Geolake Search (el futuro de las IDE está en mejorar su catálogo)

Geolake Search (the future of the SDIS is in improving its catalog)

Sergio Martín, Francisco J. López-Pellicer, Juan Valiño, F. Javier Zarazaga-Soria

REVISTA **MAPPING**
Vol. 28, 194, 24-30
marzo-abril 2019
ISSN: 1131-9100

Resumen

¿Por qué son los catálogos espaciales como son? ¿Por qué nos quejamos tanto de su comportamiento? ¿Realmente sirven para su propósito o ya deberíamos considerarlos una deuda técnica? Este artículo cuestiona los catálogos espaciales actuales proponiendo una aproximación diferente centrada en buscar metainformación construida a partir de los objetos espaciales contenidos en los conjuntos de datos. Para dar un soporte a esta idea este artículo explica cómo se puede implementar sobre uno de los motores de búsqueda más avanzados del momento, Elasticsearch, capaz de encontrar información relevante entre varios miles de millones de objetos espaciales.

Abstract

Why are spatial catalogs as they are? Why do we complain so much about their behavior? Do they really serve their purpose or should we already consider them technical debt? This article questions the current spatial catalogs proposing a different approach centered on searching metainformation constructed from the spatial objects contained in the data sets. To give support to this idea, this article explains how it can be implemented on one of the most advanced search engines of the moment, Elasticsearch, capable of finding relevant information among several billions of spatial objects.

Palabras clave: catálogo, IDE, Big Data, búsqueda, Elasticsearch.

Keywords: catalog, SDI, Big Data, search, Elasticsearch.

IAAA, Universidad de Zaragoza
segura@unizar.es
fjlopez@unizar.es
juanv@unizar.es
javy@unizar.es

Recepción 08/01/2019
Aprobación 24/01/2019

1. INTRODUCCIÓN

Cuando uno se acerca al mundo de las IDE sin un conocimiento preconcebido puede actuar sin verse influido por las convenciones de la comunidad. Por ello, el grupo IAAA de la Universidad de Zaragoza propuso a un recién graduado en Ingeniería Informática, Sergio Martín, diseñar e implementar, aplicando sólo sus conocimientos sobre el desarrollo de aplicaciones, uno de los elementos claves de una IDE: un buscador para un catálogo de datos espaciales ¿Cuál era su conocimiento del dominio? En sus propias palabras, «información espacial» le sonaba sólo a satélites y cohetes, y en cuanto a «SIG», asumía que era el nombre de un grupo de investigación de otra universidad.

Presentamos en este artículo el resultado de dicho trabajo: un prototipo de buscador de catálogo desarrollado en apenas un mes/persona utilizando aproximaciones funcionales y tecnológicas diferentes a las habituales. Pero afrontar el desarrollo de una solución desde una perspectiva diferente puede dar lugar a la sorpresa. Una de las premisas básicas del proyecto era desarrollar una solución dirigida por las necesidades del usuario. Sergio pronto detectó que uno de los requisitos fundamentales del buscador del catálogo era, precisamente, ayudar a encontrar nuevos conjuntos de datos. Parece un requisito obvio, pero tras analizar sistemas parecidos implantados en las IDE, Sergio llegó pronto a la conclusión que lo que se ofrecía a los usuarios eran aplicaciones de búsqueda con funcionalidad muy limitada, especialmente si se las analizaba a la luz de su propia experiencia como usuario de aplicaciones de uso común en la web o en el móvil. En la práctica, los buscadores de catálogo parecían diseñados para que sólo encontraran lo que el usuario ya sabía que estaba en el catálogo. ¿Quién es el responsable de este comportamiento? ¿Es resultado de un mal diseño? ¿O es debido a cómo son los catálogos que usamos en una IDE?

2. ¿QUÉ SUCEDE CON LOS CATÁLOGOS?

Es difícil construir una explicación de por qué los catálogos usados en una IDE tienen la forma que tienen actualmente. Por un lado, se puede argumentar que son el resultado de las capacidades tecnológicas de su momento de implementación y de su inspiración en otros sistemas análogos, como los catálogos de bibliotecas (Béjar, Nogueras-Iso, Latre, Muro-Medrano, & Zarazaga-Soria, 2009). Por otro lado, podríamos pensar que su naturaleza se debe a un pasado en el cual se han tomado una serie de decisiones que no han tenido en cuenta criterios de usabilidad (Larson, Olmos Siliceo, Pereira dos Santos, Klien, & Schade, 2006). ¿Son nuestros catálogos productos no acabados, productos con errores conocidos? Debemos comenzar a pensar que es así cuando vemos que es difícil o incluso imposible convencer a personas bien formadas en sistemas de información, como es el caso de Sergio, que los catálogos usados en una IDE sirven para su propósito:

Sergio: *«Si busco un conjunto de datos medioambientales sobre Zaragoza, ¿por qué no puedo preguntar al catálogo por los conjuntos de datos que efectivamente tengan datos sobre Zaragoza?»*

IAAA: *«El catálogo tiene una descripción de cada conjunto de datos que dice precisamente eso».*

Sergio: «No. Sólo si los metadatos mencionan explícitamente a Zaragoza. E incluso aunque el ámbito descrito en los metadatos cubra Zaragoza, puede que el conjunto de datos en sí no contenga ninguna información sobre Zaragoza».

A lo mejor no tenemos catálogos de datos geoespaciales. Sergio, como muchos usuarios, está pensando en el modelo de Google y otros buscadores masivos de información que lo que ofrecen es el acceso directo al recurso (web, fichero) el cual tienen previamente indexado y no solo a su descripción. Bajo esa perspectiva, nuestros catálogos no son catálogos de conjuntos de datos y servicios. Son catálogos de metadatos que describen conjuntos de datos y servicios. Unos sistemas que, si bien pudieran parecer análogos a un catálogo de datos espaciales, en la práctica, no llegan a serlo del todo cuando lo que queremos hacer es algo más que encontrar lo que ya sabemos que está en dicho catálogo.

3. BUSCADORES Y CATÁLOGOS

Cuando un usuario utiliza un buscador generalista, su búsqueda habitualmente cae en uno de estos dos tipos:

- Quiere encontrar algo que sabe que existe y conoce.
- Quiere encontrar algo que no sabe si existe.

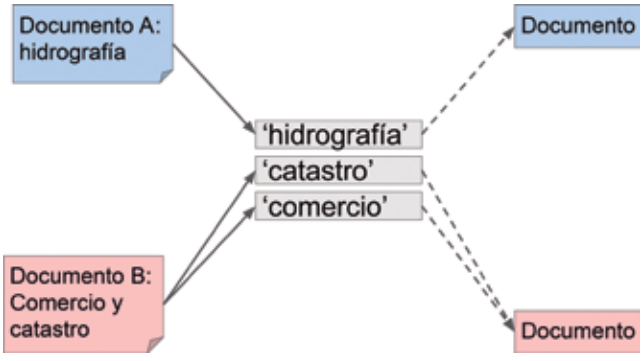

Figura 1. Un índice en un buscador

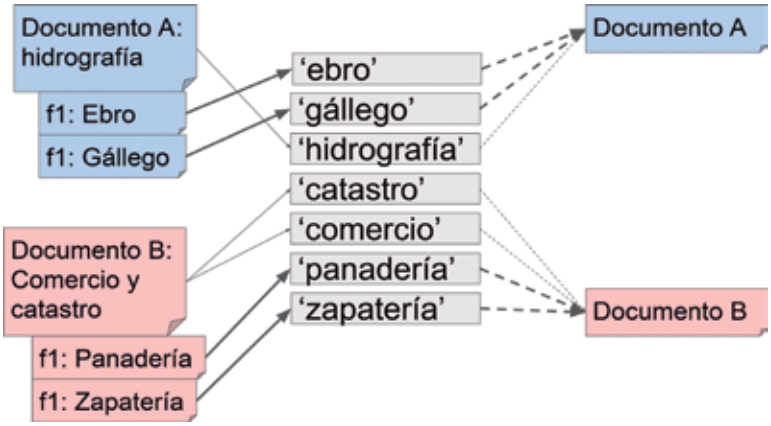

Figura 2. Índice enriquecido con los datos

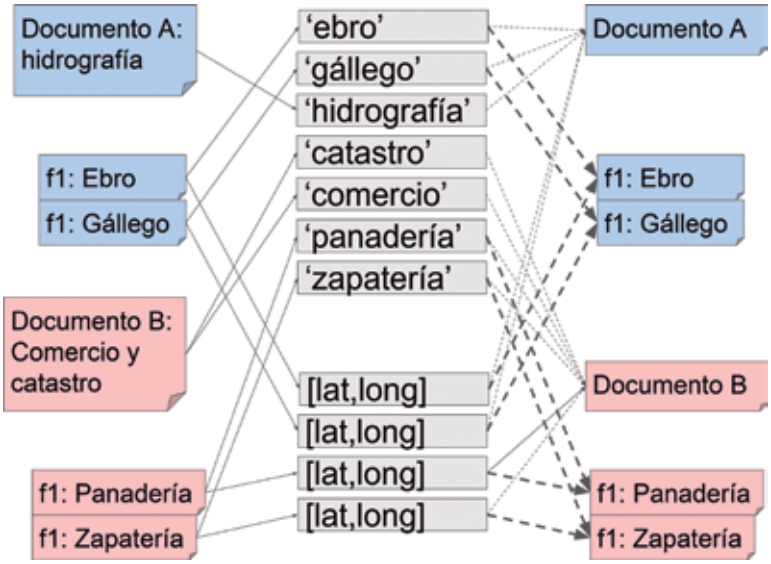

Figura 3. Índice enriquecido combinado con índice espacial

Este comportamiento se repite en un catálogo de datos geoespaciales. Si el usuario tiene claro el recurso que busca y lo conoce, no es difícil localizarlo introduciendo el nombre exacto o aproximado del recurso, ver que está en los primeros resultados de la búsqueda, y a continuación acceder a su descripción. Si quiere encontrar algo que no sabe si existe, el usuario hará una búsqueda más exploratoria. Los buscadores generalistas brillan en este tipo de búsquedas cuando, por ejemplo, un usuario busca «enciclopedias en línea» y el primer resultado es la Wikipedia. ¿Cómo lo logra el buscador generalista? Indexando, analizando y clasificando todos los contenidos de todas las páginas web para, quedémonos ahora con esta visión simplificada, crear índices invertidos de palabras que apuntan a documentos relevantes (figura 1).

¿Qué indexa actualmente un catálogo de datos espaciales? Metadatos y sólo metadatos. ¿Podría indexar algo más para dar mejores respuestas? Sí, por supuesto. No hay nada que lo impida ya que su diseño responde a la filosofía de los sistemas de recuperación de información.

4. ¿CUÁNDO PUEDE FALLAR UNA BÚSQUEDA SOBRE METADATOS?

Vamos a analizar dos casos de búsqueda en los que los metadatos no son suficientes para ofrecer un buen resultado. Estos dos casos son reales y fácilmente reproducibles por el lector.

Búsqueda por contenido. El primer caso es el de una búsqueda en la que el texto a encontrar no está presente en los metadatos, pero sí en los propios datos. Pudiera ser, por ejemplo, «busco conjunto de datos que contengan aguas superficiales del río Ebro». Pero si «Ebro» no aparece como palabra o concepto en los metadatos de un conjunto de datos relevante para esa búsqueda, dicho conjunto de datos no aparecerá en el resultado, aunque la palabra «Ebro» aparezca multitud de veces en los datos. Sin embargo, si «Ebro» aparece muchas veces en su metadato, probablemente será uno de los primeros resultados devueltos.

Búsquedas espaciales. El segundo caso es el verdaderamente interesante. Cuando buscamos datos espaciales en un catálogo estamos haciendo una búsqueda sobre rectángulos envolventes que delimitan los contenidos de las colecciones de datos. Aun suponiendo que dichos rectángulos sean correctos, ese método conlleva un inconveniente mayúsculo cuando en el catálogo hay conjuntos de datos autonómicos, nacionales, europeos y mundiales. Siguiendo el ejemplo anterior, «busco colecciones de datos que contengan aguas superficiales en el área definida por el rectángulo envolvente del río Ebro» genera muchos resultados no relevantes al incorporar respuestas de conjuntos de datos fuera de la cuenca del río Ebro. No solo eso, los resultados no estarán ordenados teniendo en cuenta el grado de relevancia espacial ya que todos los esfuerzos en mejorar la calidad de la recuperación de información en las aplicaciones de catálogo se han enfocado en analizar mejor el contenido textual de los metadatos. Esta situación lleva al caso extremo de irrelevancia en los resultados el cual se da cuando al examinar un conjunto de datos dado como respuesta se comprueba que en la práctica ni siquiera tiene datos en la zona buscada.

5. HERRAMIENTA DE EDICIÓN DE DATOS PARA LOS COLABORADORES DEL PROYECTO

Si queremos que los usuarios puedan aprovechar toda la potencia de un buscador, necesitamos analizar mejor los recursos que tenemos. Del mismo modo que podemos indexar un registro de metadatos codificado en formato XML, podemos analizar conjuntos de datos codificados en formatos como CSV, SHP, GML, WKT, etc. y extraer de ellas información valiosa para el índice (figura 2). Además, si se añade la localización de cada objeto espacial de un conjunto de datos a un índice espacial del buscador, es posible realizar preguntas cuya respuesta contenga un conjunto de datos como resultado si y sólo si tiene al menos un objeto espacial en el área que estamos buscando (figura 3). De ese modo se cubre el caso de búsqueda espacial antes mencionado.

Una vez que un catálogo se ha convertido en un buscador se puede de forma natural convertirlo en un servicio de descarga. Un catálogo donde no sólo se pudiera recuperar información sobre los conjuntos de datos sino también información sobre los propios datos permitiría hacer búsquedas de datos transversales, cruzando fronteras, colecciones de datos y temáticas, y permitiría crear colecciones de datos sintéticas construidas a partir de los resultados que al usuario le interesan.

6. DISEÑO E IMPLEMENTACIÓN

Lo primero que le puede a uno venir a la mente con esta propuesta es el impacto en tiempo de computación que supondría hacer una búsqueda sobre todos y cada uno de los datos indexados. Para atajar ese problema de manera integral, incluimos como principio fundamental la escalabilidad horizontal. En el mundo de los sistemas informáticos quiere decir que si el volumen de trabajo/datos crece se puede reaccionar añadiendo más computadoras en lugar de migrar a otras más potentes. Si tenemos, por un lado la necesidad de escalar horizontalmente y por otro la necesidad de ofrecer resultados con baja latencia, la respuesta casi directa es utilizar un sistema distribuido orientado a grandes volúmenes de datos, o como se le llama ahora, de Big Data.

6.1 Herramientas

Tras estudiar las posibles alternativas, la solución escogida para la prueba de concepto ha sido Elasticsearch. Una solución ampliamente utilizada en la industria, también en catálogos de metadatos, que además de ofrecer las típicas operaciones de búsqueda por texto permite realizar búsquedas espaciales (Corti, Lewis, Kralidis, & Mwenda, 2016). Para entender cómo funcionan esa herramienta, lo mejor es visualizar el índice mostrado en las figuras anteriores, solo que de un tamaño considerablemente mayor (varios terabytes) y después imaginar ese índice troceado y repartido en cada una de las máquinas que forman el clúster. Cuando se lanza una pregunta a cualquiera de esas máquinas, la máquina receptora la propaga al resto y cada una trata de ofrecer los mejores resultados basándose en «trozo de índice» que ella tiene. Cuando la máquina receptora recibe las respuestas de cada una de las máquinas, escoge los mejores resultados y se los envía al usuario que había lanzado la pregunta.

6.2 El flujo de datos

Para entender bien el funcionamiento del buscador, debemos primero conocer el resto de componentes que actúan detrás: almacenamiento, tratamiento de datos, etc. Sin entrar en demasiados detalles, el objetivo es diseñar un proceso de transformación que parta de un almacén de conjuntos de datos en diversos formatos y termine en el índice de Elasticsearch.

En su implementación, se trata de una cadena de procesos que transforman y redirigen los datos:

1. El primer proceso lee una pareja de ficheros <conjunto de datos, metadatos>.
2. Se extrae de los metadatos la información básica necesaria para formar un modelo común enfocado a la búsqueda: dirección del recurso, descripción, publicador, fecha, etc.
3. Se infiere de cada conjunto de datos su formato y se redirige a un proceso especializado: CSV, GML, Esri Shapefile, etc.
4. El proceso especializado:
    a. Lee el conjunto de datos
    b. Extrae los datos
    c. Los transforma a un modelo común
    d. Los envía individualmente al índice de Elasticsearch

De este modo, cualquier pareja de <conjunto de datos, metadatos> que esté bien formada y cumpla las especificaciones (en este caso ISO 19115) puede ser introducida en el sistema de búsqueda.

Para el prototipo se ha optado por utilizar la herramienta de integración de datos Talend, que ofrece herramientas para la creación de cadenas de procesos

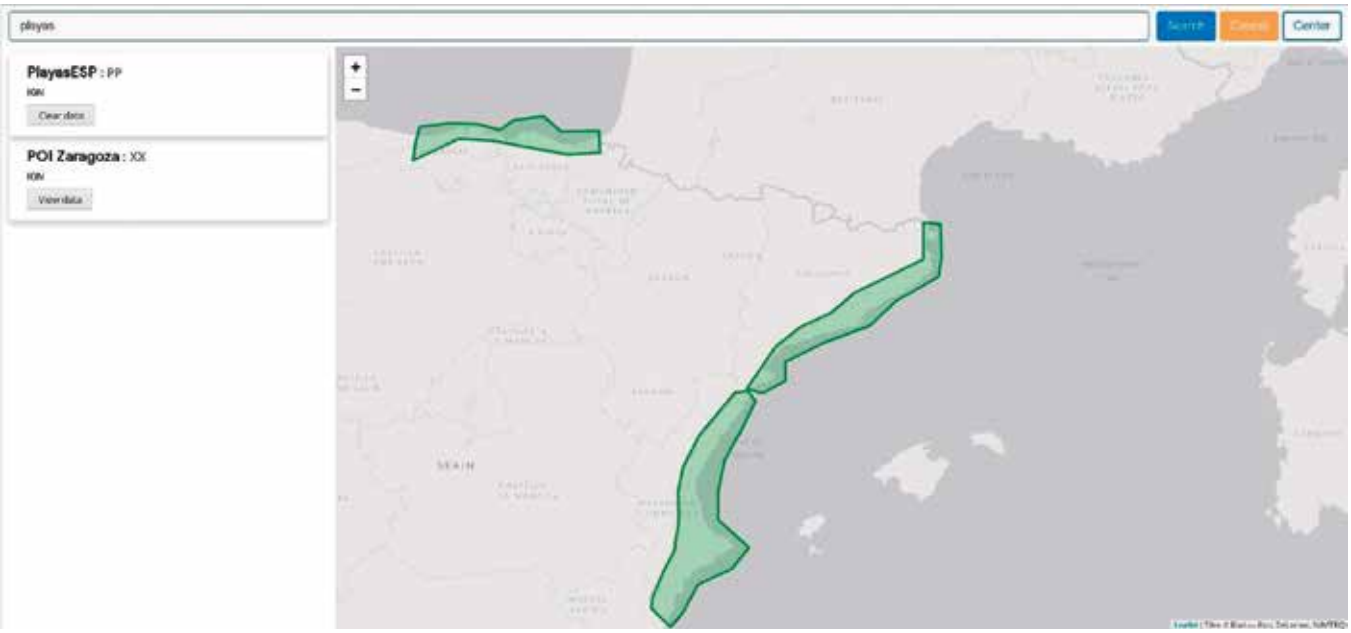

Figura 4. Prototipo buscando playas: el primer resultado es «Playas de España»

Figura 5. Prototipo buscando playas: el primer resultado es «POI de Zaragoza»

de integración. No obstante, se han tenido que adaptar manualmente esas cadenas para soportar las funcionalidades espaciales que se requerían en este proyecto.

6.3 La pregunta

Una vez está poblado el índice, es necesario conocer el lenguaje de consulta. Del mismo modo que una consulta en una base de datos relacional se hace con el lenguaje SQL, una búsqueda en Elasticsearch se hace en el lenguaje de búsqueda proporcionado por Elasticsearch. Llegado este momento es necesario recordar un concepto clave que subyace en este proyecto: la diferencia entre seleccionar datos y buscar información. Al preguntar a una base de datos SQL decimos informalmente que estamos «buscando» pero lo correcto sería decir que estamos «filtrando». Esto se debe a que en SQL le estamos diciendo al sistema que nos devuelva:

- Todos los resultados que encuentre que cumplan una serie de criterios.
- En orden aleatorio u ordenados por el valor de alguno de sus campos.

En un motor de búsqueda, lo que hacemos es proveerle de información (palabras clave, lenguaje natural) y el sistema nos devolverá:

- Los N primeros resultados que encuentre que más

se aproximen a nuestros criterios.
- Ordenados de mayor a menor relevancia.

Esta diferencia es la clave para entender el por qué de estos sistemas y por qué un gestor de base de datos no nos sirve como motor de búsqueda.

La pregunta entonces, en lenguaje natural, sería:
«Busco conjuntos de datos que:
- *bien sean relevantes y tengan datos en una zona*
- *bien tengan datos que sean relevantes y estén en una zona».*

Esta pregunta compleja se puede describir perfectamente con Elasticsearch y como resultado ofrece:
- Una lista de los conjuntos de datos que mejor se ciñen a la pregunta.
- Ordenadas de mayor a menor relevancia.
- Cada conjunto de datos ofrece la información básica recogida de sus metadatos: título, descripción, enlace al recurso original, etc.

Además, se puede enriquecer la visualización solicitando una agregación por cuadrículas espaciales, lo que permite dibujar un «mapa de calor» que muestre la cobertura real de cada conjunto de datos.

6.4 Prototipo

Con el flujo de datos implementado para procesar conjuntos de datos en formato CSV y Esri Shapefile y con metadatos en ISO 19115 y en Dublin Core, los datos cargados en Elasticsearch y la pregunta definida, sólo quedaba implementar el componente visual, el interfaz de búsqueda que incorpora un buscador facetado y un mapa. Para ello se ha utilizado el framework de desarrollo de aplicaciones web ReactJS por el interesante paradigma de diseño que ofrece, basado en componentes reutilizables. Esta elección ha permitido concluir la tarea de implementación del componente visual en poco más de dos semanas, lo que sumado al tiempo de desarrollo del proceso de carga, supone un total de un mes a cargo de un solo desarrollador. Algo realmente a tener en cuenta a la hora de valorar la facilidad de implantación de este modelo.

6.5 Resultados

Este prototipo fue presentado en las IX Jornadas Ibéricas de Infraestructuras de Datos Espaciales. En demostración se optó por utilizar datos sintéticos que reflejaran los casos de uso en lugar de buscar datos reales para evitar problemas relacionados con la carga de datos (objetos espaciales mal formados, metadatos incompletos, etc.). Podemos apreciar cómo al buscar «playas» sobre la península, aparece como primer resultado un conjunto de datos que contiene, efectivamente, playas de España (figura 4).

Sin embargo, al reducir la caja espacial de búsqueda a un área sobre Zaragoza, el conjunto de datos de las playas desaparece al no contener datos en esa área y en su lugar, aparece en primer lugar el conjunto de datos de «puntos de interés en Zaragoza», que contiene el objeto espacial llamado «Playas de Zaragoza» (figura 5). Eso demuestra el potencial de la búsqueda enriquecida en profundidad.

Los resultados del prototipo son prometedores. Por parte del interfaz de usuario, más que satisfactorios, ya que cubre las necesidades básicas de forma robusta y simple. En la parte del proceso, en cambio, es donde habrá que focalizar los futuros esfuerzos. Esto es debido a la gran heterogeneidad y complejidad de los metadatos, así como a los diversos formatos de datos espaciales.

7. CONCLUSIONES

Ahora conocemos la diferencia entre un catálogo de datos y un catálogo de metadatos. Ahora sabemos que con un catálogo de metadatos nunca podremos llegar a ofrecer los resultados que ofrecería un catálogo de datos. Ahora podemos imaginar una nueva IDE en la que los propios datos cobren la importancia que merecen y jueguen el papel que les corresponde. Ahora queda una pregunta: «¿Podemos ir más allá?»

Creemos que sí. El objetivo de este proyecto es trazar un posible rumbo para el futuro de las IDE y demostrar que es tecnológicamente viable ofrecer un mejor servicio. En este caso, nos hemos centrado en el primer elemento que un usuario se encuentra cuando trata de acceder a información espacial y hemos mostrado la viabilidad de una aproximación distinta, implementando con el esfuerzo de una persona/mes una prueba de concepto que cumple las necesidades básicas.

Con esta nueva aproximación la transformación puede ir mucho más allá: si deconstruimos las ideas hasta sus principios, aprovechamos herramientas modernas y tomamos como referentes a otros actores de la industria, podremos volver a ofrecer servicios valiosos tanto a nuestras propias administraciones como al resto de los usuarios.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Gobierno de Aragón (proyecto T59_17R) y el Gobierno de España (proyectos RTC-2016-4790-2 y TIN2017-88002-R).

REFERENCIAS

Béjar, R., Nogueras-Iso, J., Latre, M. Á., Muro-Medrano, P. R., & Zarazaga-Soria, F. J. (2009). Digital Libraries as a Foundation of Spatial Data Infrastructures. En Y.-L. Theng, S. Foo, D. Goh, & J.-C. Na (Eds.), Handbook of Research on Digital Libraries (pp. 382-389). Hershey, New York: IGI Global. https://doi.org/10.4018/978-1-59904-879-6.ch039

Corti, P., Lewis, B. G., Kralidis, A. T., & Mwenda, N. J. (2016). Implementing an open source spatio-temporal search platform for Spatial Data Infrastructures. En Proceedings of the 4th Open Source Geospatial Research and Education Symposium (OGRS2016). Perugia, Italia. https://doi.org/10.30437/ogrs2016_paper_37

Larson, J., Olmos Siliceo, M. A., Pereira dos Santos, M., Klien, E., & Schade, S. (2006). Are geospatial catalogues reaching their goals? En AGILE Conference on Geographic Information Science. Visegrád, Hungary.

Sobre los autores

Sergio Martín

Graduado en Ingeniería Informática y Estudiante del Máster homónimo. Desde que comenzó su labor como investigador para el grupo IAAA , se ha dedicado al análisis y desarrollo para la mejora de herramientas y procedimientos del ámbito de las TIC. Está especializado en el desarrollo ágil de pruebas de concepto innovadoras adoptando siempre los últimos avances de la industria para concebirlas. Su experiencia abarca desde el desarrollo software tradicional, hasta la creación y administración de infraestructuras cloud, pasando por el desarrollo de sistemas de información distribuidos y el análisis de datos. Actualmente enfocado en los sistemas de información espacial, busca los límites de las soluciones actuales y trata de esbozar el paisaje de lo que serán las herramientas del futuro.

Francisco J. López-Pellicer

Ingeniero Informático y Doctor Ingeniero en Informática por la Universidad de Zaragoza. Ha colaborado en diversos proyectos de investigación, desarrollo y transferencia centrando sus esfuerzos de investigación en el uso de la semántica geoespacial dentro del área multidisciplinaria de las Infraestructuras de Datos Espaciales. Sus intereses de investigación actuales son el desarrollo de ontologías geoespaciales, vocabularios y buscadores geográficos, el descubrimiento e indexación de recursos geo web y la publicación de información geo en la web de datos vinculados. Es autor y coautor de varios artículos publicados en revistas, libros y actas de congresos nacionales e internacionales. También ha contribuido en diversas convocatorias de licitaciones públicas en I+D+i, y en contratos de tecnología de transferencia de investigación, tanto nacionales como europeos.

Juan Valiño García

Licenciado en Matemática, Postgrado en Informática y Doctor Ingeniero en Informática por la Universidad de Zaragoza. También es Profesor Titular de la Universidad de Zaragoza estando adscrito al área de Lenguajes y Sistemas Informáticos del Departamento de Informática e Ingeniería de Sistemas. El campo de trabajo de este investigador son las tecnologías que permiten dar soporte al desarrollo de los sistemas de información geográfica en la nube y en sistemas distribuidos. En este marco ha participado en numerosas publicaciones y proyectos de investigación y transferencia en el ámbito de las Infraestructuras de Datos Espaciales a nivel local, autonómico y nacional.

F. Javier Zarazaga-Soria

Ingeniero Informático por la Universidad de Valencia y Doctor Ingeniero en Informática por la Universidad de Zaragoza. Director del grupo de investigación IAAA, ha colaborado en más de 40 proyectos de I+D+i, entre los que se destacan 4 proyectos financiados por la Comisión Europea, siendo coordinador en uno de ellos, y más de 20 proyectos financiados en convocatorias nacionales competitivas, siendo el investigador principal 7 de ellos. Además, ha colaborado en más de 20 contratos de investigación con organizaciones nacionales e internacionales y en varios proyectos de transferencia de tecnología. Además, como resultado de su actividad de investigación en el ámbito de los sistemas de información geográfica y de las Infraestructuras de Datos Espaciales es coautor de más de 50 artículos de investigación.