Análisis de herramientas para la búsqueda semántica de documentos

Leonardo de-Matteis – Karina Cenci – Pablo Fillottrani

Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
kmc,ldm,prf@cs.uns.edu.ar

Octubre 2015

BIBLIOTECA NACIONAL

V Encuentro Nacional de Catalogadores

Estructura

Escenario actual

Enormes cantidades de datos se encuentran distribuidas en diferentes servidores y servicios sobre Internet.

Escenario actual

Enormes cantidades de datos se encuentran distribuidas en diferentes servidores y servicios sobre Internet.

La información está dirigida al usuario final a través de páginas en la web pero no está disponible de forma estructurada y manipulable por procesos computacionales.

Web semántica

La **web semántica** surge para promover:

- actividades,
- procesos,
- estándares,
- y tecnologías.

Desafios de la web semántica

Mediante las herramientas que provee la Web Semántica se plantean nuevos desafíos para:

- crear servicios,
- desarrollar sistemas
- y generar valor agregado

con la información obtenible de diferentes fuentes de datos en Internet.

OAI - Protocol for Metadata Harvesting

Componentes básicos:

- *Data providers*: repositorios que permiten acceder a sus metadatos a través del protocolo.
- *Service providers*: realizan solicitudes para extraer información de los metadatos.
- Utiliza peticiones básicas provistas por el protocolo HTTP: GET o POST.

Peticiones

En los *Service providers* se definen e implementan las siguientes solicitudes:

- GetRecord
- Identify
- ListRecords
- ListIdentifiers
- ListSets
- ListMetadataFormats

Componentes

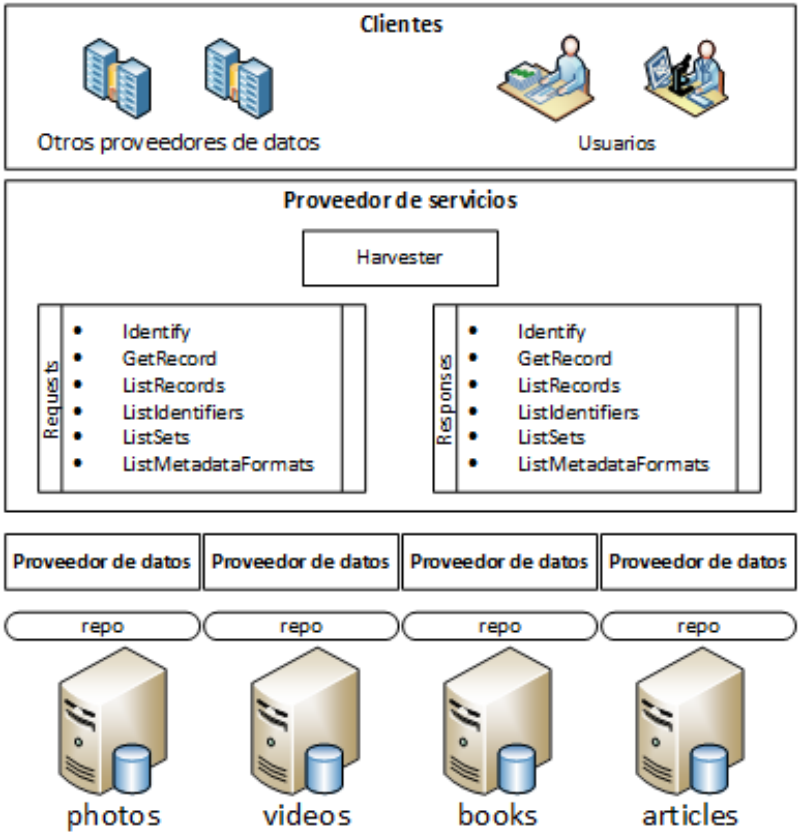

Figura: Gráfico estructural de un repositorio OAI-PMH.

WWW clásica y actual

La World Wide Web 'primitiva' consta básicamente de documentos simples enlazados.

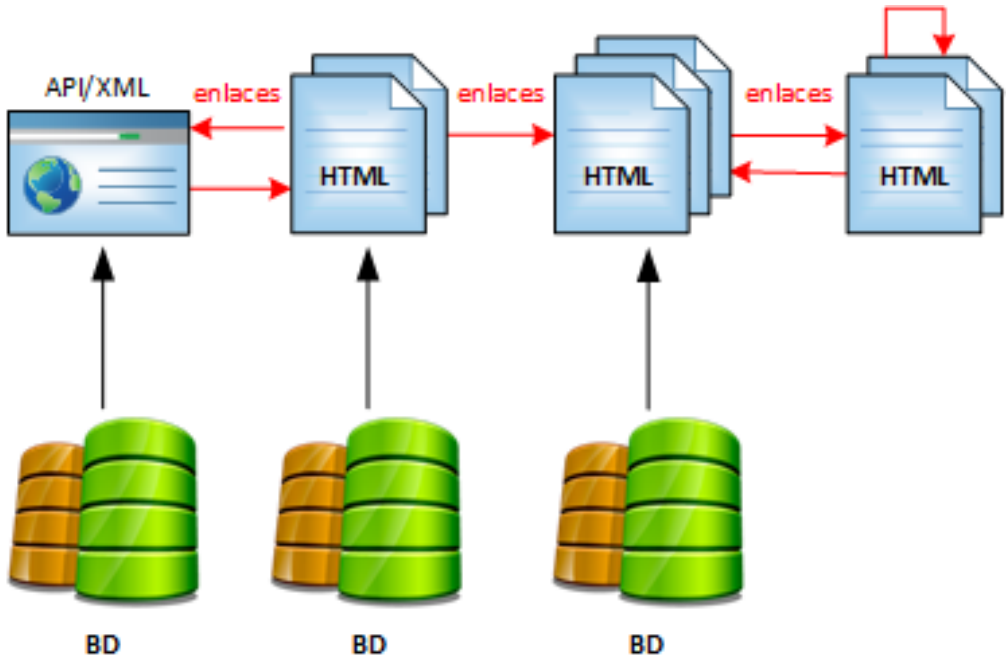

Figura: Web actual.

Web de documentos simples enlazados

Características:

- similar a un sistema de archivos gigante;
- diseñada para uso de seres humanos;
- basada en documentos y enlaces entre ellos (o partes de ellos);
- semántica implícita entre contenido y enlaces.

Web de documentos simples enlazados (cont.)

Desventajas:

- Simplicidad: enlaces sin tipos, baja estructuración y no relación entre datos.
- Integración: no brinda la posibilidad de realizar búsqueda globales.
- Consultas: imposibilidad de realizar búsquedas relacionadas con filtros.

Web de datos enlazados

La WWW donde se enlazan cosas/recursos:

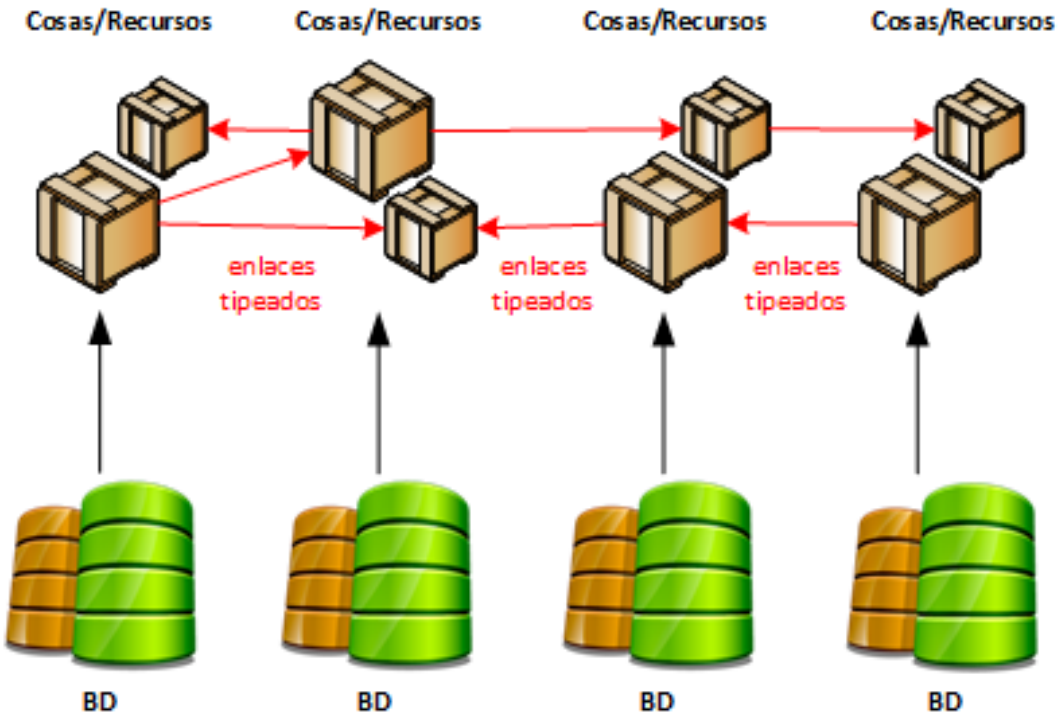

Figura: Web basada en la utilización de Linked Data.

Web de datos enlazados (cont.)

Características:

- diseñada para uso de máquinas y luego de humanos;
- consta de recursos (o descripciones de los mismos);
- enlaces entre recursos;
- semántica explícita entre recursos y enlaces.

Web de datos enlazados (cont.)

Ventajas:

- Se basa en la reutilización.
- Disminuye la redundancia.
- Estimula el crecimiento de la interconexión (enlaces entre recursos).
- Trata de facilitar la búsqueda y utilización de los metadatos.

Contraste

Acceso y filtrado de los metadatos

OAI-MPH: Los metadatos se pueden 'cosechar' de los repositorios.

Linked Data: Se realizan consultas a través del lenguaje estructurado SPARQL (visto también como protocolo de consulta basado en los metadatos accesibles sobre la web).

Diferencias relevantes

URIs

OAI-MPH: Utilizados para identificación.

Linked Data: Se utilizan como identidades con información enlazada.

Diferencias relevantes (cont.)

Protocolo

OAI-MPH: Se definen verbos en el protocolo y un conjunto de parámetros. Estos deben ser conocidos por los clientes para recuperar metadatos de un repositorio.

Linked Data: Se apoya en las funcionalidades suministradas por las tecnologías existentes en la web, protocolo HTTP y métodos asociados, URI, HTML o RDF respectivamente.

Diferencias relevantes (cont.)

Entrega de datos

OAI-MPH: Entrega datos en formato XML, siendo el único válido con respecto a la fase de negociación y entrega del contenido.

Linked Data: Depende de los mecanismos de negociación del protocolo HTTP para entregar datos en diferentes representaciones posibles.

Diferencias relevantes (cont.)

Recuperación de datos en lotes

OAI-MPH: Provee la posibilidad de recuperación de metadatos en lotes, sobre una sola transacción HTTP.

Linked Data: Indirectamente se provee mediante el uso de SPARQL como lenguaje de consulta y protocolo, permitiendo además la recuperación de metadatos en base a criterios de búsqueda complejos.

Diferencias relevantes (cont.)

Formato de los datos obtenidos

OAI-MPH: Puede devolver registros con metadatos en diferentes formatos.

Linked Data: No se puede solicitar un formato específico. Es posible describir un recurso con diferentes vocabularios y consultar obteniendo metadatos en algún formato específico (utilizando SPARQL).

Diferencias relevantes (cont.)

Filtrado de datos en consultas

OAI-MPH: Provee una manera básica de control de versiones. Permite a los clientes obtener solamente metadatos creados o modificados en un determinado rango de tiempo.

Linked Data: 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.

Situación actual: OAI-PMH

La propuesta de la OAI:

- Logra la difusión eficiente de contenido en Internet sobre la WWW.
- Permite un incremento cuantitativo de la disponibilidad global de publicaciones científicas.
- Permite el desarrollo de aplicaciones por fuera de la comunidad OAI.
- Permite incorporar en los repositorios metadatos para cualquier material que pueda ser almacenado electrónicamente.
- Adopción amplia del protocolo para intercambiar datos bibliográficos/digitales.

Situación actual: Linked Data

La cantidad de datos publicados a través Linked Data crece en forma acelerada.

Situación actual: Linked Data

La cantidad de datos publicados a través Linked Data crece en forma acelerada.

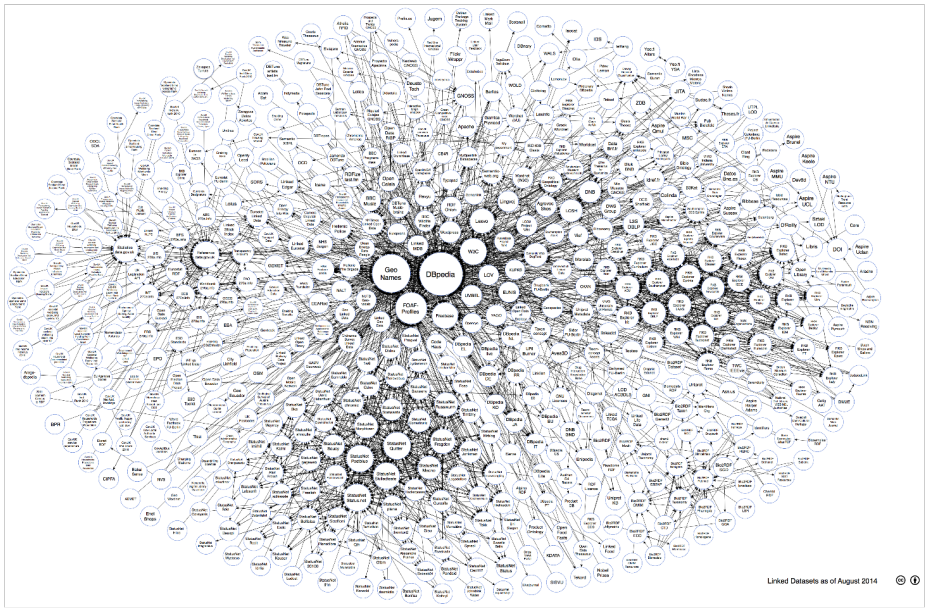

Figura: Grafo correspondiente a datasets de Linked Open Data 2014

Situación actual: Linked Data

La cantidad de datos publicados a través Linked Data crece en forma acelerada.

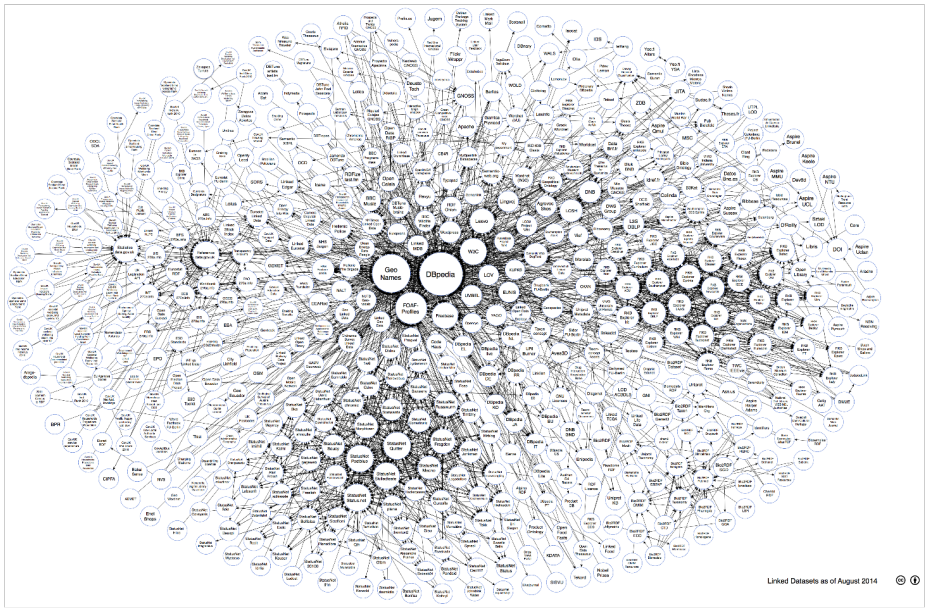

Figura: Grafo correspondiente a datasets de Linked Open Data 2014

Situación actual: Linked Data (cont.)

La tecnología requiere de desarrollos para:

- Lograr una mayor integración para los datos enlazados de diferentes fuentes.
- Alcanzar un descubrimiento dinámico de datos y fuentes de metadatos.
- Crear y brindar ambientes de desarrollo para aplicaciones.
- Generar interfases apropiadas para el usuario final.

Conclusiones

El trabajo permite realizar las siguientes observaciones:

- Existen propuestas para aquellos interesados en permitir que repositorios implementados con OAI-PMH puedan exponer sus metadatos siguiendo los principios de Linked Data.
- Sería necesario evaluar 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.

Conclusiones (cont.)

- Se pueden implementar mecanismos automáticos de acceso (transformación) pero se deben establecer criterios de calidad que deben cumplir los ítems almacenados en los repositorios.
- Debe aumentarse el nivel de relación entre datos/recursos (enlaces), para permitir obtener mejores resultados en procesos de búsqueda de información.
- SPARQL es solamente uno de los varios formalismos propuestos para consultar la web semántica (hay otras propuestas: NautiLOD, SPARQLeR, Linked Data QL, etc).
- Deben desarrollarse aplicaciones para el aprovechamiento de los diferentes conjuntos de metadatos disponibles.

Preguntas

Preguntas ... ¿?

V Encuentro
Nacional de
Catalogadores