

OPEN LINKED DATA: LA NUEVA FRONTERA

CARLOS DE LA FUENTE - MARTÍN ÁLVAREZ ESPINAR

Introducción

En este capítulo profundizamos en el fundamento técnico del buen uso de los estándares de la web para lograr la máxima eficiencia en los procesos de publicación y reutilización de la información del sector público. Para dar una visión lo más precisa de los últimos avances en representación y combinación de información basados en la estructura de la web, es necesario explicar los entresijos técnicos en los que se fundamenta. Por este motivo consideramos importante la introducción en este libro de divulgación sobre *open government*, de este capítulo que, aunque de contenido y expresión técnica, aborda un recorrido por las tecnologías actuales más avanzadas en *open data* y, por ello, aunque resulte de lectura árida, pensamos que es fundamental a la hora de tomar decisiones --sobre todo para los responsables técnicos *open data*-- en cuanto al enfoque metodológico de representación y publicación de datos. Al final del capítulo se contará con una visión amplia sobre la oportunidad que supone publicar datos contextualizados semánticamente y que facilitan la reutilización y la creación de servicios avanzados, frente a la mera publicación de datos, tal cual existan, que añaden complejidad al proceso de reutilización frente al de publicación.

Los responsables *open data* normalmente deben decidir cómo publicar, y cómo se balancea en términos de complejidad técnica la carga de trabajo, para lograr el máximo beneficio en los procesos de reutilización de la información. La complejidad técnica puede caer del lado del reutilizador, al ser necesario un procesamiento exhaustivo de los datos publicados, o caer del lado del publicador si la exposición de los datos se realiza con la mayor calidad técnica posible.



CARLOS DE LA FUENTE

Es ingeniero técnico en informática y representante de CTIC Centro Tecnológico en el Comité Asesor de W3C. Desde hace 20 años ha estado vinculado a proyectos tecnológicos y en la actualidad funge como director del Área de Servicios Tecnológicos de CTIC, que centra su actividad en tres ámbitos fundamentales: *open data*, en virtud del liderazgo ejercido por CTIC dentro de W3C al lanzar el Grupo de Trabajo Internacional sobre eGovernment, donde uno de los focos principales es *open government data*; el segundo ámbito es la Calidad Web, que tiene por objeto facilitar a la sociedad herramientas y servicios tecnológicos relacionados con los estándares W3C, especialmente accesibilidad web e independencia de dispositivo; por último, se avoca al desarrollo de procesos de madurez digital en entornos empresariales. Carlos de la Fuente es coautor del libro *El desafío hacia el gobierno abierto en la hora de la igualdad*, publicado por CEPAL.



MARTÍN ÁLVAREZ ESPINAR

Es ingeniero en informática por la Universidad de Oviedo, con amplia experiencia en el desarrollo de aplicaciones web basadas en estándares. Funge como responsable de la Oficina del W3C en España y es consultor en CTIC. Compagina su trabajo en la Oficina del W3C con labores de consultoría sobre administración electrónica, concretamente ayudando a las administraciones públicas españolas en el desarrollo e impulso de iniciativas open data y reutilización de información del sector público, a través de los mecanismos de la Web Semántica. Participa activamente en grupos de trabajo de estandarización del W3C y del programa ISA (Interoperability Solutions for the European Public Administrations) de la Comisión Europea. En 2010 fue nombrado por el diario El Mundo como una de las 25 personas más influyentes en internet en España.

Esto implica que los costes del proceso de reutilización los soporta una u otra parte, en función del mecanismo de publicación elegido.

Publicar y reutilizar información se puede hacer de la forma más básica, pero a la vez menos eficiente. Lo que puede resultar inmediato, minimizando el esfuerzo para el organismo público, se torna ineficiente y costoso para el reutilizador, sino se utilizan los estándares mas avanzados para la publicación de datos (Web semántica). Por ejemplo, publicar una planilla de cálculo en formato propietario incrementa el coste de reutilización que debe soportar el reutilizador para procesarla, ya que debe utilizar herramientas específicas para ello y la explotación de estos datos, en combinación con otros, resultará compleja. Sin embargo, si el publicador realiza el esfuerzo e introduce calidad en el proceso de publicación, incorporando tecnologías que permitan consultar, procesar y combinar datos para implementar servicios más complejos, permitirá al reutilizador implementar soluciones más eficaces, aumentando la eficiencia del organismo publicador y logrando un máximo rendimiento en procesos de interoperabilidad.

También es cierto que es fundamental el conocimiento de estas tecnologías por ambas partes, para que el proceso y la sostenibilidad de las iniciativas *open data* aporten los máximos beneficios en el medio plazo. Esto implica contar con la capacitación necesaria que, a la vez, resulta positiva ya que todo avance en el conocimiento, apropiación y buen uso de los estándares y estas tecnologías permite alcanzar el máximo potencial de la Web.

Como se verá más adelante en el capítulo, la solución está en el balanceo de las técnicas a utilizar en los procesos de publicación. Es necesario combinar adecuadamente ambas técnicas: publicar datos en crudo, tal cual están, lo que se conoce como *raw data*, o publicar datos en formatos estructurados no propietarios, que incluyan información de contexto y que faciliten su explotación gracias a la combinación con otros datos, lo que se conoce como *linked data*. Determinados conjuntos de datos, de alto valor y potencial para su reutilización, deben seguir el camino de máxima calidad en el proceso de publicación. Para lograr la mayor eficiencia, los procesos de publicación deben ir acompañados de la

información técnica necesaria, para obtener el óptimo rendimiento de estas tecnologías.

En cualquier proyecto de apertura de información la exposición de los datos no se debe limitar a la publicación de los mismos, a través de cualquier medio o formato, sino que es necesario buscar soluciones eficaces y eficientes, que sean robustas y tengan en cuenta los principios de modularidad, redundancia mínima, escalabilidad e interoperabilidad, entre otras buenas prácticas.

La publicación de la información a través de formatos estándar y abiertos es uno de los fundamentos del *open data*, ya que es básico para permitir la libre reutilización de la información por terceros. Este requisito es necesario, pero no siempre suficiente, ya que, como complemento y para facilitar la reutilización, se deberían ofrecer a los reutilizadores unos esquemas de representación interna consistentes. Es por ello que cualquier iniciativa de apertura de datos debería intentar seguir una estrategia gradual, orientada a llegar al más alto nivel tecnológico, dentro de la clasificación de las cinco estrellas para la publicación de *open data*¹ propuesta por Tim Berners-Lee, el inventor de la Web, donde se reconoce el uso de estándares para la gestión adecuada de la información que se expone. El nivel más alto sería cinco estrellas, y se consigue identificando y definiendo, a través de descripciones semánticas, la información a exponer.

★ Cosas disponibles en la Web (en cualquier formato)

★★ Datos estructurados (Por ejemplo, Excel en lugar de una imagen de una tabla)

★★★ Formatos no propietarios (Por ejemplo, CSV en lugar de Excel)

★★★★ URLs que identifiquen cosas

★★★★ Enlazar los datos con datos de otros lugares

Escala de las 5 estrellas de Tim Berners-Lee

.....
¹ Linked Open Data star scheme by example: <http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/>

Aunque la excelencia tecnológica se consigue al exponer la información como datos enlazados, el tipo de formato de representación ideal dependerá de las necesidades de los agentes reutilizadores. No siempre el nivel más elevado –cinco estrellas– es deseable, ya que no todos los colectivos reutilizadores están familiarizados con esas tecnologías, o la información a publicar tiene un carácter especial y debería estar representada mediante mecanismos concretos, por ejemplo, servicios web o estándares XML, para representación de información geográfica.

Siguiendo una estrategia de publicación gradual, y teniendo en cuenta el paradigma de los datos enlazados como objetivo, se conseguirán los mayores beneficios, tanto para las entidades que publican la información como para los consumidores finales. Si se representa la información mediante estas tecnologías, y respetando los principios de la arquitectura Web,² se conseguirá que sea posible la navegación, tanto por personas como por máquinas, entre los distintos datos relacionados, y explorar nuevas vías para encontrar información relevante, además de proporcionar opciones de búsqueda y consulta mucho más sofisticadas. Gracias también a que los resultados de las consultas serán datos estructurados, y no solo simples textos o enlaces, podrán a su vez ser reutilizados por cualquier aplicación de forma automatizada.

Al unir la filosofía de los datos abiertos del sector público y las tecnologías *linked data*, se consigue el *government linked data*, una herramienta perfecta a la hora de favorecer la reutilización de información por parte de terceras personas, en formas completamente nuevas e impredecibles, y al proporcionar un modelo de datos flexible que permite integrar fácilmente la información proveniente de distintas fuentes.

Historia de la interoperabilidad

A lo largo de la historia de la informática, los estándares han sido la piedra angular en la interoperabilidad entre sistemas

.....

² Architecture of the World Wide Web, Volume One: <http://www.w3.org/TR/webarch/>

distribuidos y heterogéneos. De la misma forma, la interoperabilidad entre reutilizador y publicador es crucial para la apertura de datos, ya que el objetivo de la exposición de los datos es la reutilización, muchas veces de forma automática, de los mismos por parte de cualquier agente.

El Lenguaje de Marcado eXtensible, o XML, fue uno de los primeros estándares publicados por el principal organismo estandarizador de la Web, el World Wide Web Consortium (W3C). La primera versión data de 1998, y desde entonces se ha convertido en parte esencial de las comunicaciones entre máquinas. El XML ha sido considerado por muchos expertos como el formato definitivo para la interoperabilidad, siempre y cuando se considere un esquema de representación de la información básico, compartido entre ambas partes: esquema XML. Estos esquemas están hechos *ad-hoc* para solucionar problemas concretos, lo que permite una interoperabilidad limitada por las siguientes razones:

- Existe ambigüedad lingüística. Ambas partes deben compartir exactamente el mismo esquema, por lo tanto, el idioma de representación del esquema.
- No existe flexibilidad a la hora de realizar una actualización o extensión del esquema. Cualquier modificación sobre el esquema anterior puede afectar a los sistemas que estén funcionando. Solo se considera una buena solución para problemas acotados, y sobre los cuales exista certeza que no cambiarán.
- La información solo se puede estructurar en forma de árbol, con las limitaciones que ello conlleva al no permitir conexiones flexibles entre la información representada en los documentos.
- Los esquemas XML están pensados para ser interpretados por máquinas, y es complicado ofrecer una representación fácilmente comprensible para los humanos.

OPEN LINKED DATA: LA NUEVA FRONTERA

A finales de los 90, Tim Berners-Lee, en busca del perfeccionamiento de la web tradicional –representada como una web de documentos enlazados– propuso una evolución de la misma como solución para el manejo de los datos, basándose en la representación semántica de la información y utilizando la arquitectura básica de la World Wide Web.³ De esta forma, la Web se comportaría como una gran base de datos, que permitiese la interoperabilidad universal entre cualquier sistema procesado por máquinas, independientemente de las tecnologías utilizadas para el tratamiento de la información, la representación de la propia información (formato), y ofreciendo la potencia de la Web para hacer infinitamente escalable dicha base de datos.

Este nuevo concepto de web, acuñado como Web Semántica –o Web 3.0– es únicamente una adaptación de la web tradicional, basada en los mismos principios tecnológicos y funcionando de forma transparente para el usuario, pero ofreciendo gran potencia de procesamiento automatizado. El paradigma de vinculación entre datos representados de forma semántica da lugar al término *linked data*.

La Web Semántica ofrece la solución a las limitaciones expresadas anteriormente, ya que evita las ambigüedades del idioma, ofrece una flexibilidad plena a la hora de extender los vocabularios y esquemas, así como en la representación de la información.

Pioneros del open data

Dos son las iniciativas que a nivel mundial, y en un ejercicio de transparencia sin precedentes, iniciaron el camino hacia modelos de gobierno más abiertos, en los que el objetivo es utilizar las ventajas que ofrecen las tecnologías de la Web para impulsar la rendición de cuentas y la evaluación de políticas públicas, evolucionando la administración de tener la consideración de propietario de la información a cumplir el papel de gestor de la misma.

.....
³ Architecture of the World Wide Web, Volume One : <http://www.w3.org/TR/webarch/>

- **Estados Unidos**, a través de la figura de Barack Obama, quien impulsó mediante una directiva de gobierno abierto la creación de un sistema de transparencia, participación pública y colaboración, con el convencimiento de que dicha estrategia fortalecerá la democracia y promoverá la eficiencia y efectividad en el gobierno. Los pilares de la estrategia de Estados Unidos se basan en la publicación en línea de información gubernamental, la mejora en la calidad de dicha información, la promoción e institucionalización de una cultura de gobierno abierto y la creación de un marco común que le de soporte.
- **El Reino Unido**, mediante la figura de Gordon Brown y posteriormente de David Cameron, ha iniciado también un proceso de apertura, mediante la publicación de información que obra en poder de la administración, con el objetivo de ayudar a los ciudadanos a entender cómo funciona el gobierno y cómo se llevan a cabo las distintas políticas. La principal motivación, en el caso del Reino Unido, es el convencimiento de que esta información pública, de fácil acceso, ayudará a que los ciudadanos puedan tomar mejores decisiones, y realicen sugerencias sobre las políticas de gobierno, basadas en la información recibida.

Ambas iniciativas comparten el objetivo común de publicar los conjuntos de datos de los que dispone la administración, de forma proactiva, y de que esa información vuelva a sus verdaderos dueños –la ciudadanía– en forma de servicios innovadores, creando al mismo tiempo procesos más eficientes de funcionamiento interno y externo que mejoren las operaciones e intercambios dentro de la propia administración.

Así pues, las iniciativas pioneras de gobierno abierto se basaron en un mismo concepto, los datos abiertos de gobierno o, lo que es lo mismo, la publicación de la información que obra en poder de las distintas agencias de la administración, haciéndola accesible a todo el mundo bajo las siguientes premisas:⁴

.....
⁴ Open Government Data Principles: http://resource.org/8_principles.html

OPEN LINKED DATA: LA NUEVA FRONTERA

1. La información tiene que ser completa.
2. Los datos deben provenir directamente de la fuente, sin transformación alguna.
3. La publicación debe realizarse a tiempo para preservar el valor de la información.
4. La información tiene que estar disponible para el mayor número de usuarios posibles.
5. Los datos deben presentar una estructura mínima, que permita su tratamiento automatizado.
6. La publicación tiene que estar ausente de toda discriminación y disponible para cualquiera.
7. Los formatos de publicación deben ser no propietarios.
8. La información debe estar libre de restricciones de derechos de autor o patentes (o ser reutilizable mediante licencias de tipo “atribución”).

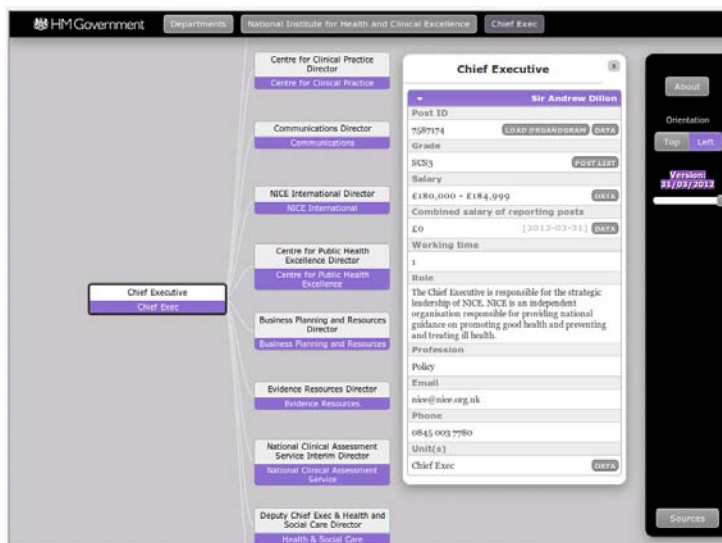
Si bien ambas iniciativas comparten el objetivo común de facilitar el acceso a la información y la interoperabilidad, y permitir su reutilización, su estrategia a seguir difiere en un aspecto fundamental ya que, mientras la iniciativa de Estados Unidos, bajo la dirección estratégica de Vivek Kundra, se centró inicialmente en la publicación de la mayor cantidad de información en el menor plazo posible, volcando la información en crudo, tal y como está disponible, en el caso del Reino Unido se realiza un tratamiento previo de los datos, para reducir los esfuerzos necesarios en su posterior reutilización.

Linked Open Data en data.gov.uk

El gobierno británico, confiando en el asesoramiento de grandes expertos de la talla del propio Tim Berners-Lee, o el Profesor Nigel Shadbolt (Universidad de Southampton), diseñaron el proyecto de apertura de datos más completo de los que se conocen hasta el momento: data.gov.uk. Este no era un portal en el que se listaban más conjuntos de datos, ni que estos fueran más útiles que otros, sino que habían expuesto una gran cantidad de información básica –infraestructuras, servicios públicos, educación o geografía– en *linked data*, lo que se constituye como la iniciativa pionera de *government*

linked data, o datos del sector público enlazados. El gran éxito e impacto de este portal llevó al resto de las iniciativas incipientes a plantearse la estrategia de exposición de los datos.

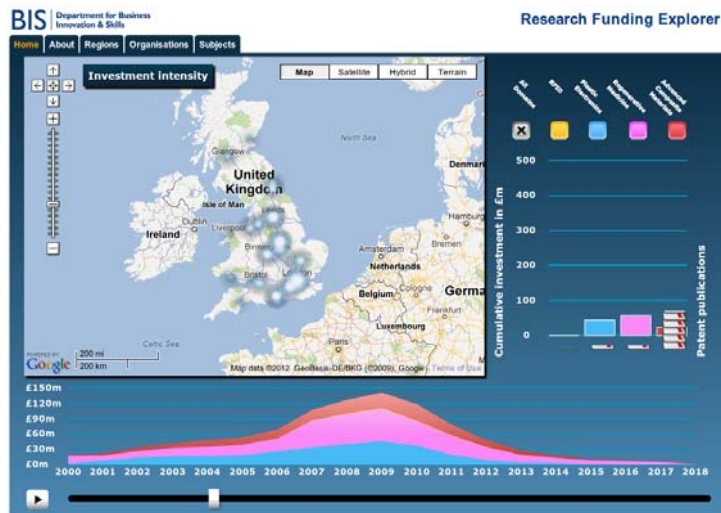
Aunque el beneficio de la aplicación de los principios de los datos enlazados era patente, la ausencia de conocimiento en el manejo de estas tecnologías ponía en entredicho la utilidad de estos formatos. La respuesta por parte del gobierno fue formar a los desarrolladores para el uso de estas tecnologías semánticas y, a su vez, ofrecer un sistema alternativo para hacer las consultas eficientes, utilizando los servicios web tradicionales. Para ello se creó un mecanismo –interfaz de programación– que permite la consulta sobre el almacén semántico, y el procesar la información en formatos “tradicionales”, como CSV o XML.



Organigrama del gobierno Británico⁵, cuya información está representada en *linked data*.

⁵ Organogram Viewer: <http://reference.data.gov.uk/gov-structure/organogram/>

OPEN LINKED DATA: LA NUEVA FRONTERA



Visualización⁶ de las inversiones para la investigación en Reino Unido.

Como evolución natural, tras meses de funcionamiento, la iniciativa estadounidense de datos abiertos (data.gov) apostó por la creación de una sección semántica dentro de su portal.⁷

Una base de datos global

Las tecnologías de la Web Semántica permiten representar, de una forma estándar, cualquier tipo de información, para que pueda ser gestionada por procesos automáticos: consultas, combinación de datos, inserciones, etcétera. El objetivo de todo esto es tener una gran base de datos global, con acceso desde cualquier dispositivo conectado a la Web.

⁶ Research Funding Explorer: <http://bis.clients.talis.com/>

⁷ Semantic Web Data.gov: <http://www.data.gov/semantic>

Esta base de datos es mucho más potente que una relacional, ya que ofrece la flexibilidad de extensión, sin modificar las estructuras existentes, brinda también mecanismos para añadir lógica a las relaciones entre recursos, lo que permitirá realizar operaciones de consulta avanzada, incluso inferencias. La representación de los datos no se hace de forma tabular, o en árbol –como en XML–, sino en forma de grafo. Esto permite relacionar recursos directamente, de cualquier forma.

Cómo se hace *linked data*

La clave para conseguir esta gran base de datos global es la identificación de los acervos y los vínculos existentes entre los distintos recursos de información y representarlos adecuadamente, para que personas y máquinas puedan procesar el contenido representado. Para ello se cuenta con las tecnologías de la Web Semántica del W3C,⁸ una gran pila de tecnologías multidisciplinares que permiten realizar todas las operaciones sobre los recursos: identificación, definición, consulta, combinación, inferencias lógicas, etcétera. Para publicar información siguiendo los principios del *linked data*,⁹ se debe cumplir con una serie de premisas:

- Identificación unívoca de los elementos de información mediante Identificadores de Recursos Uniformes (URI¹⁰).
- Proporcionar información útil cuando se resuelve un URI, describiendo los recursos, tanto en formato legible para los humanos –documentos HTML–, como mediante representaciones semánticas, siguiendo el modelo de la Infraestructura para la Descripción de Recursos (RDF).¹¹

.....
⁸ W3C Semantic Web: <http://www.w3.org/2001/sw/>

⁹ Linked Data design issues: <http://www.w3.org/DesignIssues/LinkedData.html>

¹⁰ Naming and Addressing: <http://www.w3.org/Addressing/>

¹¹ Resource Description Framework (RDF): <http://www.w3.org/RDF/>

- Añadir enlaces a otra información para crear una auténtica web de datos enlazados, que permita descubrir nuevos elementos relacionados, a través de las referencias en forma de enlaces RDF.

Todas estas características son las que hacen del *government linked data*, una herramienta perfecta a la hora de favorecer la reutilización de información por parte de terceras personas, de formas completamente nuevas e impredecibles, al proporcionar un modelo de datos flexible que permite integrar fácilmente la información procedente de distintas fuentes.

Identificación de recursos mediante URI

Los identificadores basados en URI permiten establecer referencias en la Web para cualquier recurso, lo que ofrece una forma de identificación universal por medio de la cual cualquier máquina o persona puede designar conceptos, elementos físicos o virtuales, de forma unívoca.

Para una gestión óptima de la información, es necesario establecer esquemas de definición de los identificadores, considerados como patrones de diseño de los URI, adecuados para cada tipo de información que se identifique: vocabularios u ontologías, conceptos, instancias o individuos, conjuntos de datos, etcétera. Este esquema de construcción de los URI constituye una guía de estilo en la creación de identificadores, de forma que ofrece homogeneidad, coherencia y uniformidad en la representación de recursos semánticos.

Asimismo, una buena política en la creación de URI dictará recomendaciones para componer las cadenas de caracteres, donde los términos que las compongan incluyan información comprensible y significativa, que describa la naturaleza del recurso que se identifica, el organismo de procedencia u otra información que permita a los agentes reutilizadores interpretar el identificador, e inferir intuitivamente el tipo de recurso esperado tras el URI.

Las características básicas a tener en cuenta para la definición de un esquema de URI se recogen en la especificación de buenas prácticas del W3C para la utilización de URI en la Web Semántica.¹² Los requisitos necesarios para diseñar el esquema de URI son los siguientes:

- Se debe utilizar el protocolo HTTP,¹³ de forma que se garantiza la resolución de cualquier URI en la web.
- La estructura de composición de URI debe de ser consistente y extensible.
- Los URI deben seguir una estructura de composición comprensible y significativa.
- No se debe exponer información sobre la implementación técnica de los recursos que representan los URI.
- Los URI deben cumplir el principio de persistencia, lo que significa que los que ya han sido creados previamente nunca deberían variar, y el contenido al que hacen referencia debería ser accesible. En el caso de necesidad de cambio o eliminación del recurso al que apunta un identificador, se debería establecer un mecanismo de información sobre su estado, usando los códigos de estado de HTTP. En el caso de poder ofrecer una redirección a la nueva ubicación del recurso, se utilizarían los códigos de estado HTTP 3XX, o para indicar que un recurso ha desaparecido permanentemente se utilizaría el código de estado HTTP 410.

A continuación se muestra un ejemplo de patrón de URI, que sigue estas buenas prácticas, así como un ejemplo de implementación que identifica a un recurso geográfico, como la Provincia de Madrid, donde se señalan la base del descriptor específica, la procedencia del dato – en este caso, Gobierno Español (data.gob.es)–, el carácter de la información (un recurso), el sector al que corresponde esa información (sector-público), la temática del conjunto de datos (información sobre el territorio), el nombre del

.....

¹² Cool URI for the Semantic Web: <http://www.w3.org/TR/cooluris/>

¹³ HTTP – Hypertext Transfer Protocol: <http://www.w3.org/Protocols/>

OPEN LINKED DATA: LA NUEVA FRONTERA

conjunto de datos –o *dataset*– (provincia), y el identificador concreto de la provincia (Madrid).

Patrón de URI	<code>http://{BASE}/{TIPO-RECURSO}/{SECTOR}/{TEMA}/{DATASET}/{ID}</code>
Ejemplo	<code>http://datos.gob.es/recurso/sector-publico/territorio/Provincia/Madrid</code>

Cualquier URI puede representarse de forma simplificada, a través de las abreviaturas de nombres en sus espacios, comportándose como los espacios de nombres en XML. Por ejemplo, la propiedad “seeAlso”, definida por el vocabulario del esquema RDF, se identifica a través del URI “`http://www.w3.org/2000/01/rdf-schema#seeAlso`”, permitiéndose la representación simplificada “`rdfs:seeAlso`”.

Descripción de los recursos

La definición de recursos se realiza semánticamente mediante el modelo RDF, o infraestructura de descripción de recursos, que permite representar a estos y sus relaciones entre sí. RDF no es un formato en sí, sino que existen múltiples variantes de representación, como RDF/XML¹⁴, N3,¹⁵ o Turtle,¹⁶ lo que permite el procesamiento automatizado de las representaciones.

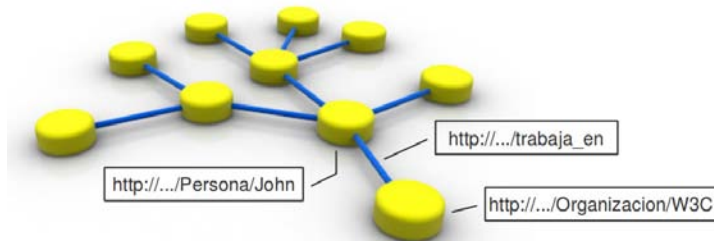
Este modelo se basa en tripletas (sujeto, predicado, objeto) que permiten enlazar grafos de recursos en forma escalable y distribuida; por ejemplo, “John trabaja_en W3C”. Cada uno de los elementos de las tripletas se identifica mediante URI, y se describe individualmente con RDF.

.....

¹⁴ RDF/XML Syntax Specification (Revised): <http://www.w3.org/TR/REC-rdf-syntax/>

¹⁵ Notation 3 Logic: <http://www.w3.org/DesignIssues/Notation3.html>

¹⁶ Terse RDF Triple Language: <http://www.w3.org/TR/turtle/>



Ejemplo de grafo RDF.

Para la definición en formato semántico es necesario disponer de un vocabulario u ontología que recoja, de una manera procesable automáticamente, la información sobre los tipos de recursos a los que se refiere el conjunto de datos (clases) y las características que definen a cada uno de estos recursos (propiedades). En ocasiones, para determinadas casuísticas, ya existirán vocabularios estándar, o de uso frecuente, que reflejen el ámbito semántico al que se refiere el conjunto de datos; otras veces, será necesario crear el vocabulario desde cero, o evolucionar los ya existentes.

Vocabularios

Desde el punto de vista técnico, estos vocabularios se definen utilizando el lenguaje RDF Schema¹⁷ (RDF-S), Web Ontology Language (OWL¹⁸), o más frecuentemente una combinación de ambos. Por su menor expresividad, RDF-S se utiliza sobre todo en vocabularios orientados a la compartición de datos entre distintas fuentes de información, mientras que OWL permite formular expresiones semánticas más ricas, pero técnicamente resulta más complejo y puede tener un soporte limitado por las aplicaciones de

¹⁷ RDF Vocabulary Description Language 1.0: RDF Schema: <http://www.w3.org/TR/rdf-schema/>

¹⁸ Web Ontology Language (OWL): <http://www.w3.org/2004/OWL/>

destino. Los vocabularios, al estar creados con estos mecanismos, son extensibles, y se posibilita la modificación o actualización futura sin la necesidad de grandes impactos en las implementaciones existentes.

De forma general, cuando exista un vocabulario aplicable al conjunto de datos, resulta preferible reutilizarlo o extenderlo para ajustarlo a la naturaleza de los recursos, en lugar de crear uno nuevo desde cero. Al reaprovechar un vocabulario ya existente se permite la interoperabilidad entre los agentes que los conozcan, y que los usuarios reutilicen los datos compartidos y puedan ser cruzados con datos de terceros. Hay varios criterios que deberemos tener en cuenta a la hora de seleccionar un vocabulario externo para representar los datos:

- **Uso frecuente.** Existen pocos vocabularios publicados como estándar, pero hay una serie de ellos que son internacionalmente reconocidos y ampliamente usados, como Dublin Core,¹⁹ FOAF,²⁰ SKOS,²¹ o vCard.²² Los vocabularios de uso frecuente facilitan en gran medida la reutilización, por lo que deben tenerse en cuenta, siempre que su significado sea adecuado.
- **Cobertura y expresividad.** Debe valorarse si el grado de expresividad es suficiente para el nivel de desagregación de la información a representar, y si permite definir todos los conceptos y relaciones.
- **Actualización.** Debe tenerse en cuenta si el vocabulario a reutilizar está siendo mantenido activamente, y si existe una política clara de actualización, mecanismos de recepción de incidencias, etcétera.

.....
¹⁹ DCMI Metadata Terms: <http://dublincore.org/documents/dcmi-terms/>

²⁰ The Friend Of A Friend Project: <http://www.foaf-project.org>

²¹ SKOS Simple Knowledge Organization System: <http://www.w3.org/2004/02/skos/>

²² Representing vCard Objects in RDF: <http://www.w3.org/TR/vcard-rdf/>

- **Exposición adecuada.** El vocabulario a reutilizar debe estar publicado de la misma forma que cualquier recurso semántico, es decir, empleando URI resolubles para identificar las clases y propiedades que lo definan, y con descripción, tanto para máquinas, como documentación para las personas.

En el caso de que se cree o se extienda un vocabulario, siempre que se considere maduro y se comience a usar para el modelado de recursos, debería ser publicado en la Web, para posibilitar posteriores reutilizaciones por otras partes interesadas.

Sistemas de organización del conocimiento

SKOS (Simple Knowledge Organization System) es un vocabulario estándar que permite definir sistemas de organización del conocimiento, tales como glosarios, taxonomías, tesauros, etcétera. A través de SKOS se pueden definir modelos de la realidad, representados de forma semántica, que servirán para establecer esquemas de conceptos sobre dominios específicos.

En la actualidad ya son muchos los organismos que han puesto a disposición los esquemas de conceptos y vocabularios que gestionan. Entre estos destacan los vocabularios y taxonomías publicadas por la Biblioteca del Congreso (de Estados Unidos)²³, las bibliotecas nacionales alemana, húngara y francesa, la NASA, la FAO, el New York Times, y varios gobiernos en todo el mundo. El W3C mantiene una lista²⁴ de conjuntos de datos modelados con SKOS por todas estas organizaciones, y que expuestos en la Web para su reutilización.

.....
²³ Linked Data Service – Library of Congress: <http://id.loc.gov>

²⁴ SKOS/Datasets: <http://www.w3.org/2001/sw/wiki/SKOS/Datasets>

Datos enlazados

Tras la representación de los recursos con datos propios de la organización –creación del grafo de recursos semánticos del dominio de información determinado– se recomienda el enriquecimiento de esta información mediante enlaces a otros conjuntos de datos existentes en la propia organización o en fuentes de datos externas. A esto se le conoce como *linking data*, o vinculación de datos, y es el objetivo principal del paradigma de los datos enlazados, ya que este enlace, a través de propiedades RDF, complementará y enriquecerá las descripciones de los recursos con información expuesta por expertos en dominios concretos.

La distribución de conocimiento permite a cada usuario de esta web de datos especializarse en su dominio, y utilizar libremente la información relevante para complementar sus recursos. Esto reduce costes y fomenta la innovación; cada agente podría centrarse únicamente en su ámbito y destinar sus esfuerzos a mejorar solo sus productos.



Varios dominios de información distribuidos en la web de los datos.

La relación que se establezca con las fuentes externas se definirá mediante propiedades acordes al tipo de información, y a la confianza en los datos enlazados. Puede utilizarse cualquier propiedad que se considere adecuada, preferentemente procedente

de los vocabularios recomendados o de los propios, considerando estrictamente el significado de la propiedad o término que establece el vínculo. Ejemplos de propiedades ampliamente utilizadas, que permiten hacer referencia a recursos con similitudes son: “*rdfs:seeAlso*”²⁵ y “*owl:sameAs*”.²⁶ La primera (*see also* o “ver también”) indica una referencia a un recurso similar al actual, mientras que la segunda (*same as* o “igual que”) define equivalencia entre los recursos enlazados. Ante la necesidad de indicar que un recurso es exactamente igual que otro, siempre que se confíe en la fuente, se recomienda usar la propiedad “*owl:sameAs*”.

Los datos enlazados en la práctica

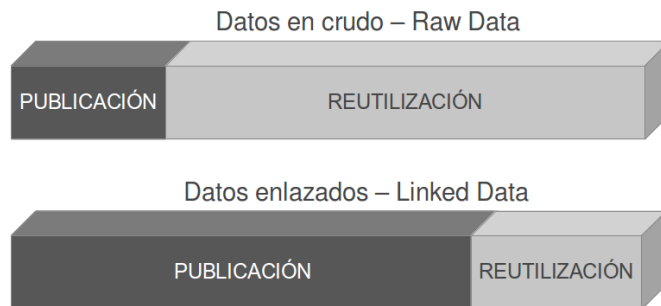
Como se ha comentado anteriormente, uno de los principales retos para solucionar el paradigma de los datos enlazados es la interoperabilidad entre sistemas, algo necesario en las administraciones públicas debido al carácter multinivel y distribuido que tienen sus distintos órganos y a la necesidad de comunicación con agentes externos (otras administraciones públicas, ciudadanía o empresas).

Por lo general, los datos enlazados permiten una reutilización de la información con un esfuerzo menor que el que se requiere para la de datos en crudo y en formatos no semánticos. SPARQL es el lenguaje y protocolo que permite consultar y acceder a las bases de datos y procesar los datos enlazados, independientemente de donde se encuentren o de su carácter.

.....
²⁵ <http://www.w3.org/2000/01/rdf-schema#seeAlso>

²⁶ <http://www.w3.org/2002/07/owl#sameAs>

OPEN LINKED DATA: LA NUEVA FRONTERA



Esfuerzo de publicación y reutilización en función del formato del dato.

Muchos organismos ya se están beneficiando de estas tecnologías y mecanismos, ya que son una gran ayuda en la fusión de empresas o en la reestructuración de las administraciones públicas, manteniendo los sistemas de información interoperables entre sí y permitiendo una plena integración de los datos sin costosas plataformas difíciles de mantener.

Otro caso de uso de los datos enlazados en la industria es la gestión eficiente de contenidos, basándose en los de agentes externos a la propia organización. El ente de comunicación británico BBC es uno de los pioneros en la adopción de los datos enlazados, no solo en su utilización interna para construcción de estructuras de la información o para la apertura de sus propios datos, sino reutilizando contenido procedente de fuentes externas. Portales temáticos, como BBC Music,²⁷ utiliza información procedente de la DbPedia (proyecto que ha permitido representar semánticamente los artículos de la Wikipedia)²⁸ y de MusicBrainz²⁹ para enriquecer su contenido sobre artistas, obras, etcétera.

.....
²⁷ BBC Music: <http://www.bbc.co.uk/music>

²⁸ Wikipedia: <http://en.wikipedia.org>

²⁹ MusicBrainz: <http://musicbrainz.org>

Evolución hacia el *government linked data*

El W3C está impulsando el *government linked data*, y prueba de ello es la existencia del grupo³⁰ que lleva el mismo nombre –o el acrónimo GLD– que está desarrollando directrices, vocabularios y tecnologías que faciliten la adopción de los datos enlazados por parte de las iniciativas de apertura de datos de las administraciones públicas de todo el mundo.

Aunque existen grandes beneficios para las administraciones asociados a la adopción de estas tecnologías, aún son escasas las iniciativas de apertura de datos basadas en *linked data*. Esta tendencia está cambiando y se experimentará una evolución notable en los próximos años ya que, con la introducción de nuevos estándares, como el HTML5 (lenguaje de creación de páginas web) que permiten incrustar datos en los documentos HTML, cualquier administración pública puede realizar la exposición de datos enlazados, asociados a la información ya existente en sus gestores de contenidos.

De forma transparente para las personas, los datos estarían incluidos en el documento web como microdatos,³¹ o RDFa³² (Resource Description Framework in Attributes), que es un mecanismo genérico para incrustar definiciones modeladas en RDF en los atributos de documentos basados en XML o en el propio HTML; ello propicia la generación de datos enlazados sin necesidad de infraestructura adicional al propio gestor de contenidos. Aunque son numerosos los casos de adopción de RDFa por parte de la industria, como son los buscadores, tiendas de comercio electrónico o redes sociales, como Facebook, aún son pocos los proyectos impulsados por iniciativas públicas que hacen apertura de datos usando esta tecnología.

.....
³⁰ W3C Government Linked Data Working Group: <http://www.w3.org/2011/gld/>

³¹ Microdata HTML5: <http://www.w3.org/TR/microdata/>

³² RDFa Primer: <http://www.w3.org/TR/rdfa-primer/>

OPEN LINKED DATA: LA NUEVA FRONTERA

El portal legislation.gov.uk,³³ impulsado por los Archivos Nacionales, órgano dependiente del Ministerio de Justicia del Reino Unido, es uno de los proyectos pioneros en la utilización de RDFa para abrir los datos públicos. Este portal ofrece toda la información sobre la legislación británica a través de un sitio web con servicios adecuados para la búsqueda de información por parte de personas y, a su vez, publica toda la información estructuradamente, y marcada de forma semántica, para que pueda ser procesada automáticamente.

.....

³³ [legislation.gov.uk](http://www.legislation.gov.uk): <http://www.legislation.gov.uk>