Error! No text of specified style in document. Open Data

LinkedData como modelo de datos abiertos

Índice

1 LinkedData y web semántica como modelo de datos abiertos

Linked Data y web semántica como modelo imprescindible en un proyecto de datos abiertos en un caso real para el portal Aragón Open Data

La mayoría de los usuarios de internet son conocedores de la gran cantidad de información que existe hoy en día en la red. Se puede encontrar cualquier tipo de conocimiento en tan solo un par de clics. Sin embargo, se tiene tan interiorizada la acción de buscar, ya sea a través de una caja de texto o por voz, que se obvia dónde se está buscando y cómo se realiza.

Parece que es algo de lo que no debemos preocuparnos, pero es importante entender que la web no distinguía hasta hace poco entre los distintos significados o contextos de la palabra que hemos introducido. Es verdad que esto a nivel web, buscadores, etc… ha evolucionado mucho, sobre todo con la llegada de las redes sociales. No están al mismo nivel de éxito los resultados de un usuario con tres redes sociales y un uso medio de cinco horas al día, que una persona que se conecta a su PC menos de una hora a la semana y desconoce la existencia de estas aplicaciones.

La primera persona cuando busque por un término, tendrá mayor probabilidad de que los resultados estén relacionados con sus gustos y actividades, es decir, estará en un contexto más acertado. El segundo usuario, por su parte, cuando realice la búsqueda, se podrá encontrar con resultados más ambiguos.

Esto se debe principalmente a la cantidad de información y el volumen de ficheros de texto, imágenes, videos y otros tipos de documentos que abundan en la red. Elementos que son subidos por distintos agentes, cada uno con su estructura, estilo, formato, codificación, etc…

¿Y qué tiene que ver todo esto con los datos abiertos?

La proliferación de las plataformas de datos abiertos, de la mano de políticas de transparencia en gran medida, provocaron que en un corto período de tiempo se subieran gran cantidad de ficheros, la mayor parte de carácter público, ya que la mayoría provenían de organizaciones e instituciones públicas.

Estos ficheros se almacenaron en distintos formatos (XML, JSON, CSV, TXT, PDF), que unos son más reutilizables que otros. Lo más seguro es pensar que un PDF es más legible para un humano y que un JSON será más fácil de interpretar por una máquina.

La web semántica nació para ofrecer más valor a la web tradicional, que se construía en lenguaje HTML. Esta nueva web quería otorgar a cada archivo, recurso o contenido de la red, un contexto o definición que permitieran a las máquinas comprender el significado de estos. El objetivo final no era solo el de poder mostrarlo a través de una pantalla, sino que pudiera ser interpretado o reutilizado. Para ello fue imprescindible complementar la tecnología actual con el lenguaje estructurado XML (Extensible Markup Language) y el RDF (Resource Description Framework).

Los metadatos, ontologías o vocabularios ganaron importancia al ser pieza clave para representar a nivel de maquina el conocimiento de la web. Antes de avanzar en estos conceptos semánticos, es necesario aterrizar el término de Linked Data, un término informático que ayuda a materializar la teoría.

Poco a poco, el concepto Linked Data es menos desconocido en la web. Y eso es porque cada vez se utiliza más, sobre todo los gigantes de la tecnología y las redes sociales. El Linked Data o datos enlazados, describe un método de publicación de datos estructurados para que puedan ser interconectados.

Linked Data surgió a partir del proyecto de la Web Semántica, de la necesidad de introducir información en forma de metadatos en la Web.

Al principio se creía que el poder residía en la cantidad de datos que componían la red, pero con el tiempo se vio que la fuerza estaba en la calidad de estos datos, y la calidad se conseguiría siguiendo unas reglas estrictas en el momento de inserción de los datos.

La web semántica demandaba utilizar un identificador único para cada dato, para ello se necesitaba utilizar una forma estandarizada de distinguir los datos unos de otros y se escogió el URI (Uniform Resource Identifier).

Por ejemplo, haciendo referencia a la especificación Dublin Core, en el momento que queramos especificar el publicador de un recurso, en vez de poner el término Publicador y seguido de la persona, se debería utilizar el URI **http://purl.org/dc/elements/1.1/publisher** (http://purl.org/dc/elements/1.1/publisher). De esta manera, cualquier persona o máquina que lea este metadato sabrá que el nombre de la persona, organización o servicio que se indica es el publicador de dicho recurso. Al utilizar estándares se garantiza que los datos estén interrelacionados.

Como se mencionaba antes, las ontologías o vocabularios, como también se les denominan, son una de las herramientas más importantes de la web semántica. Las ontologías públicas son el estándar para evitar definir entidades ya creadas y no replicarlas de nuevo, facilitando la lectura y la legibilidad a las máquinas.

Con el objetivo de hacer más entendible la teoría de las ontologías, se va a intentar dar forma con dos casos concretos que se utilizan en Aragón Open Data. Por un lado, se hablará de la Estructura de Información Interoperable de Aragón (Ei2A) y, por otro, del Identificador Europeo de Legislación.

Estructura de Información Interoperable de Aragón en Aragón Open Data

Dentro del proyecto de Aragón Open Data nació una iniciativa por Acuerdo de 17 de julio de 2012 del Gobierno de Aragón. En virtud del mismo se ordenó el inicio del proyecto de apertura de datos públicos y el 6 de febrero de 2013 se implantó a través del Portal opendata.aragon.es. A lo largo de este tiempo se han realizado numerosos trabajos para conseguir la automatización en la publicación de la información para asegurar que terceros puedan reutilizarla de la mejor manera.Y es que debido al volumen de datos que se comenzó a almacenar y a ofrecer en el portal, se valoró la necesidad de automatizar la gestión de la información. Esto empezó a tener una especial relevancia de todos aquellos elementos que ayuden en la mejoría de la estructuración de la información y la estandarización de los datos que contienen las bases de datos.

En base a esto, dentro de la Dirección General de Administración Electrónica y Sociedad de la Información, surgió la idea de generar un conjunto de reglas técnicas y legales que permitieran profundizar en esa estandarización y que llevan a pensar en la creación de la Estructura de Información Interoperable de Aragón (EI2A).

El objetivo principal es que el EI2A sea el marco en el que los datos abiertos y en general la información del Gobierno de Aragón pueda comenzar a automatizarse en base a una estructura común. Para ello es necesario de seguir una serie de elementos técnicos, organizativos y legales.

Aunque se explica más adelante en este mismo artículo, una de las herramientas más utilizadas y que se ejecuta a diario, llamada Datacube, se basa en la **Estructura de Información Interoperable de Aragón** (https://opendata.aragon.es/def/ei2a/index.htm). Esta ontología posee la capacidad de apoyar la interoperabilidad de datos del dominio del Gobierno de Aragón, con el objetivo de estandarizar información y explotarla en un presente y en un futuro.

De acuerdo a las recomendaciones de la Web Semántica, la ontología EI2A reutiliza diversas ontologías, esquemas y vocabularios para describir entidades que se adecuen al dominio del Gobierno de Aragón. Reutilizar ontologías facilia el intercambio de conocimiento y la comunicación entre personas, agentes inteligentes y sistemas.

De esta forma, se consigue describir los conceptos deseados del dominio del Gobierno de Aragón introduciendo un número mínimo de nuevos elementos.

Algunas de las ontologías y vocabularios usados para el desarrollo de la propuesta ontológica de Informacion Interoperable de Aragón son las siguientes:

- **Simple Knowledge Organization System (SKOS)** (https://www.w3.org/TR/vocab-org/)**:** Ontología desarrollada por W3C para describir organizaciones.
- **Owl** (http://www.w3.org/2002/07/owl): un lenguaje de marcado para publicar y compartir datos usando ontologías en la WWW. OWL tiene como objetivo facilitar un modelo de marcado construido sobre RDF y codificado en XML.
- **RDF Schema** (http://www.w3.org/2000/01/rdf-schema): Vocabulario de uso general que se utiliza en el modelado de esquemas en RDF para la creación de otros Vocabularios.
- **XML Schema** (http://www.w3.org/2001/XMLSchema): Lenguaje de esquema utilizado para describir la estructura y las restricciones de los contenidos de los documentos XML de una forma muy precisa.
- **Dublin Core Metadata Terms** (http://dublincore.org/): Conjunto complete de términos elaborado por la iniciativa de metadatos de Dublin Core, entidad de referencia en el desarrollo de metadatos de amplio ámbito de actuación, así como en las buenas prácticas para su gestión.
- **ELI**: legislación en un formato normalizado, de manera que puede localizarse, intercambiarse y reutilizarse por encima de las fronteras. Es la siguiente ontología que se explica en el artículo.

Identificador Europeo de Legislación como formato del Boletín Oficial de Aragón

El Identificador Europeo de Legislación (ELI) es una iniciativa, adoptada en el año 2012 conjuntamente por los países y las instituciones de la Unión Europea, que permite acceder online a la legislación en un formato formalizado, de manera que pueda localizarse, intercambiarse y reutilizarse por encima de las fronteras.

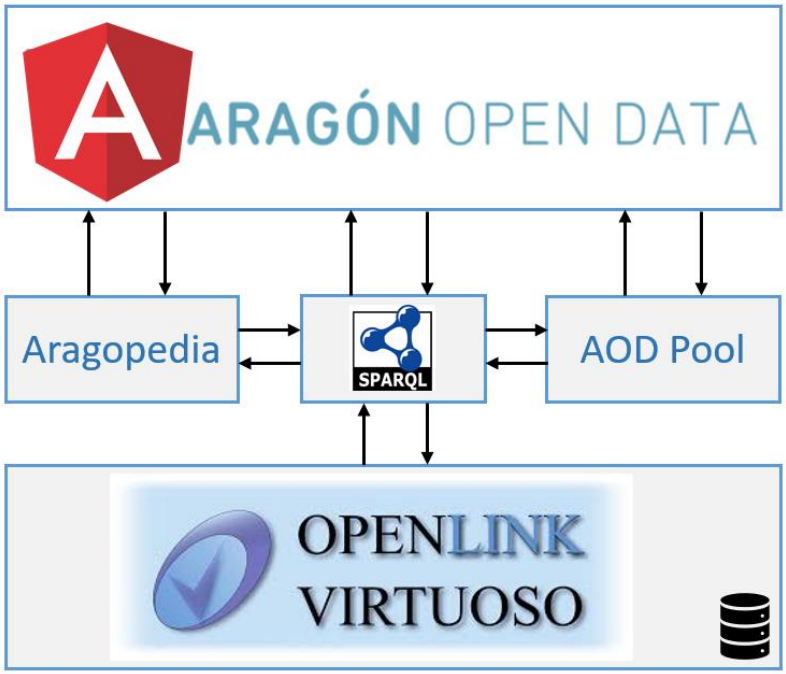
Al igual que se explicaba la importancia de una estructura común para los datos del Gobierno de Aragón, es sumamente importante utilizar identificadores permanentes y metadados estructurados para mejorar la calidad y la fiabilidad de la información jurídica online.

Esto favorece la interoperabilidad entre los sistemas de información estructurando la legislación de manera normalizada, pero teniendo en cuenta al mismo tiempo las características específicas de los distintos ordenamientos jurídicos.

El Boletín Oficial de Aragón está colaborando con la iniciativa europea ELI, que permite acceder online a la legislación en un formato normalizado. Desde el proyecto de Aragón Open Data se desarrollaron unos esquemas para leyes, decretos y órdenes que relacionaba los datos del BOA con el formato del ELI.

Una vez se desarrollaron dichos esquemas, se introdujeron en una herramienta llamada AOD Pool para insertar en la base de datos semántica de la plataforma y pudiera ser explotada desde fuera.

La web semántica en Aragón Open Data



Todos los datos enlazados que se exponen a través del portal de datos abiertos del Gobierno de Aragón son almacenados en Virtuoso. Es un servidor universal, un híbrido de Servidor de Aplicaciones Web y Sistema de Gerenciamiento de Banco de Datos Objeto-Relacional (ORDBMS).

Su arquitectura permite la persistencia de datos en los formatos relacional, RDF, XML, texto, documentos, Datos Conectados, etc.

Virtuoso es una de las plataformas de Linked Data más utilizadas en la actualidad. Además posee un banco de tripletas nativo, como la DBpedia, un repositorio de conjuntos de datos, abierto y gratuito, con información estructurada proveniente de Wikipedia.

Una tripleta semántica es la entidad atómica de datos en el modelo de datos Resource Description Framework (RDF). Como su nombre indica, una tripleta o una terna es un conjunto de tres entidades que codifica una declaración sobre datos semánticos en forma de expresiones sujeto-predicado-objeto.

Por ejemplo, una de las tripletas que representa el tipo de un documento del BOA ELI sería:

Sujeto:

<http://www.boa.aragon.es/eli/es-ar/d/2020/10/14/89/dof>

Predicado:

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

Objeto:

<http://data.europa.eu/eli/ontology#LegalResource>

Esto indica, utilizando URIs como se puede observar, que el documento indicado es de tipo Recurso Legal (LegalResource). En este caso, para el objeto, se está utilizando una de las numerosas definiciones que contribuye la ontología pública del ELI.

En la actualidad, en Aragón Open Data, hay dos servicios que insertan tripletas semánticas en Virtuoso. Una gran parte de los datos disponibles para cada municipio, provincia, comarca y para la Comunidad

Autónoma de Aragón, están disponibles utilizando un vocabulario estándar del consorcio W3C, como es DataCube.

Este vocabulario se utiliza en Aragopedia para ofrecer distintos bloques de información sobre cada una de estas divisiones administrativas. Lo más destacable es que cada dataset se expresa como un cubo de datos (bidimensional, tridimensional, o en general de cualquier número de dimensiones). Cada una de las celdas de este cubo de datos se denomina qb:Observation, donde el prefijo qb se refiere a http://purl.org/linked-data/cube# . Y cada una de estas observaciones tiene asociadas una serie de dimensiones o atributos, como pueden ser el año al que se refieren las estadísticas, el municipio, comarca o provincia al que se refieren, la propiedad que se está midiendo, etc.

Hasta el momento, Aragopedia contiene los siguientes cubos de datos, obtenidos con la siguiente consulta SPARQL:

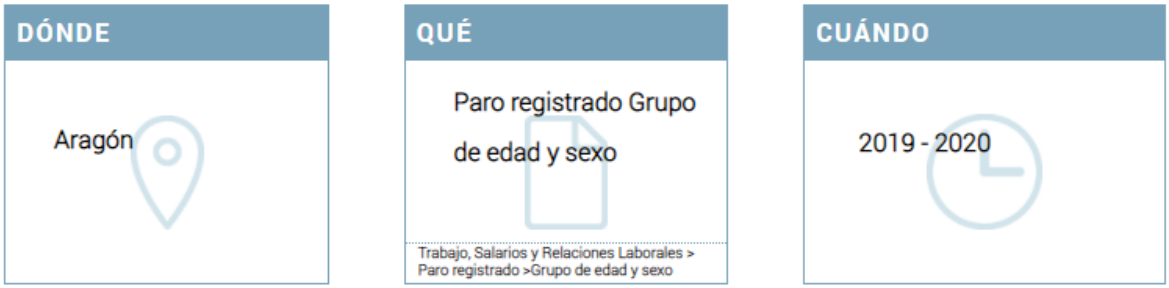
PREFIX qb: <http://purl.org/linked-data/cube#> SELECT DISTINCT ?y WHERE {?x qb:dataSet ?y }

La herramienta dentro de Aragón Open Data se encarga de descargar estos cubos de datos del BI del Instituto Aragonés de Estadística, trasnformarlos a formato de terna e insertarlo en la base de datos semántica.

Desde el servicio de Aragopedia, que es un frontal estático, se hace consultas SPARQL a los distintos cubos de datos insertados para ofrecer una herramienta de consulta para cualquier usuario sin conocimientos previos de web semántica o consultas SPARQL.

ARAGOPEDIA: LOS DATOS

Busca toda la estadística local en tres pasos



PARO REGISTRADO GRUPO DE EDAD Y SEXO

CSV JSON

Trabajo, Salarios y Relaciones Laborales > Paro registrado >Grupo de edad y sexo

Informe metodolológico

Fuente: Instituto Aragonés de Estadística (IAEST)

☑ Sexo ☑ Edad (grupos quinquenales) ☑ Nº parados

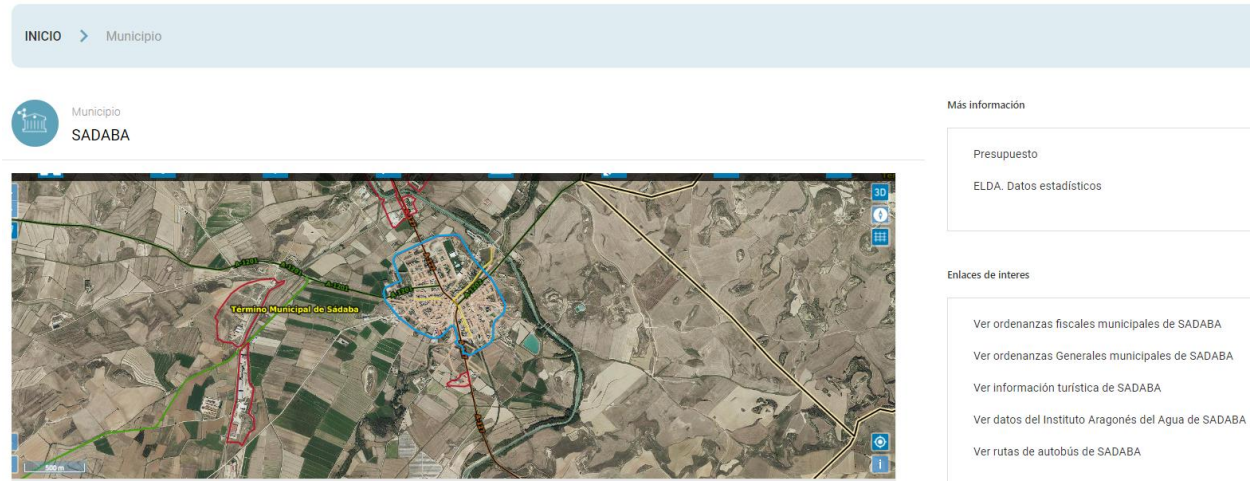
Solo se muestran los primeros 300 resultados. Para disponer de la totalidad de los datos descargue CSV / JSON

Área	Mes	Sexo	Edad (grupos quinquenales)	Nº parados
Aragón	2020-09	Hombres	30-34	2833
Aragón	2020-09	Hombres	16-19	1449
Aragón	2020-09	Hombres	35-39	2813
Aragón	2020-09	Mujeres	16-19	1072

El otro servicio que nutre Aragón Open Data de datos enlazados es el proyecto Aragón Open Data Pool, una herramienta innovadora dentro del panorama de los datos abiertos nacionales, que demuestra la importancia de centralizar datos y servirlos para favorecer su uso y explotación.

El AOD Pool, con ayuda de esquemas XML previamente construidos, relaciona con la ontología EI2A los datos existentes en Aragón Open Data bajo la temática de transporte, gestión del agua, turismo y agricultura-PAC (Política Agrícola Común de la Unión Europea), genera tripletas y las almacena en Virtuoso.

Una vez que las ternas están cargadas en Virtuoso, desde un frontal desarrollado en PHP con el framework Symfony, se puede visualizar, por ejemplo, todos los datos relacionados de un municipio de Aragón, como se aprecia en la siguiente imagen.



Gracias a las URIs utilizadas de lo municipios, se puede relacionar toda la información de un mismo municipio y mostrarla de manera conjunta. Así, si un usuario quiere informarse sobre los presupuestos, datos estadísticos, ordenanzas fiscales, información turística o las rutas de autobús de un municipio de Aragón, no tiene que navegar entre distintas pantallas ni aprender a programar complejas consultas en SPARQL para relacionar sus búsquedas.

Además, el AOD Pool también se encarga de insertar las Leyes del Boletín Oficial de Aragón como datos abiertos enlazados y así ofrecer la posibilidad de realizar consultas cruzadas con otros organismos bajo la misma estructura.

Desde un panel de control interno para la administración del equipo de Aragón Open Data se gestiona la actualización de los datos enlazados. En la siguiente imagen se aprecian las vistas que existen en la actualidad para cargar y actualizar, de forma semantizada, los distintos datos del BOA.

Nombre Vista	Periodicidad	Criterio	Fecha ↓	Hora	Estado	Logs Carga	Archivos Carga	Acción	Valido
boa_eli	Diaria		25 Oct 2020	16:00	Tipo (Actualización): Formulario Web Clase: dc_type (http://opendata.aragon.es/def/ei2a#recurso_legal)				
boa_eli_correcciones	Diaria		25 Oct 2020	16:05	Tipo (Actualización): Formulario Web Clase: dc_type (http://opendata.aragon.es/def/ei2a#recurso_legal)				
boa_eli_ordenes	Diaria		25 Oct 2020	16:10	Tipo (Actualización): Formulario Web Clase: dc_type (http://opendata.aragon.es/def/ei2a#recurso_legal)				
boa_eli_ordenes_correcciones	Diaria		25 Oct 2020	16:15	Tipo (Actualización): Formulario Web Clase: dc_type (http://opendata.aragon.es/def/ei2a#recurso_legal)				

Para los usuarios más avanzados o para los desarrolladores que buscan reutilizar y ofrecer nuevos servicios con los datos de Aragón Open Data, es necesario que sepan que todas estas ternas semánticas pueden ser explotadas a través de la interfaz web: **SPARQL Endpoint** (https://opendata.aragon.es/sparql), desde la que se pueden ejecutar consultas. Estas consultas tienen que ser escritas bajo el lenguaje SPARQL, que es un acrónimo recursivo del inglés SPARQL Protocol and RDF Query Language. Es un lenguaje estandarizado para la consulta de grafos RDF.

La potencia de una herramienta SPARQL es su habilidad para crear consultas completas que referencian múltiples variables al mismo tiempo. Y no solo datos bajo un mismo portal, sino datos de distintos portales que comparten estructura semántica.

En resumen, para la creación de una web semántica, lo primero es tener acceso a los datos abiertos que se quieren conectar. Es imprescindible conocer la estructura de los datos y tratar de agruparlos bajo esquemas comunes de información.

La implantación de una base de datos RDF como Virtuoso y un editor de consultas complejas SPARQL es sencilla, si bien, es importante valorar y analizar el volumen de datos que se van a almacenar y la potencia de la máquina que va a soportar la arquitectura semántica.

La complejidad reside en la creación de nuevas herramientas que sean capaces de leer un gran volumen de datos con estructuras diferentes, homogeneizarlas y crear ternas para almacenar en bases de datos semánticas. Es un trabajo de campo tedioso y que requiere de un amplio esfuerzo.

Si el proyecto es de datos abiertos y se requiere desarrollar una estructura de web semántica es una buena decisión trabajar con el stack de CKAN, que ya se explicó en un artículo anterior, Virtuoso y consultas SPARQL. Ya que te permiten desarrollar, evolucionar tu plataforma y ofrecer esta posibilidad a otros desarrolladores que busquen explotar estos datos.

1.1 Autor

Gabriel Alcober Fuertes

Equipo de Aragón Open Data | Delivery Lead en DXD

Gabriel inició su carrera profesional en 2011, y durante este tiempo ha participado en proyectos con gran diversidad de tecnologías. Ha trabajado en proyectos Salesforce y Java, y en los últimos tres años ha trabajado en desarrollos enfocados a datos abiertos y web semántica.