Open Government Data en el mundo

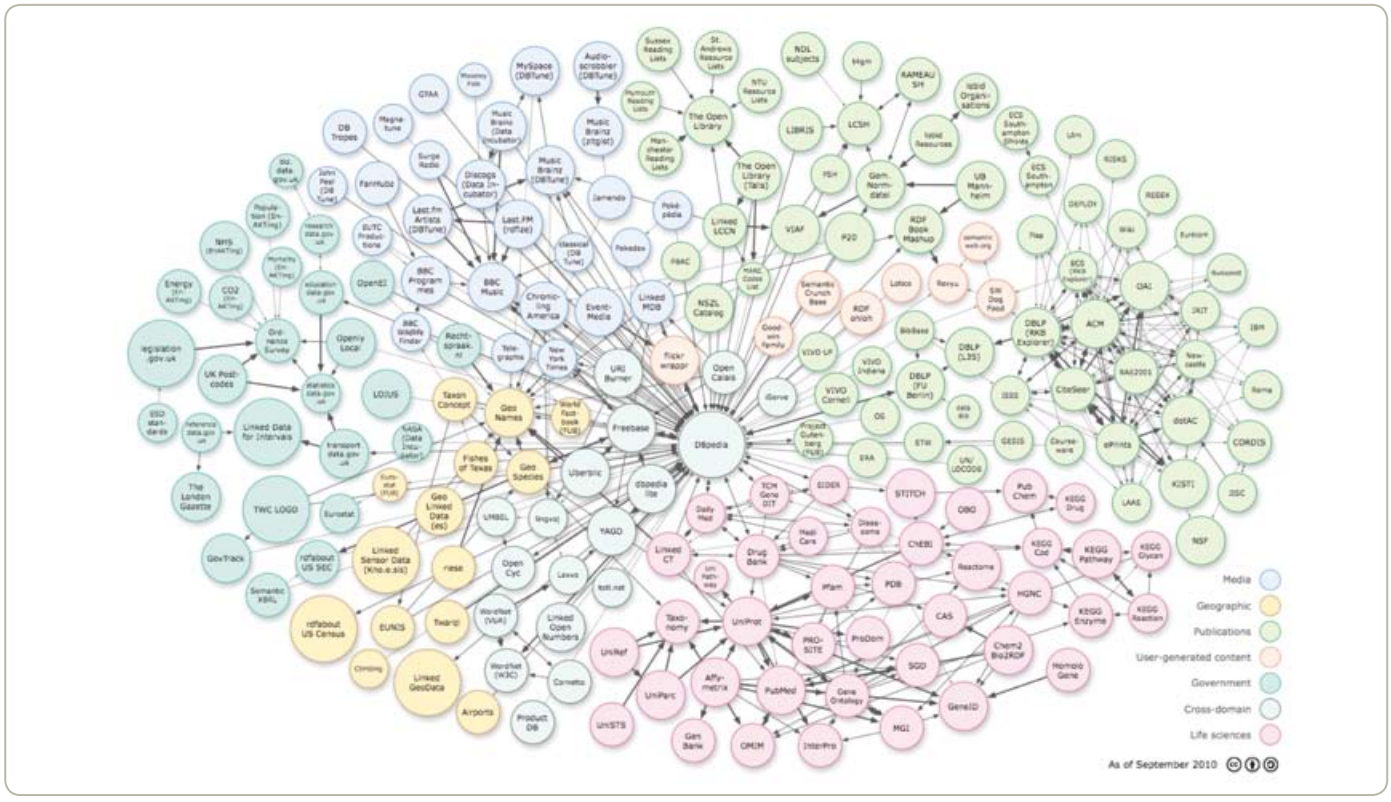
Álvaro Graves
Ingeniero Civil en Computación y Magíster en Ciencias mención Computación, Universidad de Chile. Estudiante de PhD en Cognitive Science, Tetherless World Constellation, Rensselaer Polytechnic Institute, Estados Unidos.
alvaro@graves.cl

Desde hace algunos años ha tomado fuerza la idea de publicar libremente datos de gobierno de distintos países, tanto a nivel nacional, regional y municipal. Este movimiento, conocido como Open Government Data (OGD) se ha extendido durante los últimos años y actualmente más de una veintena de países, incluyendo Estados Unidos y el Reino Unido, implementan portales de publicación de datos. Asimismo, este movimiento se ha visto fuertemente asociado a Linked Data, que consiste en una serie de principios para publicar datos usando tecnologías de la Web Semántica, que los hacen fácilmente procesables por máquinas. Esta simbiosis ha beneficiado a distintas organizaciones al interior del gobierno, así como a académicos, investigadores y ciudadanos en general. El presente artículo describe cómo los países han comenzado a adoptar OGD y cómo el uso de Linked Data ha ayudado en la publicación de datos.

¿QUÉ ES OPEN GOVERNMENT DATA?

Open Government Data consiste en un conjunto de principios que apuntan a que los datos generados o usados por los gobiernos debiesen estar a libre disposición y uso por parte de los ciudadanos. Existen varias razones que justifican esto: en primer lugar, los datos generados por el gobierno son financiados con los impuestos de todos. ¿No deberían todos los ciudadanos poder usarlos, dado que han pagado por ellos? En segundo lugar, el reuso de estos datos permite que otras personas se beneficien directa e indirectamente de estos, aumentando su valor y utilidad. En Estados Unidos, empresas como BrightScope.com (que reporta información sobre consejeros financieros) y aplicaciones como Roadify.com (que entrega información sobre transporte público de Nueva York en tiempo real) utilizan

Figura 1



La nube de Linked Data (los datasets y sus enlaces) en septiembre de 2010. Los datasets en verde de la izquierda, corresponden a datos de gobierno.

datos publicados por el gobierno para sus operaciones. En tercer lugar, la publicación de datos gubernamentales permite que la ciudadanía esté más informada sobre cuáles son las actividades del gobierno y cómo se realizan, aumentando la transparencia y el accountability de este último: por ejemplo, en Estados Unidos es posible ver qué funcionarios de la Casa Blanca han sido visitados, cuántas veces y por quién[1]. Finalmente, tecnologías como la Web permiten que el costo de publicar datos sea muy bajo: una vez que los datos han sido recolectados o generados y usados por el gobierno, el proceso de publicarlos es generalmente sencillo y simple.

En 2007 un grupo de expertos definió un conjunto de ocho principios que reflejan cómo los gobiernos debiesen publicar datos[2]: los datos deben ser completos, primarios, estar disponibles a tiempo, ser accesibles, fácilmente procesables por máquinas, no se debe discriminar a quienes lo soliciten, no deben estar en formatos propietarios y deben usar licencias abiertas. Existen por supuesto una serie de restricciones sobre qué cosas no pueden considerarse OGD: por ejemplo, es usualmente aceptado que la información personal de ciudadanos, así como datos que puedan afectar la seguridad nacional no deben ser publicados. A pesar de estas excepciones, se entiende como una buena práctica el que la opción por omisión sea publicar datos y el no hacerlo sea el caso particular.

LINKED DATA

Uno de los movimientos de mayor impacto en la Web Semántica es Linked Data[3], el cual consiste en una serie de principios para publicar datasets acerca de distintos temas, donde cada "cosa" (un auto, una persona, el día de ayer) tiene asignada una URI (similar a una dirección Web o URL). Como cada cosa es identificable por estas URIs, el siguiente paso es enlazar estos datasets, identificando qué URIs se refieren a la misma "cosa" o están relacionadas de alguna manera: de esta forma, es posible navegar por distintos datasets para obtener más información que la provista por una organización solamente. Por ejemplo, un médico puede buscar en DBpedia -una versión "semantificada" de Wikipedia- acerca de la proteína P53, encargada de la supresión de tumores y encontrará una descripción de ésta en varios idiomas, así como temas relacionados (oncología, proteínas, etc.). Luego, desde DBpedia es posible obtener las URIs con que esta proteína es descrita en otros datasets. Al acceder a estos nuevos datasets es posible encontrar qué enfermedades están asociadas a P53.

Es claro que las comunidades de OGD y Linked Data tienen mucho en común y pueden beneficiarse mutuamente, la primera usando Linked Data como plataforma, la segunda mostrando en OGD un caso de uso real. Actualmente, un porcentaje importante de la "nube" de Linked Data (los conjuntos de datos conectados) son datos de gobierno, como puede verse en la Figura 1.

Figura 2

Portal Data.gov del Gobierno estadounidense, permite buscar datasets relacionados con agricultura, defensa, medio ambiente y presupuestos, entre otros.

DESARROLLO DE OGD EN EL MUNDO

Las historias de OGD en Estados Unidos y en el Reino Unido son ilustrativas de cómo los gobiernos han adoptado distintos modelos de OGD y cuál ha sido su relación con Linked Data.

Estados Unidos: modelo Bottom-Up

(Disclaimer: estoy asociado con Tetherless World Constellation y he participado activamente en éste como parte del trabajo realizado por este laboratorio en colaboración con Data.gov).

En mayo de 2009, la administración del Presidente Barack Obama lanzó el sitio Data.gov[4] que fue la primera plataforma centralizada de publicación de datos en el mundo, construida por un gobierno. Comenzando con cerca de 40 datasets, actualmente provee sobre los 300.000, los cuales describen información relacionada con temas de energía, salud, migraciones, seguridad pública y muchos más. El uso de los datos ha sido aprovechado por una serie de aplicaciones, como DataMasher[5] (sitio especializado en crear *mashups*, es decir visualizaciones de cruza de datos) y Fly On Time[6] (sitio que permite saber cuántas son las demoras de vuelos en Estados Unidos), por nombrar algunas. Una de las prioridades de Data.gov era liberar la mayor cantidad de datos, bajo un proceso de publicación simple, por lo que se dio flexibilidad a los funcionarios de gobierno en cuanto a los mecanismos de publicación: es así que los datos han sido publicados principalmente como archivos XML, Excel, Comma-Separated Values (CSV), Really Simple Syndication (RSS), Keyhole Markup Language (KML o KMZ) y archivos Shapefile (SHP).

Otra medida tomada para simplificar el proceso de publicación fue apuntar a los datos localizados en los servidores de los organismos gubernamentales correspondientes, en vez de replicarlos en Data.gov; de esta forma se evitan problemas técnicos y se puede reusar buena parte de la infraestructura existente (por ejemplo, servicios que proveen feeds RSS). Desde hace algún tiempo, Data.gov (Figura 2) ha comenzando a publicar datos en RDF (Resource Description Framework), el lenguaje para datos en la Web Semántica. Este trabajo se ha hecho en conjunto con Tetherless World Constellation y ha implicado dos procesos paralelos: por un lado la conversión textual de los datos de manera automática, donde estos se extraen desde las tablas Excel y archivos CSV y se aplica una transformación genérica para generar RDF. El segundo proceso consiste en la publicación de datos mejorados, curados manualmente, donde se busca una representación más fidedigna de lo que los datos representan en el mundo real, que a la estructura de la tabla desde la que fueron sacados. Por ejemplo: la conversión automática de una tabla con nombre, apellido y dirección de una persona considerará los tres valores asociados a la misma entidad (la fila de la tabla); una versión mejorada considerará qué nombre y apellido pertenecen a una persona, mientras que la dirección está asociada a un lugar, el cual está relacionado con la persona, como se puede ver en la Figura 3.

Reino Unido: modelo Top-Down

En enero de 2010, el Gobierno del Reino Unido lanzó Data.gov.uk[7] (Figura 4). El enfoque británico fue diferente: se usó tecnología semántica y Linked Data desde un principio, por lo que en muchos casos (no todos) los datos están disponibles en RDF así como en su contraparte en formato CSV. Por ejemplo, cada escuela en el Reino Unido tiene una URI (por ejemplo, http://education.data.gov.uk/id/school/103335). El uso de Linked Data permite que al acceder a esta URI (sea posible obtener información relevante para la escuela: al usar un navegador como Firefox o Chrome obtenemos un documento HTML, pero también es posible escribir programas que vean los datos "puros" en RDF usando esta URI, los cuáles serán más fáciles de procesar que extraerlos desde el HTML.

Asimismo, el Gobierno británico dispuso de SPARQL endpoints (servicios Web donde es posible ejecutar consultas en SPARQL, el equivalente a SQL para datos semánticos) con información sobre distintas áreas (educación, transporte, etc.), de manera

que en muchos casos no es necesario descargar la información, sino que es posible consultarla directamente en los servidores del Gobierno.

CATÁLOGOS DE OGD

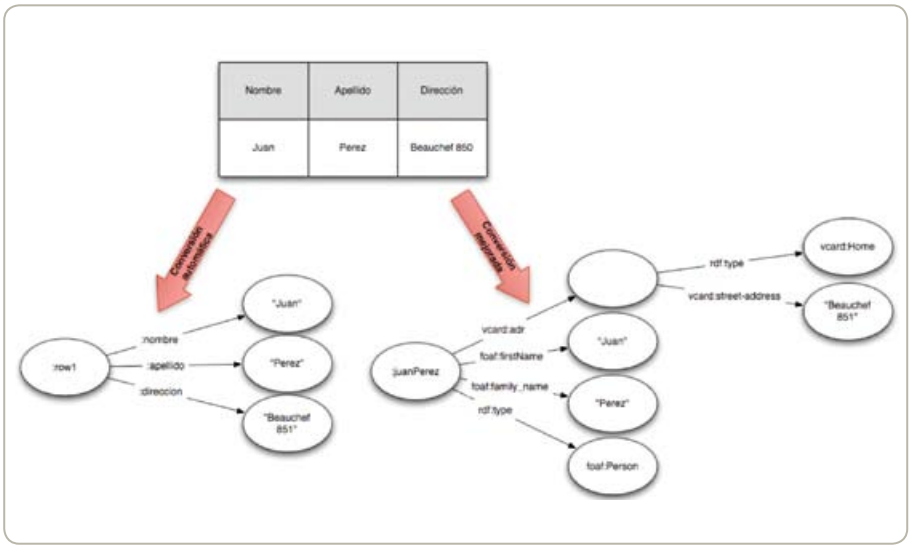
Como una forma de facilitar el acceso a esta gran cantidad de datos, existen varios esfuerzos por crear "metacatálogos" donde sea fácil buscar datasets disponibles en distintos portales. Así, por ejemplo, la Comunidad Europea ha trabajado en los últimos años para disponer de un portal centralizado que liste los datos de los países que la componen, tanto a nivel local, regional, así como nacional. Uno de los problemas es que al haber cientos de catálogos, no es fácil para los usuarios encontrar los datos que buscan, de manera que han creado PublicData.eu[8], el cual permite buscar en diversos portales de la Comunidad Europea. De esta forma, no se intenta replicar el trabajo hecho por otras organizaciones gubernamentales, sino agregarlo para facilitar la búsqueda por parte de los usuarios.

A nivel internacional, Tetherless World Constellation ha creado un catálogo de fuentes de datos de gobierno de diversos países y organizaciones internacionales, el cual se puede explorar seleccionando diversos criterios como país de origen y temas relacionados, entre otros[9]. Un esfuerzo similar ha realizado la fundación CTIC, la cual también provee un navegador[10] para buscar catálogos de datos por país y tipo, como se puede ver en la Figura 5.

DESAFÍOS

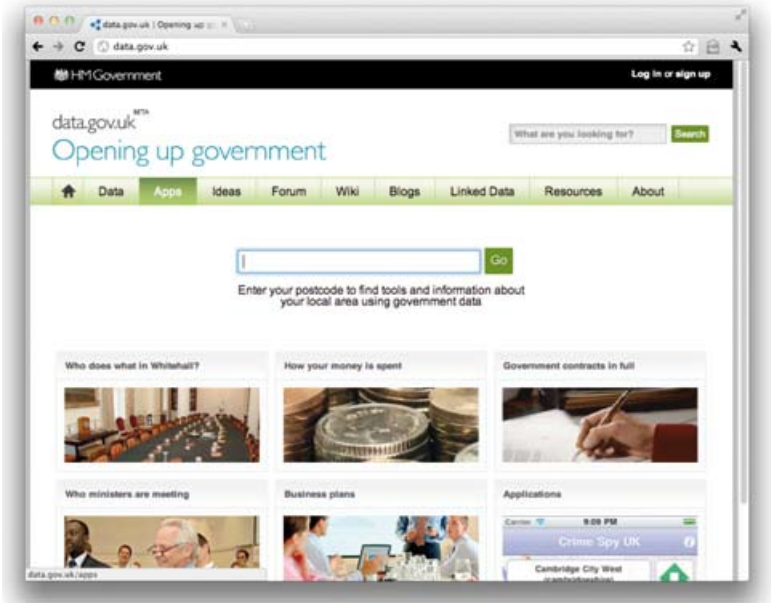
Se puede decir que OGD presenta desafíos en varios frentes, pero por brevedad sólo mencionaré los que parecen más relevantes. En primer lugar, es necesario un fuerte apoyo político y considerar a OGD como una política fundamental para mejorar la transparencia de un gobierno. Sin una valoración desde el mundo político, cualquier esfuerzo se va a quedar sólo en buenas intenciones. Más aún, este apoyo debe verse reflejado en una asignación de recursos, ya que como cualquier otra política pública, implementar OGD requiere tiempo y dinero para que los encargados lo puedan llevar a cabo. En segundo lugar, la cantidad, variedad y distribución de datos disponibles implican que se requiere especial preparación por parte de los organismos públicos: la experiencia en distintos países muestra que es necesario capacitar a quienes serán los encargados de OGD en cada órgano del Estado, lo que toma tiempo. En tercer lugar está el asunto de la calidad. Es claro que no todos los datos son igual de "buenos" en términos del "ruido" que poseen, cuán confiables son, cómo son representados, etc. Para ayudar a resolver esto, es necesario establecer una serie de

Figura 3



El proceso de conversión desde una tabla a RDF puede ser automático, llevando a una representación más cercana a la tabla (a la izquierda) o una conversión mejorada, reusando vocabularios (en este caso FOAF y vcard) y más cercana a lo que los datos describen (grafo a la derecha).

Figura 4



Portal Data.gov.uk del Gobierno del Reino Unido. En la versión actual se facilita la búsqueda de datos y aplicaciones relevantes para localidades específicas, basado en el código postal.

Figura 5

Países que poseen catálogos de datos públicos, según la Fundación CTIC.

métricas que ayuden a los consumidores de estos datos. Finalmente, quizás las preguntas más importantes que debiésemos tratar de resolver son: ¿cómo hacemos para que los ciudadanos comunes y corrientes puedan sacar el máximo provecho de estos datos sin tener que convertirse en hackers?, ¿qué tipo de servicios debiesen ofrecer los gobiernos para aumentar la participación ciudadana en las iniciativas de OGD?

Con todo lo anterior, queda la pregunta sobre cómo poder replicar estas iniciativas en otros países e instituciones. Por una parte, la experiencia muestra que no es necesario centralizar todos los datos, sino centralizar las búsquedas: los usuarios no tienen por qué cargar con la responsabilidad de saber dónde están los datos, sólo saber que pueden buscarlos en un solo sitio. Esto conlleva a que el repositorio debe coordinar con los diversos organismos proveedores de datos; lo anterior es posible usando vocabularios para describir catálogos de datos, tales como dcat[11] para comunicar qué datasets están disponibles. Asimismo, es importante publicar los datos en la mayor variedad de formatos posible, de manera de llegar a diferentes audiencias y disminuir las barreras para la creación de aplicaciones. Para lograr esto es recomendable tener un modelo de datos flexible desde el cual sea posible traducir y exportar a diferentes formatos. Es aquí donde RDF aparece como una excelente alternativa: convertir desde RDF a otros formatos resulta más fácil que desde, por ejemplo, CSV o Excel. Por otro lado, una crítica importante que se le ha hecho a Data.gov es la falta de recursos para mantener una comunidad de hackers y desarrolladores. El acceso a ejemplos de código, APIs (Application Programming Interface), tutoriales, documentación, etc. facilita el uso de los datos por parte de programadores, particularmente quienes desarrollan software en su tiempo libre. Otra crítica hecha a Data.gov (y en menor grado a Data.gov.uk) ha sido la calidad del sistema de búsquedas. Encontrar la información que se busca no resulta fácil, lo que desmotiva a los usuarios. Un esfuerzo para mejorar esto ha sido alpha.gov.uk, el cual ofrece sugerencias en una forma similar a lo que hace Google Instant[13]. Ésta y otras alternativas para mejorar las búsquedas pueden ser críticas para garantizar el éxito de un portal de OGD.

CONCLUSIONES

Este artículo ha hecho una breve revisión sobre qué es Open Government Data, su relación con Linked Data, así como ejemplos exitosos de la aplicación de estas tecnologías en gobiernos de distintas partes del mundo. Existen una serie de desafíos a la hora de implementar OGD: en general existe un conflicto natural entre la simplicidad de publicación y simplicidad de consumo de los datos y cada gobierno ha buscado un camino diferente para lidiar con este problema. Más aún, hacer fácil para la ciudadanía el usar estos datos sigue siendo un problema abierto. Sin embargo, ya es posible ver beneficios en el uso de estos datos por parte de empresas y desarrolladores para creación de aplicaciones y servicios. Asimismo, OGD ha mostrado que es posible transparentar las actividades del gobierno, facilitando la detección de potenciales fraudes e ineficiencias en la gestión.

Todavía hay mucho camino por recorrer para aprovechar todo el potencial que ofrece Open Government Data, pero la tendencia en el mundo es que poco a poco los gobiernos van abriendo más sus datos para que la ciudadanía pueda hacer uso de ellos tanto a nivel nacional, regional como local. BITS

REFERENCIAS

[1] http://bit.ly/WHvisitors

[2] http://www.opengovdata.org/home/8principles

[3] http://linkeddata.org

[4] http://data.gov

[5] http://www.datamasher.org/

[6] http://flyontime.us/

[7] http://data.gov.uk

[8] http://publicdata.eu

[9] http://logd.tw.rpi.edu/demo/international_dataset_catalog_search

[10] http://datos.fundacionctic.org/sandbox/catalog/faceted/

[11] http://vocab.deri.ie/dcat

[12] http://drupal.org

[13] http://www.google.com/instant/