatial Data Infrastructures Research (IJSDIR).
- Beltran A, Abargues C, Granell C, Núñez M, Díaz L, Huerta J. 2013. A virtual globe tool for searching and visualizing geo-referenced media resources in social networks. International Journal on Multimedia Tools and Applications. Volume 64, Issue 1, pp 171-195. Springer US. ISSN: 1380-7501. DOI: 10.1007/s11042-012-1025-0

A.2 Capítulos de Libro

- Manso M, Beltran A. 2013. Automatic Metadata Generation for Geospatial Resource Discovery. Geographic Information Systems: Concepts, Methodologies, Tools, and Applications. IGI Global, 2013. 2176-2207. Web. 28 Feb. 2013. ISBN: 978-146-662-038-4. DOI:10.4018/978-1-4666-2038-4.ch129
- Garcia I, Rodríguez L, Benedito M, Trilles S, Beltran A, Díaz L, Huerta J. 2012. Mobile Application for Noise Pollution Monitoring through Gamification Techniques. Lecture Notes in Computer Science, 2012, Volume 7522/2012, Entertainment Computing – ICEC 2012, pp. 562-571, DOI: 10.1007/978-3-642-33542-6_74.
- Manso M, Beltran A. 2012. Automatic Geospatial Metadata Generation for Geospatial Resource Discovery. Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications. IGI Global, 2012. pp. 78-110. Web. 30 Mar. 2012. ISBN: 9781466609457. DOI: 10.4018/978-1-4666-0945-7.ch005
- Abargues C, Beltran A, Granell C. 2010. MIMEXT: a KML extension for georeferencing and easy share MIME type resources. In M. Painho, M.Y. Santos, H. Pundt (Eds): Geospatial Thinking. Lecture Notes in geotecrmation and Cartography, Springer Verlag, Berlin, May 2010, pp. 315-334, ISBN 978-3-642-12325-2.
- Beltran A, Díaz L, Granell C, Huerta J, Abargues C. 2009. Description and publication of Geospatial Information. Book chapter in Pei-Gee Peter Ho (Ed): Geoscience and Remote Sensing, In-Tech (Vukovar, 2009), pp. 133-152, ISBN 978-953-307-003-2.
- Abargues C, Granell C, Díaz L, Huerta J, Beltran A. 2009. Discovery of user-generated geographic data using web search engines. In G. Jedlovec (Ed): Advances in Geoscience and Remote Sensing, In-Tech (Vukovar, 2009), pp. 207-228, ISBN 978-953-307-005-6.

A.3 Conferencias Internacionales

- Larizgoitia I, Beltran A, Llaves A, Toma I, Maué P. 2013. Environmental service discovery based on semantically annotated OGC service descriptions. Proceedings of the ACM Symposium on Appliced Computing (SAC) 2013, Semantic Web and Applications (SWA) Technical Track. Coimbra, Portugal, March 2013.
- Larizgoitia I, Beltran A, Toma I, Maué P. 2012. Publication and Discovery of Semantically Annotated Geospatial Web Services. Proceedings of the 26th International Conference on Informatics for Environmental Protection, Sustainable Development and Risk Management (EnviroInfo 2012), Dessau, Germany, August 2012, Shaker Verlag, Aachen 2012, ISBN: 978-3-8440-1248-4.
- Rodriguez L, Tamayo A, Beltran A, Huerta J. 2012. Visualization of Sensor Data in Virtual Globes. Proceedings of the 15th AGILE International Conference on Geographic Information Science (AGILE 2012). Avignon, France. April 2012. ISBN: 978-90-816960-0-5.
- Michels H, Roth M, Beltran A. 2012. Semantic DESCaaS – Extending the Description as a Service Concept to Enable Semantic Annotations. Proceedings of the 15th AGILE International Conference on Geographic Information Science (AGILE 2012). Avignon, France. April 2012. ISBN: 978-90-816960-0-5.
- Beltran A, Granell C, Huerta J. 2011. Describing heterogeneous resources through Apache Tika and OSGeo FDO. Proceedings of the 14th AGILE International Conference on Geographic Information Science – Advancing geotecrmation Science for a Changing World, Utrecht, 2011. AGILE. ISBN 978-90-816960-1-2. Utrecht, The Netherlands, April 2011.
- Beltran A, Granell C, Huerta J. 2010. OSGeo FDO y Apache Tika: construyendo una plataforma para la descripción de recursos multimedia. Actas de las I Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE 2010). Lisbon, Portugal, October 2010.
- Beltran A, Abargues C, Fonts O, Granell C. 2010. VisioMIMEXT: Incorporando contenidos multimedia en globos virtuales. Actas de las I Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE 2010). Lisbon, Portugal, October 2010.

- Tamayo A, Abargues C, Beltran A, Granell C, Huerta J. 2010. Gathering statistics of XML Schema Usage in OGC Web Services. Second Open Source GIS UK Conference (OSGIS 2010). Nottingham, UK, June 2010.
- Díaz L, Gould M, Beltran A, Llaves A, Granell C. 2008. Multipurpose metadata management in gvSIG. In Academic Proceedings of the 2008 Free and Open Source Software for Geospatial Conference (FOSS4G 2008). Cape Town, South Africa, Sep 2008, pp. 90-99, ISBN 978-0-620-42117-1

A.4 Conferencias Nacionales

- Beltran A, Díaz L, Huerta J. 2012. Construyendo un sistema de indexación y búsqueda de recursos georreferenciados. Actas de las VI Jornadas de SIG Libre (SIG Libre 2012). Girona, Spain, Mar 2012. ISBN: 978-84-694-9927-6
- Beltran A, Granell C, Huerta J. 2011. Descripción de recursos multimedia georreferenciados. Actas de las V Jornadas de SIG Libre (SIG Libre 2011). Girona, Spain, Mar 2011. ISBN: 978-84-694-1624-2
- Beltran A, Martín C. 2010. Generando descripciones de recursos para gvSIG desde GeoCrawler. 6th International gvSIG User Conference (gvSIG 2010). Valencia, Spain, November 2010.
- Abargues C, Beltran A, Granell C. 2010. Extensión y uso de KML para la anotación, georreferenciación y distribución de recursos de tipo MIME. Actas de las IV Jornadas de SIG Libre 2010 (SIG Libre 2010). Girona, Spain, Mar 2010.
- Beltran A, Huerta J, Díaz L, Granell C. 2009. GeoCrawler y gvSIG: un Tándem para la Generación Automática de Metadatos. 5th International gvSIG User Conference (gvSIG 2009).
- Díaz L, Granell C, Beltran A, Llaves A, Gould M. 2008. Extracción semiautomática de metadatos: hacia los metadatos implícitos. In Proceedings of the II Jornadas de SIG Libre (SIG Libre 2008). Girona, Spain, Mar 2008.
- Beltran A, Llaves A, Martín C, Díaz L, Gould M, Granell C. 2007. Extracción y gestión semiautomática de metadatos en gvSIG. 3rd gvSIG User Conference (gvSIG 2007). Valencia (Spain), November 2007.

A.5 Estancias de Investigación

- Semantic Annotation of Geospatial Web Services. Realizada en el Institute for Geoinformatics (ifgi) de la Westfälische Wilhelms - University of Münster (Alemania) bajo la supervisión de Prof. Dr. Werner Kuhn desde el 15 de Junio del 2011 al 15 de Diciembre del 2011.