Análisis de Herramientas para la Búsqueda Semántica de Documentos

Leonardo de-Matteis, Karina Cenci, Pablo Fillottrani[1]
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
{ldm,kmc,prf}@cs.uns.edu.ar

Resumen

Tanto los repositorios institucionales como los de contenido libre se benefician del uso de *frameworks* que permiten el acceso a metadatos dentro del marco general de la Web Semántica. Se destacan *OAI-PMH*, que permite la recolección de metadatos, y *Linked Data*, que mediante bases de datos *RDF* y el lenguaje *SPARQL*, constituye una mejora sustancial sobre el primero, definiendo prácticas para publicar, compartir y enlazar información disponible en la web mediante *URIs*, *HTTP* y modelos de grafos *RDF*. Este trabajo se propone mostrar comparaciones y aspectos relevantes entre ambos *frameworks*.

1. Introducción

En la actualidad, la información se encuentra disponible en grandes volúmenes de datos distribuidos en diferentes servidores y servicios sobre Internet. En la mayoría de los casos, esta información está dirigida al usuario final a través de páginas en la web pero no está disponible en forma estructurada y fácilmente procesable para procesos computacionales. En este contexto, la Web Semántica promueve actividades, procesos, estándares y tecnologías para el desarrollo de sistemas que faciliten la construcción de servicios basados en información que le otorguen un valor agregado. Puede sostenerse, entonces, que su objetivo consiste en poder procesar la información de tal manera de obtener conclusiones a partir de conjuntos de datos heterogéneos disponibles en la web. El procesamiento de tales datos mediante herramientas específicamente diseñadas permite producir resultados que no podrían lograrse con la mera búsqueda de información aislada y no relacionada. Así, algunos de los servicios que pueden darse mediante herramientas de este tipo incluyen, por ejemplo, la construcción de gráficas sobre cálculos estadísticos, el cálculo de erogaciones a nivel global de empresas u organizaciones, la publicación de información de organismos públicos procesada para informar a los contribuyentes, el procesamiento de metadatos para determinar relaciones entre los mismos y su posterior publicación en formatos diversos para uso en aplicaciones específicas, etc.

Para tratar de eliminar las barreras que complican o imposibilitan el acceso a los datos de diversos tipos de repositorios (tanto los que son de carácter institucional como también otros de contenidos libres) existen diversos *frameworks*. Entre ellos podemos mencionar dos como los más destacados y extendidos. En primer lugar, se destaca *OAI-PMH* (*Open Archives*

1 Com. de Investigaciones Científicas Prov. de Buenos Aires

Initiative - Protocol for Metadata Harvesting, 'Iniciativa de archivos abiertos – Protocolo para la cosecha de metadatos') que, como su nombre lo indica, permite recolectar metadatos. El segundo *framework* más destacado, *Linked Data*, define un conjunto de buenas prácticas para la publicación, compartición y enlazado de información disponible en la web, utilizando para ello *URIs*, el protocolo *HTTP* y modelos de grafos *RDF* (*Resource Description Framework*, 'Marco para la descripción de recursos') que permiten la descripción de elementos y de sus relaciones. La Web Semántica se ve reflejada, así, en el actual uso cada vez mayor de *Linked Open Data*[2] junto con las representaciones *RDF* y el lenguaje de consulta *SPARQL*, lo que supone una mejora cualitativa sobre el estándar *OAI-MPH*.

Si bien hoy en día hay muchas formas de acceder a la información disponible en la web, desde mecanismos simples de uso general como los *RSS*, más estructurados como *microformats*[3], genéricos vía de diferentes *APIs*[4] (que permiten la entrega de resultados en *JSON* y *XML*), y más avanzados utilizando definiciones precisas como las que caracterizan a *OAI-MPH* o *Linked Data*. Estos últimos están en relación directa con las características propias de la construcción y procesamiento de metadatos que posibilita la Web Semántica y en este artículo presentamos un análisis de ambos *frameworks*, tratando de determinar las mejores oportunidades de interrelación entre ambos.

2. OAI-PMH

2.1 Reseña histórica

El protocolo *OAI-PMH* surge a finales de la década del 1990 cuando, a partir de las iniciativas de investigadores y bibliotecarios en *Los Alamos National Laboratory*, iniciaron los primeros trabajos para su implementación. Después de los primeros artículos que comenzaron a divulgar las ideas surgidas en diversas reuniones de trabajo, estas derivaron en el nacimiento de una organización denominada *Open Archives Initiative* (*OAI*) con el objetivo de desarrollar y aplicar técnicamente mecanismos comunes de interconexión de archivos para compartir información de catalogación, lo que hoy conocemos bajo el término más general de *metadatos*.

El objetivo planteado consistía en lograr la difusión eficiente de contenido en Internet para alcanzar, así, un incremento cuantitativo de la disponibilidad global de publicaciones científicas. Desde entonces, el protocolo ha sido ampliamente adoptado para intercambiar datos bibliográficos.

Ahora bien, para comprender cabalmente su desarrollo temporal, debemos remitirnos a lo que se entendía por "archivo" cuando se formuló por primera vez este protocolo en el ámbito donde se desarrollaron las reuniones de trabajo iniciales y donde nacieron las comunidades de *eprints* (artículos de investigación, tesis, artículos en conferencias, capítulos de libros o bien,

2 *Linked Open Data* como método para la publicación de datos estructurados se basa en los cuatro principios definidos por Berners-Lee para *Linked Data* (véase sección 3.1) más un quinto: el contenido abierto o datos abiertos.

3 http://microformats.org

4 *API*: *Application Programming Interface.*

sencillamente, libros). En ese entonces, los "archivos" eran un sinónimo para el depósito de documentos científicos completos, dejando a un lado el concepto tradicional referente al hecho de almacenar y proteger cualquier tipo de información para uso posterior.

Por su parte, la idea asociada firmemente al término "archivo abierto" (*Open Archive*) era la de definir interfaces que facilitaran la disponibilidad de contenidos procedentes de diferentes orígenes, lo que hoy en día podríamos concebir como diferentes bases de datos fuentes, pertenecientes a diversas instituciones u organizaciones. Debe destacarse, en este marco, que el término *open* (abierto) no implica gratuidad o acceso ilimitado a información proveniente de dichas fuentes.

Con el tiempo surgió naturalmente la idea de incluir también el acceso a diversos materiales digitales disponibles, permitiendo el desarrollo de otras aplicaciones fuera de la comunidad inicial. Así que, en la actualidad, los desarrollos propuestos por la *OAI* no se enfocan solo en publicaciones científicas sino también en la comunicación de metadatos sobre cualquier material almacenado electrónicamente.

2.2 Descripción técnica

A continuación vamos a describir en forma sintética la estructura del modelo *OAI-MPH*. En primer lugar, hay que mencionar que en la implementación se encuentran definidas claramente dos entidades principales: los *Data providers* y los *Service providers* (Gráfico 1).

El primer tipo de entidad consiste en repositorios que permiten acceder a sus metadatos a través del protocolo *OAI-MPH* mientras que las segundas son aquellas que realizan solicitudes para extraer información de los metadatos.

Por otra parte, en el protocolo se definen un conjunto de seis tipos de peticiones (o verbos) que son invocados a través del protocolo *HTTP* (el cual da soporte a toda la web). Para ello, *OAI-MPH* hace uso de las peticiones básicas provistas por este protocolo, solicitando los servicios a través de peticiones con argumentos especificados que se envían a través de los parámetros encapsulados en los clásicos métodos *GET* o *POST*. Las peticiones disponibles e implementadas en todos los *Service Providers* son las siguientes:

GetRecord: Recuperar un registro en particular. Sus argumentos son: *identifier*: identificador del registro solicitado; *metadataPrefix*: formato bibliográfico en que se lo desea obtener.

Identify: Obtener información sobre un servidor de datos, no necesita parámetros y en la respuesta se especifican diferentes elementos, entre ellos podemos citar: *repositoryName, baseURL, protocolVersion, adminEmail*, etc.

ListRecords: Recuperar los registros completos. Sus argumentos son: *from* (opcional); *until* (opcional); *metadataPrefix* (requerido); *set* (opcional); *resumptionToken* (exclusivo).

ListIdentifiers: Recuperar solo los encabezamientos de los registros, es una forma abreviada de *ListRecords* y posee definidos los mismos argumentos.

ListSets: Permite recuperar un conjunto de registros de un repositorio. Es decir, permite obtener registros creados opcionalmente por el servidor para facilitar una obtención selectiva

de registros. Por ejemplo: un cliente puede pedir que se recuperen solo los registros pertenecientes a una determinada clase.

ListMetadataFormats: Obtener una lista de formatos de metadatos que contiene el repositorio de datos. El protocolo soporta diferentes formatos para expresar los metadatos, pero requiere que los servidores de datos permitan el acceso utilizando *Dublin Core*, codificado en *XML*. Cada proveedor de datos es libre de ofrecer los registros en otros formatos adicionales.

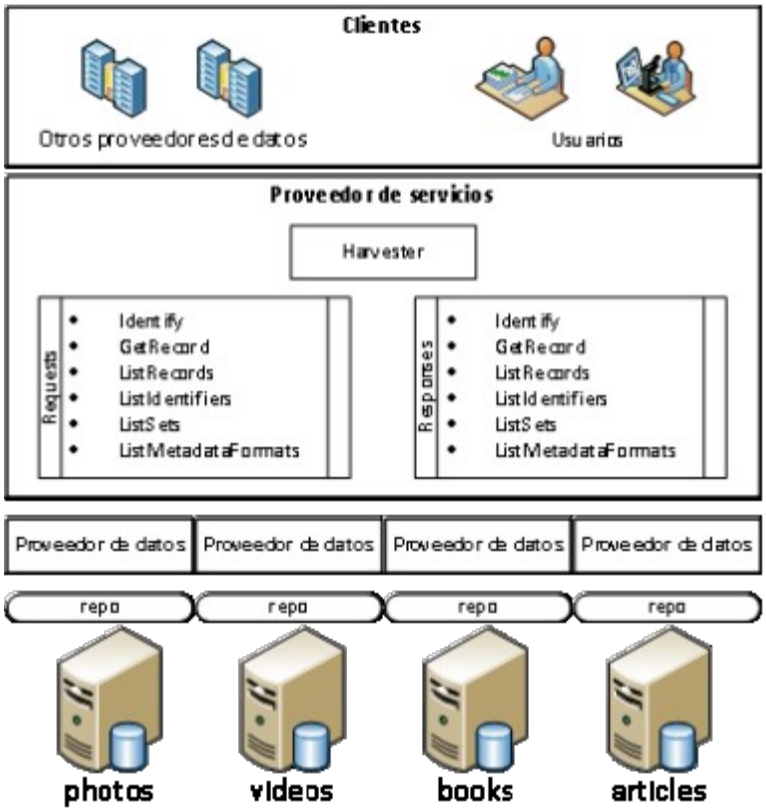

Gráfico 1 – Componentes de la propuesta en que se basa el protocolo OAI-MPH.

Los clientes que invocarán estos servicios no necesitarán utilizar todos los tipos de peticiones durante sus ejecuciones. Como ejemplo, vemos a continuación el uso de la petición *ListRecords* para extraer información de los metadatos desde el año 2011:

```
http://archive.org?verb=ListRecords&from=2011
```

Las respuestas son codificadas a través de *Dublin Core* para alcanzar una interoperabilidad básica. Por otra parte, los proveedores de datos pueden definir una jerarquía para soportar niveles de granularidad en la recolección propia de los clientes que acceden a los *Service Providers*. También se incorporan estampillas de tiempo para marcar los últimos cambios en el conjunto de metadatos y así mejorar el soporte a los clientes para realizar sus acciones de búsqueda y recolección. Y, por último, se brinda información sobre errores, los cuales están también basados en *HTTP*.

Como se aclaró anteriormente, la información derivada de la recolección de metadatos no implica su gratuidad. Es por ello que la *OAI* no define ni prohíbe ningún esquema para la gestión de derechos sobre los datos, quedando el tratamiento de las cuestiones asociadas a la propiedad intelectual de los mismos en el ámbito de los propios proveedores de datos.

En la *OAI* se ha definido una solución minimalista para la interoperabilidad de los diversos "archivos" (o proveedores de metadatos). Esta resulta necesaria si se quiere lograr una adopción global, de ahí que se deriva el concepto de recolección o cosecha de metadatos (*metadata harvesting*) para brindar a los proveedores de metadatos la posibilidad de exponerlos a través de una interfaz común de manera de lograr que se desarrollen nuevos servicios de valor agregado a través del acceso a la información provista. Cabe resaltar que el protocolo *OAI-MPH* se diseñó rechazando la búsqueda distribuida de información. Por el contrario, el concepto de este *framework* se basa en contar con diferentes servidores que proporcionan metadatos aisladamente, bajo criterios de alcance tan simples como, por ejemplo, el filtrado de registros añadidos o modificados en un período específico.

3. *Linked Data* – Descripción

3.1 Reseña histórica

La evolución natural de una web de documentos, basada en archivos *HTML* con enlaces entre ellos (Gráfico 2), es lo que se denomina hoy en día *Linked Data*, es decir, una web donde se publica información estructurada que se entrelaza y sobre la cual se pueden realizar consultas semánticas. *Linked Data* se basa en el uso de un conjunto de tecnologías de la web que son consideradas estándares hoy en día como son el protocolo *HTTP*, *URIs* y *RDF*. Desde que en 2006 Tim Berners-Lee empezó a esbozar ideas que permitieran desarrollar la Web Semántica, en la que ya no se enlazan simples documentos independientes sino que se enlazan cosas/recursos, la intención ha sido alcanzar la reutilización de estos diferentes recursos, contando con una forma no ambigua de representar datos sobre ellos (Gráfico 3). Además, se pretende que las técnicas desarrolladas permitan reducir la redundancia y maximizar la interconectividad entre diferentes conjuntos de metadatos en la web actual y futura.

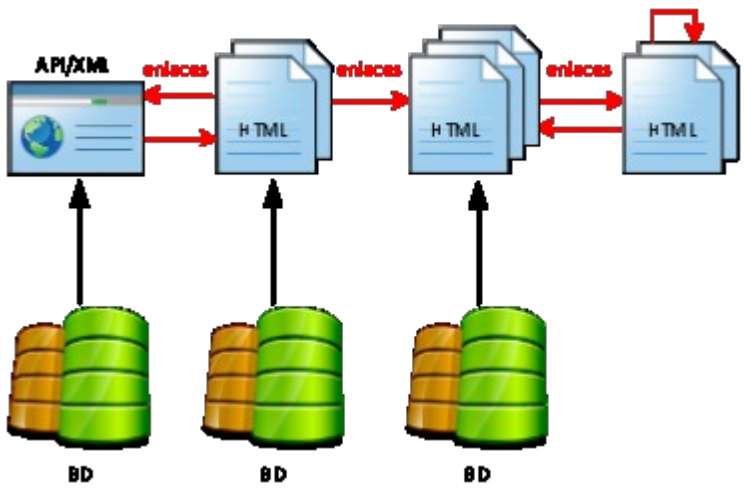

Gráfico 2 – Estructura primitiva de la WWW que consta básicamente de documentos simples enlazados.

Con estos fines, Berners-Lee formuló los cuatro principios que describen la naturaleza de los datos enlazados: emplear UiRIs para nombrar las cosas/recursos; utilizar enlaces *HTTP* para que los usuarios puedan localizar esas cosas/recursos; proporcionar información útil cuando el *URI* es consultado (a través de representaciones *RDF*) y, por último, incluir enlaces a otros

URIs con los datos que se visualizan para descubrir más información sobre otras cosas relacionadas [CITATION Tim06 \l 11274].

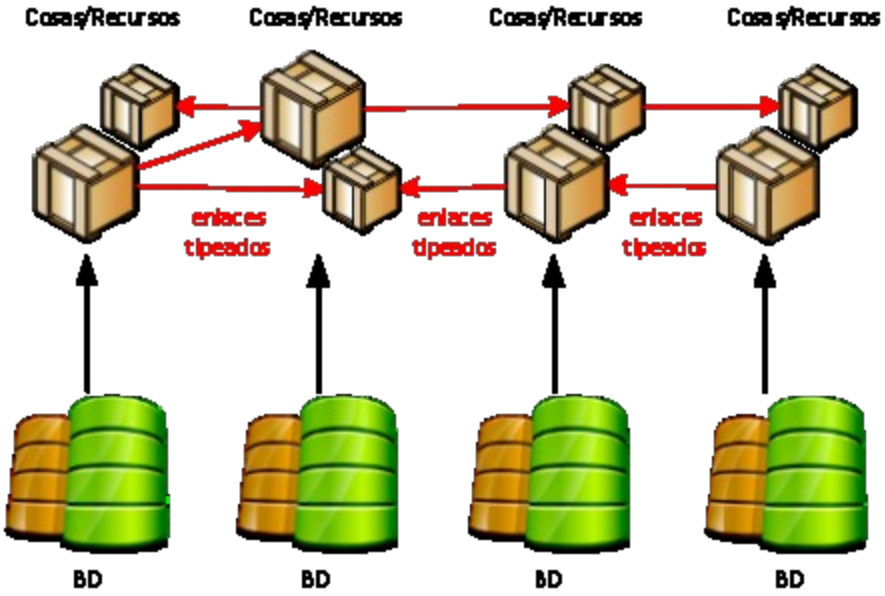

Gráfico 3 – La web compueta por Linked Data.

3.2 Descripción técnica

Para poder construir conjuntos de datos a partir de la utilización de *Linked Data*, el formato de representación de los datos empleado es *RDF*. Se trata de una arquitectura de metadatos cuyo objetivo es proveer un medio para describir y organizar conjuntos de datos a través de diferentes aplicaciones que pueden ser incompatibles entre sí. Para ello se trabaja con el etiquetado de la información. Así, una determinada página web puede ser etiquetada según el tipo de texto que mejor la describa, por ejemplo, como un `<artículo>`, y el nombre de quien lo escribió puede ser descripto mediante la etiqueta `<autor>`. De esta manera, resulta posible establecer una relación entre el autor y el artículo que permita a otros usuarios realizar búsquedas a través de varios documentos, ya no sobre páginas web como entidades individuales, a través de las diferentes relaciones establecidas.

El diseño original de *RDF* está basado en proyectos previos similares como MCF (*Meta Content Framework*) y *PICS* (*Platform for Internet Content Selection*). En *PICS* se permitía el etiquetado de datos para ser asociados con contenidos en Internet. Los objetivos originales de esta especificación eran colaborar con padres y maestros para controlar a qué contenidos de Internet podían acceder los alumnos, pero también podía servir para firmas de código y privacidad. Por ejemplo, algunos navegadores podían utilizar información en formato PICS para determinar sitios no permitidos web de acuerdo a su contenido.

RDF codifica los datos en la forma de "triples", es decir en tres componentes: sujeto, predicado y objeto. Tanto el sujeto como el objeto se representan como *URIs*, los cuales identifican un recurso cada uno, o un *URI* y una cadena de texto respectivamente. El predicado, que también se representa con un *URI*, especifica cómo se relacionan el sujeto y el objeto, que pueden pertenecer a *datasets* diferentes. En la Tabla 1 podemos ver un ejemplo, donde el objeto/sujeto es la fotografía correspondiente de una hoja de un árbol, el cual está representado por el *URI* "http://bioimages.vanderbilt.edu/kirchoff/ac1490", y se establecen

tres tripletas con otros tantos objetos (*URIs* y cadenas de texto o literales) a través de tres relaciones (predicados) diferentes: fotógrafo, fecha de toma y derechos de autor.

Tabla 1 – Datos sobre una fotografía de una hoja de un árbol.

Predicate	Object
http://xmlns.com/foaf/0.1/maker	http://bioimages.vanderbilt.edu/contact/kirchoff#coblea
http://purl.org/dc/terms/created	"2010-09-01T03:41:31"
http://purl.org/dc/elements/1.1/rights	"(c) 2011 Bruce K. Kirchoff"

Asimismo, diversos servicios de búsqueda como Google y Yahoo utilizan también datos en formato *RDF*, reconociendo a partir de determinados vocabularios la información y estableciendo calificaciones sobre los mismos para lograr mejoras en los procesos de filtrado de datos.

Al utilizar el modelo de datos *RDF* en un escenario de *Linked Data* obtenemos una serie de importantes beneficios que Bizer y Heath sintetizan como:

1) El uso de *URIs HTTP* como identificadores globales únicos para datos como así también para vocablos, el modelo *RDF* está diseñado para ser utilizado a gran escala y permite que cualquiera refiera cualquier cosa.
2) Los clientes pueden localizar cualquier *URI* en un grafo *RDF* a través de la web para recuperar información adicional. Así, cada triple *RDF* forma parte de la web y puede emplearse como punto de partida para explorar este espacio de datos.
3) El modelo de dato subyacente permite establecer enlaces *RDF* entre los datos de diferentes fuentes.
4) La información de distintas fuentes puede combinarse con facilidad mezclando los dos conjuntos de triples en un único grafo.
5) *RDF* permite representar información que se expresa con diferentes esquemas en un único grafo, lo que significa que pueden combinarse términos de diferentes vocabularios para representar los datos.
6) En combinación con lenguajes de esquemas como *RDF-Schema* y *OWL*, el modelo de datos permite emplear el nivel de estructura deseado, lo que significa que pueden representarse datos tanto muy estructurados como semi-estructurados [CITATION Hea11 \p 16-17 \t \l 1033].

Por su parte, los mismos autores enumeran a continuación una serie de características desaconsejadas dentro del contexto de *Linked Data* al emplear *RDF* y sugieren para cada una de ellas procedimientos alternativos:

1) En *RDF* la cosificación debería evitarse, puesto que las declaraciones cosificadas son difícilmente consultables mediante *SPARQL*. Entonces, se deben adjuntar los metadatos al documento que contiene las tripletas relevantes.
2) Las colecciones y los contenedores en *RDF* son problemáticos si desean ser consultados con *SPARQL*. O sea, se recomienda siempre utilizar triples múltiples con el mismo predicado.

3) El alcance de los nodos en blanco está limitado al documento en el que aparecen, lo que significa que es imposible crear enlaces a ellos desde documentos externos, reduciendo la posibilidad de entrelazar diferentes fuentes de datos. Todos los recursos en un *dataset*[5] deberían denominarse siempre utilizando URIs [CITATION Hea11 \p 17 \t \l 1033].

A partir de diferentes discusiones en el ámbito del *W3C*, se han producido documentos que especifican las mejores prácticas con respecto a la producción de *Linked Data*. Por ejemplo, en el documento *Linked Data Cookbook – A brief history on open Government Linked Data*[6] preparado por este consorcio en 2011, se detallan y ordenan las siguientes recomendaciones:

1) Modelar los datos.
2) Nombrar las cosas con *URIs*.
3) Reutilizar los vocabularios disponibles.
4) Publicar descripciones que puedan ser leídas por humanos y por computadoras.
5) Convertir los datos disponibles a formato *RDF.*
6) Especificar la licencia adecuada en cada caso.
7) Proveer repositorios de *Linked Data* públicos y anunciarlos.

En otro documento del año 2014, *Best Practices for Publishing Linked Data*[7], evidenciando el desarrollo constante en el área, se precisa un nuevo conjunto de recomendaciones complementarias:

1) Instruir a las partes interesadas sobre el proceso de preparación y mantenimiento de *Linked Data*.
2) Seleccionar un *dataset*.
3) Modelar los datos.
4) Especificar una licencia adecuada.
5) Determinar *URIs* adecuados.
6) Utilizar vocabularios estándar.
7) Convertir los datos a una representación *Linked Data*.
8) Proveer acceso a los datos mediante mecanismos estándares utilizados en la web (procesos independientes o buscadores).
9) Publicar y anunciar los nuevos conjuntos de datos.
10) Reconocer el contrato social en cuanto a la responsabilidad de mantener los datos y hacerlos accesibles a través del tiempo.

Como se puede observar, se han especificado recomendaciones sobre el manejo y creación de repositorios basados en *Linked Data* que se seguirán desarrollando y precisando ya que no se trata de una cuestión cerrada y todavía estamos viendo su evolución.

Una manera de apreciar el crecimiento que está teniendo el uso de la infraestructura basada en *Linked Data*, es visualizar los diferentes grafos correspondientes a la evolución de los

5 Algunos ejemplos de *datasets* son los correspondientes a las iniciativas: dbpedia, geonames, dbtune, DBLP bibliografía, etc. Cada de uno de los cuales posee millones de tripletas.

6 http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook#A_Brief_History_on_Open_Government_Linked _Data [Consultado el 11/9/2015].

7 http://www.w3.org/TR/ld-bp/

datasets que conforman la iniciativa *Linking Open Data*[8]. Desde el año 2007 se han confeccionado gráficos, que permiten observar un crecimiento constante. En los sucesivos gráficos el tamaño de los conjuntos muestra el número de triples de cada *dataset*, según la información provista por sus responsables, en ocasiones, valores estimados. La existencia de flechas, por su parte, representa la existencia de al menos 50 enlaces entre dos conjuntos mientras que cada enlace significa que existe un triple *RDF*. Por último, la dirección de las flechas indica el conjunto de datos que contiene a estos enlaces.

Así, el gráfico 4 corresponde al año 2007 y permite ver los primeros doce *datasets* disponibles y su interrelación. En años posteriores, se aprecia un incremento cada vez mayor hasta llegar a los 570 de 2014 en el gráfico 5.

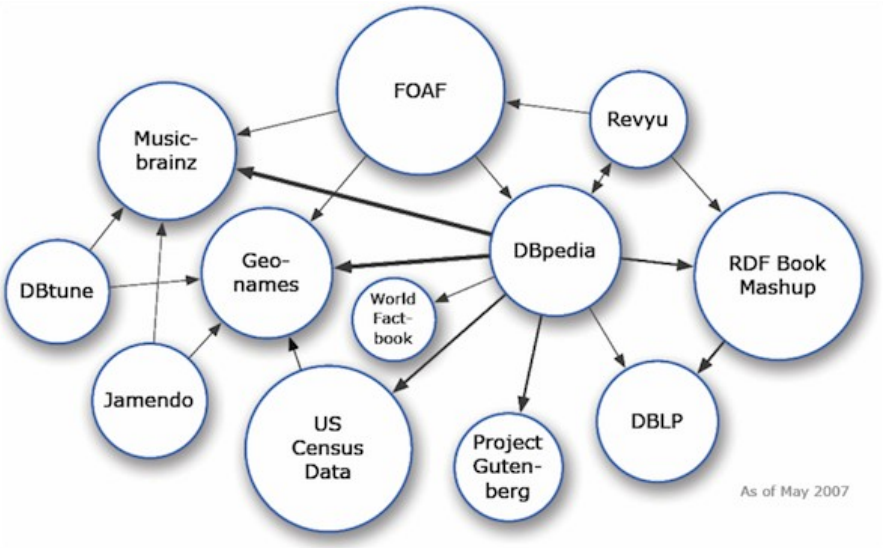

Gráfico 4 – 1/5/2007

8 http://lod-cloud.net/. Los autores no informan datos para los años 2012 y 2013.

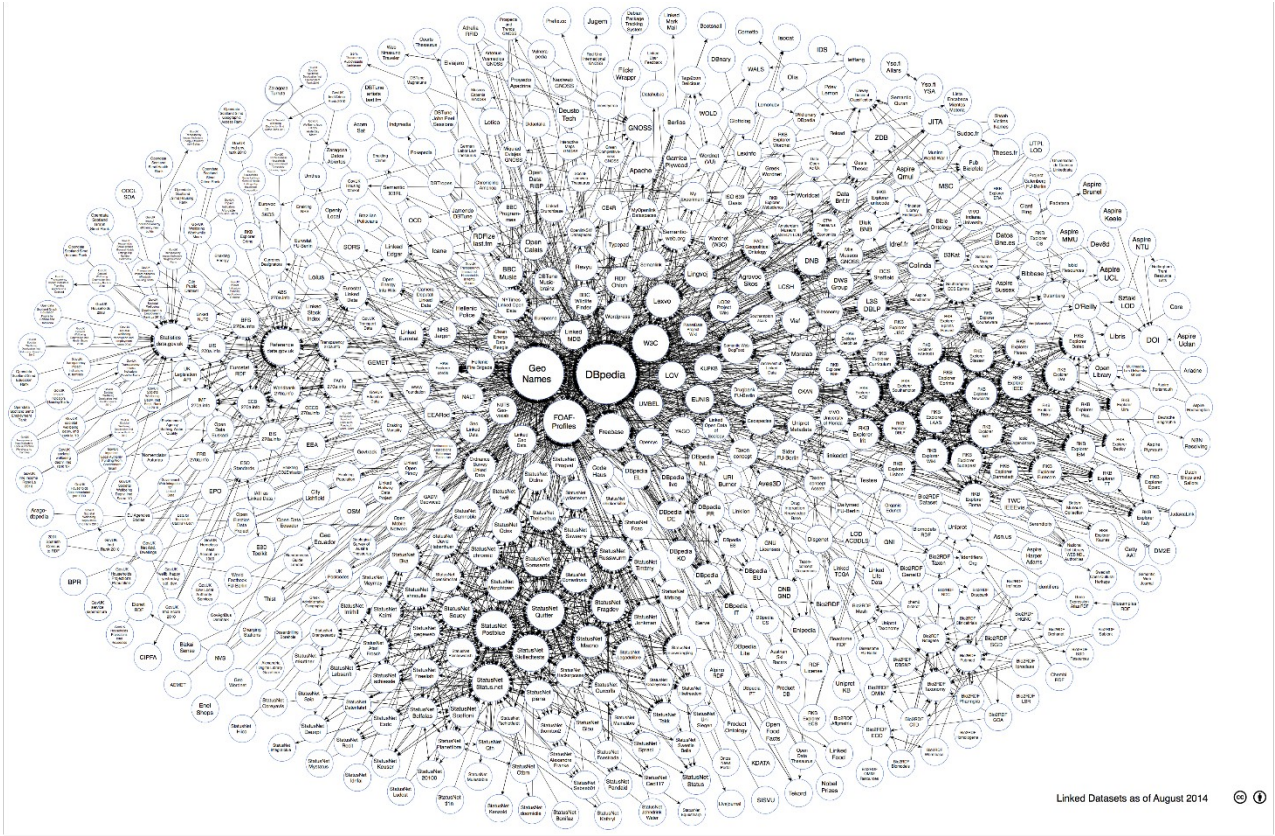

Gráfico 5 – 30/8/2014

En la siguiente tabla, podemos observar la evolución de la cantidad *datasets* en diferentes fechas de generación de los grafos de estudio:

Fecha	Cantidad de datasets
30/08/2014	570
19/09/2011	295
22/09/2010	203
14/07/2009	95
27/03/2009	93
05/03/2009	89
18/09/2008	45
31/03/2008	34
28/02/2008	32
10/11/2007	28
07/11/2007	28
08/10/2007	25
01/05/2007	12

3.3 Aplicaciones y ejemplos de uso

El constante crecimiento de las iniciativas que apelan a *Linked Data* para publicar sus datos en línea ha determinado un desarrollo paralelo de aplicaciones, campo en el que se continúa

trabajando. Bizer, Heath y Berners-Lee (2009) proponen una forma de sistematizar las herramientas, y mencionan algunos de los ejemplos más representativos dentro de cada clase, mucho de los cuales ya no se encuentran operativos.

Las tres clases que se definen son: navegadores de *Linked Data*; buscadores de *Linked Data* y, por último, las aplicaciones para fines específicos. El objetivo de los navegadores es permitir a los usuarios explorar conjuntos de datos siguiendo enlaces que se expresan como triples *RDF* en lugar de enlaces *HTML*, pero dándoles también la posibilidad no solo de explorar los distintos ítems en forma enlazada sino también de analizar grupos de metadatos, indicando –de distintas maneras (texto con enlaces o grafos) – incluso el origen de los datos, es decir, de qué *dataset* provienen. Con independencia de si logran este objetivo general o no, la revisión de los ejemplos proporcionados por los autores muestra que, en este caso, la mayoría de estas implementaciones estaban basadas en la web (*Disco, Tabulator, Marbles, FOAFNaut*), sin constituir *software* independiente (la excepción era *Fenfire*). La búsqueda de alternativas operativas en la actualidad sugiere que no existen todavía navegadores operativos con un alcance lo suficientemente abarcador, es decir, que logren relacionar de manera completamente efectiva metadatos obtenidos de diferentes *datasets*.

Los buscadores, por su parte, pueden clasificarse según se orienten hacia los usuarios (*Falcons* y *SWSE*) o hacia otras aplicaciones que los requieran (*Sindice, Swoogle y Watson*).

Además de estas dos clases de aplicaciones, que cumplen funciones de tipo genérico, un área en la que hay un gran número de desarrollos y de investigaciones en marcha se vincula con las aplicaciones para fines específicos. Algunas de las reseñadas por los autores se relacionan con datos que hacen a la industria del cine; con aplicaciones para organizar los repositorios de grandes conglomerados de medios de comunicación; con la gestión de la información académica (organismos de ciencia y universidades), entre otras[CITATION Biz09 \t \l 1033].

4. Comparaciones

Una vez implementado y puesto en marcha un repositorio *OAI-PMH* los metadatos se transfieren desde el servidor al cliente a través de la web, mediante la utilización de parámetros enviados vía *HTTP* a través de métodos *GET* o *POST*. De esta manera diferentes clientes pueden acceder y obtener información para ser utilizada o bien indexada para procesamiento posterior. Acá se presenta la primera diferencia sustancial con lo propuesto a través de *Linked Data*, ya que esta última tecnología se basa en la idea de exponer los metadatos sobre la Web propiamente dicha, de manera tal que estos puedan ser accedidos por personas o bien por programas informáticos. Por lo tanto, queda claro que en el protocolo *OAI-PMH* se utiliza la estructura propia de la web para el transporte de los datos, mientras que en *Linked Data* datos son parte de la web en sí misma [CITATION Has09 \l 11274]. Por otra parte, brinda una ventaja significativa, en el sentido que todos los metadatos representados vía *URIs* sirven para identificar recursos propiamente dichos. Para entender mejor el concepto en la iniciativa *OAI-MPH* se utilizan los *URIs* para identificar ítems, no así en *Linked Data*, donde estos adquieren el rol de identidades que se pueden relacionar y seguir para obtener información no ambigua sobre ellas, es decir obtener una representación de su descripción [CITATION Der07 \l 1033].

Con respecto a la posibilidad de consultas de los metadatos, el hecho de que se pueden "cosechar" datos de los repositorios en *OAI-PMH* es una de las cualidades de este protocolo, que se ve superado al utilizar *Linked Data*, ya que se pueden realizar consultas a través de un

lenguaje estructurado, como *SPARQL*, que naturalmente puede ser visto como un protocolo de consulta específico basado en los metadatos accesibles sobre la web en sí misma.

El contar con más y más datos disponibles en la web a través de *Linked Data* permite que la cantidad de aplicaciones a desarrollar sea mayor con el transcurso del tiempo, así como su tipo y calidad. Algunas de las aplicaciones más comunes que ya existen (en diferentes estados de avance de implementación) son aquellas que posibilitan la búsqueda y la recuperación, como los siguientes motores de búsqueda: *Swoogle, Watson, Falcons Explorer, Falcons, LOD Cloud Cache* y *sameAs*.

En Haslhofer & Schandl [CITATION Has09 \n \t \l 1033] encontramos una catalogación de las diferencias conceptuales entre la tecnología basada en repositorios propuesta por la *OAI* y la idea de universalidad detrás de *Linked Data*. A modo de síntesis, confeccionamos la siguiente tabla donde se consignan las diferencias encontradas:

OAI-PMH	Linked Data
URIs utilizados para identificación.	*URIs* como identidades con información enlazada.
Se definen verbos en el protocolo y un conjunto de parámetros. Estos deben ser conocidos por los clientes para recuperar metadatos de un repositorio.	Se apoya en las funcionalidades suministradas por las tecnologías existentes en la web, protocolo *HTTP* y métodos asociados, *URI*, *HTML* o *RDF* respectivamente.
Entrega datos en formato *XML* (siendo el único válido).	Depende de los mecanismos de negociación del protocolo *HTTP* para entregar datos en diferentes representaciones.
Provee la posibilidad de recuperación de metadatos en lotes, sobre una sola transacción *HTTP.*	Indirectamente se provee mediante el uso del *SPARQL* como lenguaje de consulta y protocolo. Permitiendo además la recuperación de metadatos en base a criterios de búsqueda complejos.
Puede devolver metadatos en diferentes formatos.	No se puede solicitar un formato específico. Solo es posible describir un recurso con diferentes vocabularios y consultar obteniendo metadatos en solo formato.
Provee una manera básica de control de versiones. Permite a los clientes obtener solamente metadatos creados o modificados en un determinado rango de tiempo.	Sería posible en *OAI-PMH* agregar términos específicos dentro del vocabulario y utilizar *SPARQL* para consultar un rango de fechas para obtener metadatos dentro del mismo. Otra alternativa sería llevar un control de actualizaciones mediante una bitácora.

5. Conclusiones

Existen propuestas con respecto a la posibilidad de contar con mecanismos automáticos que permitan exponer mediante *Linked Data* los metadatos de repositorios *OAI-PMH*, como así también la funcionalidad de enlazar metadatos de diferentes orígenes [CITATION Has09 \l 1033], pero estableciendo ciertos criterios de calidad que deben cumplir los ítems almacenados en los repositorios para poder ser expuestos en la web. El potencial nivel al que se pueden enlazar los datos en el futuro, sin lugar a dudas contribuirá a obtener mejores resultados en diferentes procesos de búsqueda de información asociada a la labor de investigadores de todas las disciplinas científicas a partir del análisis de datos estadísticos

como así también a optimizar el funcionamiento de organismos diversos y de áreas de gobierno a partir de la organización significativa de sus datos operativos. Estos dos ejemplos solamente, dan cuenta de cómo estas tecnologías pueden contribuir a una mejora en la calidad de vida de la sociedad moderna.

Actualmente SPARQL es solamente uno de los varios formalismos propuestos para consultar la web basada en *Linked Data*, hay otras propuestas como *NautiLOD, SPARQLeR* y *Linked Data QL*, entre otros.

El campo de trabajo es tan rico que además del desarrollo de las aplicaciones necesarias para el aprovechamiento de las posibilidades que ofrecen los diferentes *datasets* y la Web Semántica, amerita en el futuro también programas de investigación sobre: manejo de repositorios distribuidos para utilizar con el protocolo *OAI-PMH*; nuevos mecanismos de migración de datos de este protocolo a *Linked Data* y, por último, sobre las posibilidades de interrelacionar de forma automática las técnicas de búsqueda de plataformas como Yahoo o Google (técnicas basadas en el empleo de algoritmos de análisis de enlaces como *pagerank*) con las que *Linked Data* permite utilizar a partir de *SPARQL*.

6. Referencias bibliográficas

Berners-Lee, T. (2006, 07 27). *Linked Data.* Retrieved from W3C: http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., & Heath, T. (2011). Linked Data. Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology, 1*(1), 1-136. doi:10.2200/S00334ED1V01Y201102WBE001

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1-22. doi:10.4018/jswis.2009081901

Cruz, J., & Coll, I. (2003). Coll, I. S., & Cruz, J. M. B. (2003). Open archives initiative. Protocol for metadata harvesting (OAI-PMH): descripción, funciones y aplicaciones de un protocolo. *El profesional de la información, 12*(2), 99-106.

Dereferencing HTTP URIs . (2007, 5 31). Retrieved from W3C site: http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14.html

Fionda, V., Gutierrez, C., & Pirro, G. (2015). N auti LOD - A Formal Language for the Web of Data Graph. *ACM Transactions on the Web, 9*(1), 5.

Haslhofer, B., & Schandl, B. (2010). Interweaving OAI-MPH data sources with the linked data cloud. *International Journal Metadata, Semantics and Ontologies 5(1)*, 17-31.

Hyvönen, E. (2012). Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology, 2*(1), 1-159. doi:10.2200/S00452ED1V01Y201210WBE003

Jiewen, H., Abadi, D., & Ren, K. (2011). Scalable SPARQL querying of large RDF graphs. *In proceedings of the VLDB Endowment*, *4(11)*, pp. 1123-1134.

Kochut, K., & Janik, M. (2007). SPARQLeR: Extended SPARQL for semantic association discovery. *In proceedings of the Semantic Web: Research and applications* (pp. 145-159). Springer Berlin Heidelberg.

Pérez, J., & Hartig, O. (2015). LDQL: A Query Language for the Web of Linked Data. *In proceedings of 14th Int. Semantic Web conference.*

W3C. (2004/2014, 02 10). *RDF Primer. W3C Recommendation and Working Group Note.* Retrieved from W3C: http://www.w3.org/TR/rdf-primer/

W3C. (2008, 01 15). *SPARQL Query Language for RDF.* Retrieved from W3C: http://www.w3.org/TR/rdf-sparql-query/

W3C. (2011, 03 21). *SPARQL 1.1 Overview.* Retrieved from W3C: http://www.w3.org/TR/sparql11-overview