



VISUALIZACIÓN Y RECUPERACIÓN DE INFORMACIÓN EN LA WEB SEMÁNTICA



Jorge Morato, Sonia Sánchez-Cuadrado, Alejandro Ruiz-Robles y José-Antonio Moreiro-González



Jorge Morato es doctor en documentación y trabaja en el *Departamento de Informática* de la *Universidad Carlos III de Madrid*. Sus principales publicaciones están relacionadas con la utilización y construcción automática de tesauros y ontologías en la mejora de la recuperación de información y gestión del conocimiento.

<http://orcid.org/0000-0002-7530-9753>

Universidad Carlos III de Madrid, Departamento de Informática
Avda. Universidad, 30. 28911 Leganés (Madrid), España
jmorato@inf.uc3m.es



Sonia Sánchez-Cuadrado es doctora en documentación y ha sido profesora del *Departamento de Informática* de la *Universidad Carlos III* durante más de una década. Actualmente trabaja en sistemas de procesamiento del lenguaje natural en publicidad SEO. Su actividad investigadora está enfocada a la extracción de información, reconocimiento de patrones, procesamiento del lenguaje natural y sistemas de organización del conocimiento.

<http://orcid.org/0000-0002-7722-1982>

JOT Internet Media
Gral. Ramírez de Madrid, 6-8. 28020 Madrid, España
ssanhec@gmail.com



Alejandro Ruiz-Robles es máster en ciencia y tecnología informática por la *Universidad Carlos III de Madrid (UC3M)* y profesor a tiempo completo por la *Universidad de Piura (Perú)*. Actualmente cursa el doctorado en la *UC3M* y su principal área de investigación es la gestión del conocimiento.

<http://orcid.org/0000-0002-5280-679X>

Universidad de Piura, Dep. Ingeniería Industrial y de Sistemas
Av. Ramón Múgica, 131. Apdo. postal 352, Piura, Perú
alejandro.ruiz@udep.pe



José-Antonio Moreiro-González es catedrático de biblioteconomía y Documentación de la *Universidad Carlos III de Madrid*. Es autor de trabajos sobre técnicas de análisis de contenido documental y cuestiones conceptuales de la documentación. Ha sido director del *Departamento de Biblioteconomía y Documentación* y decano de la *Facultad de Humanidades, Comunicación y Documentación*.

<http://orcid.org/0000-0002-8827-158X>

Univ. Carlos III de Madrid, Depto. de Biblioteconomía y Documentación
C/ Madrid, 126. 28903 Getafe (Madrid), España
jamore@bib.uc3m.es

Resumen

Se cuestiona el funcionamiento de los motores de búsqueda para la web semántica y si la visualización gráfica de los resultados mejora la recuperación. Se han analizado los tipos de consultas que se pueden realizar y las principales características de un buscador semántico. Se ha seleccionado una consulta y se ha ejecutado en buscadores semánticos de acceso online con la *DBpedia*. Además, se han estudiado las funciones básicas y deseables en un buscador de la web semántica y se han revisado algunos estudios sobre la evaluación de buscadores semánticos. Por último se analizan las soluciones gráficas para mostrar los resultados de algunos buscadores. Las principales conclusiones inciden en la idea de que las soluciones de visualización de los resultados de las consultas no cumplen las expectativas propuestas en la bibliografía. .

Palabras clave

Datos abiertos, Datos enlazados, Web semántica, Buscadores, Motores de búsqueda, Consultas, Visualización, Representaciones gráficas, Evaluación, Navegación conceptual, *DBpedia*.

Artículo recibido el 31-01-2014
Aceptación definitiva: 06-06-2014

Title: Information visualization and retrieval in the semantic web

Abstract

The authors have questioned how the semantic web search engines work and if the graphical display of the search results improves the retrieval. The types of queries that can be performed, as well as the main features of a semantic search engine, have been analyzed. A query was selected and executed in semantic search engines with online access to *DBpedia*. The basic and desirable functionality in a semantic web search engine, and reviewed studies about the evaluation of semantic search engines have been studied. Finally, there have been analyzed the graphic solutions for displaying the results of some search engines. The main findings underlines the idea that graphical solutions to display query results do not meet the expectations proposed in the bibliography.

Keywords

Linked open data, LOD, Semantic web, Queries, Search engines, Visualization, Graphical display, Evaluation, Conceptual browsing, *DBpedia*.

Morato, Jorge; Sánchez-Cuadrado, Sonia; Ruiz-Robles, Alejandro; Moreiro-González, José-Antonio (2014). "Visualización y recuperación de información en la web semántica". *El profesional de la información*, mayo-junio, v. 23, n. 3, pp. 319-329.

<http://dx.doi.org/10.3145/epi.2014.may.12>

1. Introducción

El proyecto *linked open data* (LOD) se basa en la propuesta de **Berners-Lee** de publicar datos enlazados en la web mediante el uso de URIs, que permiten nombrar las entidades y establecer enlaces que relacionen recursos (**Bizer; Heath; Berners-Lee**, 2009). Esta información se formaliza con estándares como RDF, con sentencias en forma de tripletas con la secuencia recurso-atributo-valor. Posteriormente estas tripletas pueden ser interrogadas con lenguajes como *Sparql*, donde la potencia en la recuperación se produce por la combinación de tripletas de diferentes recursos.

En 2011 el proyecto *LOD* ya contaba con 295 *datasets*, con más de 31 millones de tripletas (**Bauer; Kaltenböck**, 2011) y diferentes recursos interrelacionados, caracterizándose por las enormes dimensiones de sus componentes. Por ejemplo *FreeBase* está formado por 43 millones de tópicos y 2,3 millones de hechos. *DBpedia* contiene 3,7 millones de tópicos, incluyendo 832.000 personas o 639.000 lugares.

<http://dbpedia.org/About>

<http://www.freebase.com/browse>

El elevado número de conceptos interrelacionados hace patente la necesidad e importancia de un método intuitivo para organizar, recuperar, filtrar, visualizar y navegar entre conceptos y acceder a sus recursos vinculados. En el estudio de **Méndez y Greenberg** (2012) se recogen y sintetizan diversas propuestas desde la perspectiva de *LOD*.

2. Consulta y navegación en la web semántica

Tipos de consultas y respuestas

Las consultas que caracterizan la web semántica consisten en la formulación de una única respuesta a una pregunta dada o a la obtención de un listado relacionado con una entidad, más que a una localización de palabras claves contenidas en documentos en lenguaje natural. Se interroga sobre un dato simple como una fecha, un nombre, un lugar (ele-

mentos también conocidos como entidades del nombre), o bien sobre listados de entidades con atributos o facetas comunes que permitan filtrar la información. Según la propuesta clásica de **Broader** (2002), los tipos de consultas son:

- Informacionales: pueden ser tanto de instancias de tipos de datos (listado de músicos alemanes, fecha de las olimpiadas de Londres, etc.) como de refinamiento o filtrado de resultados por características, facetas, etc. (músicos alemanes del siglo XX con más de cinco sinfonías).
- Navegacionales: para, dado un recurso encontrar otros similares (webs sobre tecnologías para almacenar energía), o bien, dado un concepto encontrar documentos relacionados (sitios con fotografías de monumentos en París).
- Transaccionales: aquellas que permiten interactuar con el sitio una vez localizado.

“ La interrogación semántica implica conocer a priori dónde preguntar y cómo se representaron los conceptos y sus relaciones ”

Obviamente esta clasificación no está orientada, por ejemplo, a la localización de recursos de conocimiento formalizados (como ontologías, vocabularios de metadatos, etc.). Un ejemplo podría ser localizar un vocabulario de metadatos que incluya conceptos relacionados con un corpus de documentos. Ante estas carencias, investigadores como **Uren et al.** (2007) sugieren otros tipos de consultas como: consulta de entidad, relacional (que busca relaciones entre entidades) y parametrizada, que sería similar a las conocidas en documentación como facetadas. De forma análoga, **Morato et al.** (2013) identifican consultas conceptuales, contextuales y las que permiten el acceso a un recurso. También se han categorizado según las necesidades de información

del usuario (Wei; Barnaghi; Bargiela, 2008; Strassunskas; Tomassen, 2010). Otras tipologías hacen hincapié en la formulación de la consulta: Fazzinga y Lukasiewicz (2010) clasificaron las consultas semánticas en estructuradas, con palabras clave y en lenguaje natural.

Los distintos tipos de consultas pueden devolver tipos de resultados diferentes: una consulta sobre un dato normalmente puede requerir una única tripleta; un listado puede precisar un conjunto de tripletas; y la petición de una o más páginas asociadas a un concepto, puede devolver una página enlazada mediante una o más tripletas. Por ejemplo, para obtener “un listado de municipios españoles y la población correspondiente a cada uno de ellos”, se crea una consulta y los datos resultantes pueden proceder de varias combinaciones de tripletas.

Tipos de buscadores semánticos

Presentan una gran diversidad según sus características. Su interfaz de búsqueda puede presentar formularios, grafos o permitir el uso de un lenguaje de recuperación. Lopez et al. (2011) centraron la clasificación en aspectos como: el ámbito, el grado de estructuración de las fuentes, los tipos de datos para expresar la consulta y el contexto de búsqueda (desambiguación, multilingüismo, ranking o fiabilidad). Además, el buscador puede tener o no una ontología subyacente para navegar semánticamente entre los conceptos.

La satisfacción del usuario ante la complejidad de la web semántica no puede medirse con métricas típicas de recuperación de información

Algunos permiten búsquedas sólo contra un *dataset* o un vocabulario, mientras que otros permiten consultas federadas contra más de un conjunto de datos. También varían en cuanto a la usabilidad y conocimiento técnico requerido por el usuario. Por último cada buscador utiliza un lenguaje de interrogación, y una ontología que puede determinar la potencia y complejidad para realizar la consulta.

Algunas propuestas plantean buscadores similares a Google en cuanto a un formulario sencillo como interfaz de búsqueda. Es el caso de Swoogle, Sindice, Falcons o Dr.Watson, que

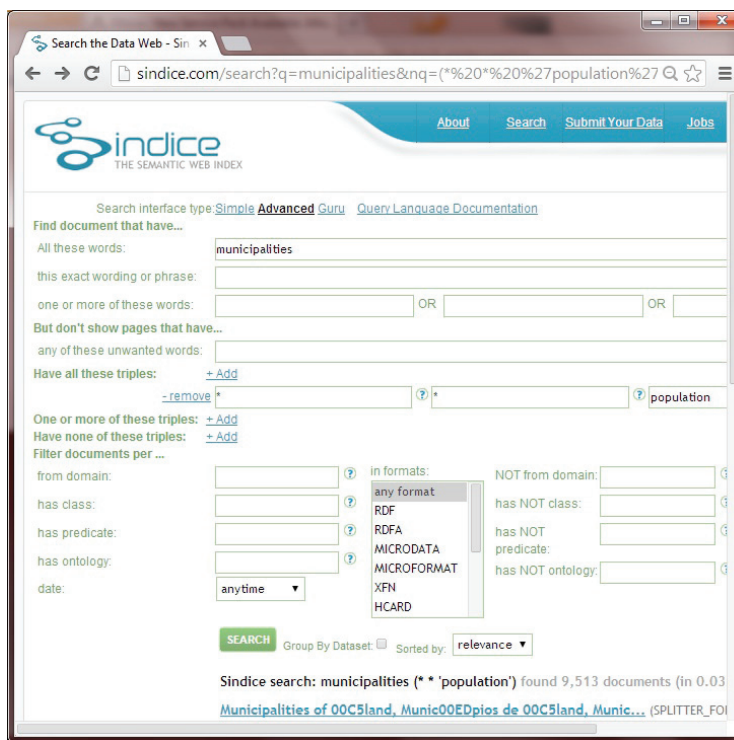


Figura 1. Búsqueda avanzada en Sindice para buscar la población de municipios. <http://sindice.com>

presentan una interfaz despejada que permite una consulta sencilla basada en una palabra clave. Sin embargo, esta aparente usabilidad contrasta con los interrogantes planteados al usuario sobre cómo se ha representado cada concepto, como se observa en la figura 1, Lo que supone un conocimiento especializado previo sobre las consultas y las respuestas que se pueden obtener.

Estos buscadores proporcionan una solución parcial para resolver los tipos de consultas de la web semántica. Por una parte por la dificultad de interrogar al buscador, cuando se desconoce el recurso o si no se tiene un conocimiento de los lenguajes de la web semántica (Miličić, 2011; Morato et al., 2013), y por otra, por la dificultad añadida que supone la interpretación de los resultados, bien por el modo en el que se ordenan, o bien por cómo se muestra el resultado y su contexto.

En cuanto a las opciones de búsqueda y el lenguaje de interrogación para los buscadores semánticos, dependiendo del tipo de buscador y de consulta se exigen algunos conocimientos a los usuarios como:

- Conocer un recurso que contenga esta información.
- Conocer la estructura de conocimiento que defina el con-

Tabla 1. Comparación de las definiciones para el elemento “agent”

Vocabulario	Definición elemento “agent”
NSDL Registry http://metadataregistry.org	The person or organization having a role with the vocabulary
MuDiCat http://www.ifla.org	A person (author, publisher, sculptor, editor, director, composer, etc.) or a group (family, organization, corporation, library, orchestra, country, federation, etc.) or an automaton that has a role in the lifecycle of a resource
FOAF http://www.foaf-project.org	An agent (eg. person, group, software or physical artifact)
VCard http://tools.ietf.org/html/rfc6350	An entity who may sometimes act on behalf of the entity associated with the vCard

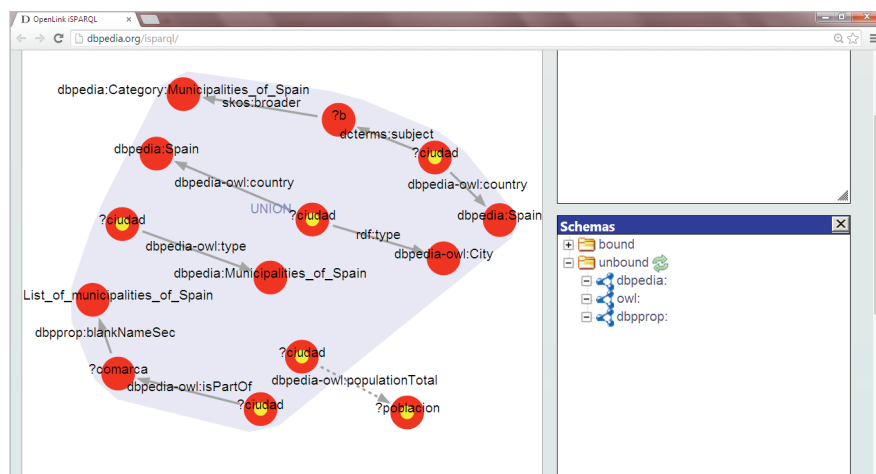


Figura 2. Grafos alternativos sobre población de municipios en España con *iSPARQL*
<http://dbpedia.org/isparql>

cepto general en ese recurso y cómo está representado el dato. Un mismo concepto puede estar representado por criterios diferentes.

- Ser consciente de las diferencias de matices entre las definiciones alternativas referidas a la representación de un concepto. Esto es, una misma codificación puede representar conceptos más genéricos, más amplios o más específicos dependiendo del vocabulario.
- Para consultas complejas, conocer el lenguaje de interrogación del buscador (como *Sparql*).
- Valorar la consulta más eficaz en el caso de que existan alternativas.

En el siguiente punto se detallan ejemplos de los elementos discutidos en esta enumeración.

Representación y codificación de los conceptos

Un mismo concepto puede ser representado en distintos vocabularios o esquemas de metadatos con diferente terminología y codificado en diferentes lenguajes de representación, por eso se habla de representación y codificación de conceptos. Si se desconoce en qué vocabulario se encuentra el concepto, una aproximación razonable para realizar la consulta sería utilizar recursos que abarquen múltiples vocabularios como *Open metadata registry (OMR)* o *Linked open vocabularies (LOV)*. No obstante, nos encontraremos con los problemas asociados a la polisemia y falta de contexto. Por ejemplo en la búsqueda para el concepto “*agent*”, el *OMR* lo encuentra en 4 vocabularios y *LOV* lo localiza en 88. Si extrae-

Tabla 2. Consultas *Sparql* para recuperar la población de municipios:
 prefijos *dbpedia*: <http://dbpedia.org/resource>; *dbpedia-owl*: <http://dbpedia.org/ontology>; *category*: <http://dbpedia.org/resource/Category>;
dbpprop: <http://dbpedia.org/property>

1	SELECT DISTINCT ?ciudad ?poblacion WHERE {{?ciudad dbpedia-owl:type dbpedia:Municipalities_of_Spain} OPTIONAL {?ciudad dbpedia-owl:populationTotal ?poblacion.}}
2	SELECT DISTINCT ?ciudad ?poblacion WHERE {{?ciudad rdf:type dbpedia-owl:City. ?ciudad dbpedia-owl:country dbpedia:Spain} OPTIONAL {?ciudad dbpedia-owl:populationTotal ?poblacion.}}
3	SELECT DISTINCT ?ciudad ?poblacion WHERE {{?ciudad dbpedia-owl:country dbpedia:Spain. ?ciudad dcterms:subject ?b. ?b skos:broader category:Municipalities_of_Spain} OPTIONAL {?ciudad dbpedia-owl:populationTotal ?poblacion.}}
4	SELECT DISTINCT ?ciudad ?poblacion WHERE {{?ciudad dbpedia-owl:isPartOf ?comarca. ?comarca dbpprop:blankNameSec dbpedia>List_of_municipalities_of_Spain} OPTIONAL {?ciudad dbpedia-owl:populationTotal ?poblacion.}}

mos y comparamos los resultados se aprecia que el solapamiento de las definiciones es parcial (tabla 1), por lo que el concepto que representan podría no ser exactamente el mismo.
<http://lov.okfn.org/dataset/lov/index.html>

<http://metadataregistry.org>

Algunos trabajos, como el de **Fuentes-Lorenzo, Morato y Gómez** (2009), proponen mostrar el contexto en lenguaje natural y anotadores que conectados a vocabularios simplifiquen su codificación. No obstante, sin visualizar las relaciones del término, el problema persiste.

Formulación de la consulta

Hemos desarrollado el ejemplo propuesto previamente sobre “un listado de municipios españoles y la población correspondiente a cada uno de ellos” en la *DBpedia*. En la tabla 2 observamos la formulación de consultas alternativas en *Sparql*, pero no equivalentes en su resultado. En el ejemplo las entidades buscadas pueden estar categorizadas como: *Municipalities_of_Spain* (consulta 1), una consulta más limitada con *city* con un valor de filtrado igual a *Spain* (consulta 2), ciudades que tengan el término genérico *Municipalities_of_Spain* (consulta 3), o una consulta equivalente a la tres para el caso de que haya una subdivisión administrativa como provincia o similar (consulta 4).

Se confirma que la formulación de esta consulta implica al menos tres dificultades:

- conocimiento de la sintaxis del lenguaje de búsqueda y la posibilidad de combinar tripletas y filtros;
- identificación de los conceptos sobre los que se han representado los datos;
- en caso de consultas alternativas, selección de la formulación con datos más confiables.

3. Resultados de las consultas

La figura 2 muestra los grafos resultado de las cuatro consultas alternativas sobre población de municipios en España del apartado anterior en *OpenLink iSPARQL*. Observamos como cada secuencia de nodos muestra la relación que los

une, y cada una de estas secuencias corresponde a una de las consultas realizadas en el apartado anterior.

Aunque todas las búsquedas devuelven resultados válidos, al compararlas se observa que no son equivalentes: no se obtienen los mismos datos en cuanto a nombres, número de municipios y población.

Los datos obtenidos se han comparado con los publicados por el *Instituto Nacional de Estadística (INE)* para 2013. El *INE* recoge 8.249 municipios en España. Se han considerado las variaciones ortográfica o idiomática. En la tabla 3 se observa el número de lugares recuperados y coincidentes con los recogidos en el *INE*. Se han calculado también las tasas de precisión y exhaustividad sin tener en cuenta el dato de población ni el año al que corresponde ese valor

Una observación más pormenorizada de los datos muestra que entre los resultados de las consultas 1 y 2 el porcentaje de ciudades coincidentes es sólo de un 3% y aproximada-

ciudad	poblacion
http://dbpedia.org/resource/Barcelona	3218071
http://dbpedia.org/resource/Barcelona	1620943
http://dbpedia.org/resource/Murcia	442573
http://dbpedia.org/resource/Palma_Majorca	401270
http://dbpedia.org/resource/Las_Palmas	383308
http://dbpedia.org/resource/Bilbao	351629
http://dbpedia.org/resource/C%C3%B3rdoba,_Andalusia	325453
http://dbpedia.org/resource/Valladolid	311501
http://dbpedia.org/resource/Gij%C3%B3n	281649
http://dbpedia.org/resource/L'Hospitalet_de_Llobregat	258642
http://dbpedia.org/resource/A_Coru%C3%B1a	246056
http://dbpedia.org/resource/Vitoria-Gasteiz	238247
http://dbpedia.org/resource/Jundiz_recycling_plant	235661
http://dbpedia.org/resource/Elche	230112
http://dbpedia.org/resource/Oviedo	224005
http://dbpedia.org/resource/Santa_Cruz_de_Tenerife	222417
http://dbpedia.org/resource/Badalona	220977
http://dbpedia.org/resource/Cartagena_Spain	218210

Figura 3. Resultado de la consulta 1 sobre la población de municipios

mente del 84% para las parejas 1 y 3, y 3 y 4. En el caso de ciudades coincidentes entre las consultas 2 y 3 es un 96%. Por ejemplo Madrid sólo se obtiene en la consulta 3 y Barcelona se obtiene en todas salvo en la consulta 2. Los datos de las poblaciones coincidentes de la *DBpedia* y el *INE* son significativamente diferentes. Con estos datos se muestra que no todas las formulaciones tienen los mismos niveles de ruido y silencio en la recuperación.

Tabla 3. Exhaustividad y precisión respecto a los datos del *INE* <http://www.ine.es>

Pregunta	Coincidentes <i>INE</i> (T test 95%)	Exhaustividad	Precisión	Ciudades encontradas sin dato población (%)
1	1.442 (p=0,022)	0,17	0,88	7
2	192 (p=0,084)	0,02	0,85	1
3	4.911 (p=0,003)	0,59	0,92	2
4	1.325 (p=0,043)	0,16	0,86	3

A modo de ejemplo, en la figura 3 se muestran los resultados de la consulta 1. En el listado completo se detectan ausencias significativas, incoherencias, e incluso datos erróneos. Otro ejemplo, que dificultaría las consultas relacionadas (*Uren et al., 2007*) sería la consulta “*Select ?a ?b Where {?a skos:broader ?b. ?b skos:broader ?a.}*” que proporciona como resultado 1.566 conceptos cuyo genérico y específico es el mismo concepto.

Los principales problemas encontrados en la recuperación son:

- dificultad para formular consultas en el lenguaje de recuperación debido a su complejidad;

Tabla 4. Propuestas orientadas a mostrar de un modo gráfico las relaciones

Propuestas	Principal función	Problema que afronta
<i>VisualSparql</i> y <i>Welkin</i> http://graves.cl/visualSparql http://simile.mit.edu/welkin	Visualizar gráficamente consultas <i>Sparql</i>	Dificultad de uso del lenguaje de recuperación
Di-Martino (2010)	Transformar consulta <i>PLN</i> en <i>Sparql</i> y representarla gráficamente	Dificultad de traducir el lenguaje natural al de consulta
<i>Cypher</i> http://docs.neo4j.org/chunked/milestone	Lenguaje de consulta gráfico con bases de datos gráficas <i>Neo4j</i>	Dificultad de uso del lenguaje de recuperación
<i>Iconvis</i> y <i>Wikimindmap</i> http://iconvis.polito.it/iconvis http://www.wikimindmap.org	Búsqueda gráfica secuencial y facetada de entidades del nombre	Problemas para restringir el ámbito de búsqueda
<i>Sgvizler</i> http://dev.data2000.no/sgvizler	Resultados con distintos tipos de gráficos (histogramas, dispersión, geolocalización, etc.)	Escasa adaptación de los buscadores a distintos tipos de consultas
Mejía-Sánchez-Bermejo (2013)	Desambiguación mediante el cálculo de la distancia semántica en el grafo	Problemas asociados a la denominación de un concepto en un vocabulario debido a sinonimia y polisemia
<i>Thinkpedia</i> y <i>Thinkbase</i> (Hirsch; Hosking; Grundy , 2009)	Visualización de vocabularios	Dificultad de navegar debido a la abundancia de conceptos

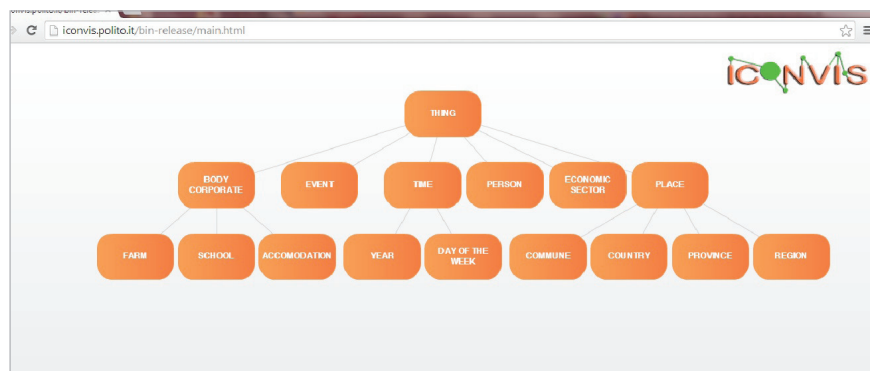


Figura 4. Consulta en Iconvis <http://iconvis.polito.it/iconvis>

visualización de información pueden interactuar entre sí para simplificar y optimizar el proceso de consulta. Es decir, se trata de mejorar la visualización de los recursos de la web semántica, dado que esta visualización tiene como objetivo representar objetos de información de forma que el acceso a la información y al contenido subyacente sea más sencillo y comprensible (Baeza-Yates; Rivera-Loaiza; Velasco, 2004). Dependiendo de la información, ésta podrá ser visualizada como un dato, una imagen, una tabla o un grafo.

- posibles inconsistencias en las definiciones entre vocabularios al definir el mismo concepto;
- problemas asociados a la denominación de un concepto en un vocabulario debido a sinonimia y polisemia;
- dificultad para navegar debido al volumen de conceptos;
- dificultad para saber a qué vocabulario/s debemos interrogar;
- escasa usabilidad y funcionalidad de los buscadores de la web semántica a los distintos tipos de consultas y usuarios no expertos;
- problemática con moléculas RDF, es decir dificultad de expresar determinadas ideas en una única tripleta.

La complejidad del lenguaje *Sparql* y las deficiencias en la visualización limitan la popularidad de los recursos semánticos

Visualización y navegación web

El volumen de datos que puede producir una consulta y la dificultad para interpretar los resultados ha dado lugar a iniciativas encaminadas a visualizar gráficamente los vocabularios, tratando de simplificar la consulta y contextualizando los conceptos. Hirsch, Hosking y Grundy (2009) señalan que la web social, la web semántica y las distintas técnicas de

Algunos ejemplos de herramientas orientadas a la visualización y la navegación web son *VisualSparql*, *Welkin*, *Iconvis*, *Cypher*, entre otras muchas. Por ejemplo, *Iconvis* (figura 4) permite navegar entre nodos asociados de forma gráfica. Cada nodo se expande y muestra las clases y propiedades asociadas, convirtiendo el evento en un conjunto de consultas *Sparql*. Tras varios nodos se muestra el resultado en la *Wikipedia*.

La tabla 4 recopila propuestas que muestran de modo gráfico las relaciones entre conceptos enlazados semánticamente, y expone la función más relevante y las principales dificultades que conlleva su uso.

Tabla 5. Elementos para el análisis de sistemas de recuperación semántica

Uren et al. (2007)	Mangold (2007)	Strasunskas y Tomassen (2010)	Morato et al. (2013)
Entorno de búsqueda: escala, heterogeneidad y portabilidad	Arquitectura	Arquitectura	
Tipos de consulta	Necesidades de información del usuario	Objetivo de la búsqueda	Tipos de consulta
Posibilidad de realizar consultas de forma iterativa y exploratoria: refinamiento, recomendación y reutilización	Posibilidad de modificación de la consulta. Transparencia e interactividad	Expresión de búsqueda Input del usuario (palabras clave, lenguaje natural, gráficos, consulta formal e interactiva)	
	Estructura de la ontología	Sistemas de organización del conocimiento (KOS)	Gestión semántica: desambiguación, multilingüismo, sinonimia, reuso
	Formalización de la ontología	Formalización de la ontología (RDFS, OWL, ...)	Lenguaje en que se expresa Posibilidad de modificar el esquema para mejorar la interoperabilidad, formalización, interactividad
	Acoplamiento entre ontología y documentos		Posibilidad de gestionar el marco semántico y capacidad de modificar ámbito, extensibilidad, modificabilidad
		Alcance (según sea aplicación de escritorio o web)	
Problemas intrínsecos: Comprensión, ranking de resultados y alineamiento			

4. Criterios para la evaluación de buscadores semánticos

Algunos autores sostienen que las medidas clásicas de precisión y exhaustividad son aplicables a la evaluación de motores semánticos de recuperación (Tümer; Shah; Bitirim, 2009; Andago; Phoebe; Thanoun, 2010). En general estos estudios no evalúan recursos de la web semántica, sino motores de documentos en lenguaje natural que aplican técnicas de extracción semántica y sistemas de organización del conocimiento.

Otros autores como Strasunskas y Tomassen (2010) señalan que la satisfacción del usuario al enfrentarse a la complejidad de una aplicación de un recuperador semántico no puede ser medida con estas métricas de recuperación de la información. La solución según estos autores es una evaluación que incluya aspectos cualitativos, como la calidad de los documentos formalizados, la calidad de la ontología subyacente, la arquitectura, usabilidad o la dificultad de la consulta. Diversos autores (Esmaili; Abolhassani, 2006; Uren et al., 2007; Mangold, 2007; Morato et al., 2013; Strasunskas; Tomassen, 2010) citan los siguientes elementos a evaluar:

- análisis de los buscadores según el tipo de consulta;
- calidad de los sistemas de organización de conocimiento (relaciones de dominio, meronimia, sinonimia, jerarquía, instancias y negación);
- similitud entre documentos y la ontología;
- criterios de indización de los documentos con la ontología;
- capacidad para representar, modificar o expandir la consulta o interactuar con el usuario;
- presentación o visualización de los resultados.

Muchos de estos factores son interdependientes, por ejemplo, los criterios de indización o la capacidad de expandir la consulta dependen de la riqueza semántica representada en el sistema de organización de conocimiento.

Como se puede observar en la tabla 5, todos los autores consideran relevante los tipos de consulta en la evaluación. Por otro lado, algunos señalan la necesidad de incluir una evaluación de la interfaz de búsqueda (Strasunskas; Tomassen, 2010), mientras otros se centran en las características de la ontología subyacente (Morato et al., 2013). Hirsch, Hosking y Grundy (2009) mencionan que dada la diferente calidad de los recursos subyacentes, se dificultan las pruebas de usabilidad

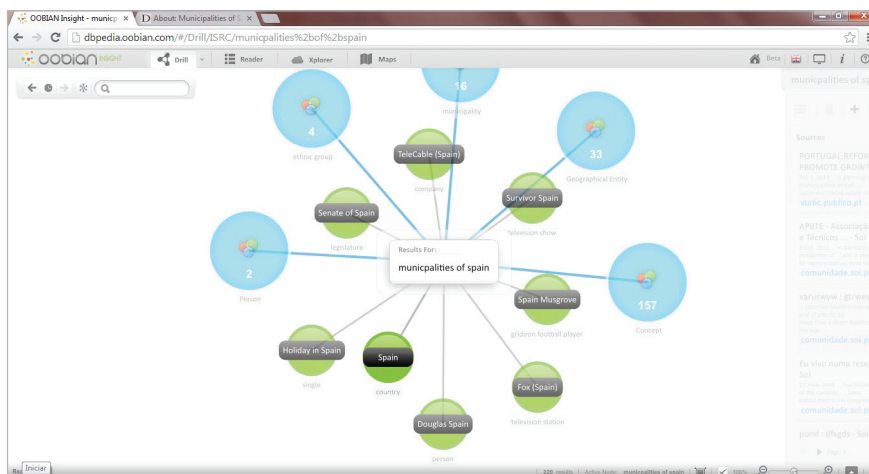


Figura 5. Oograph

heurística en los usuarios. Como alternativa, estos investigadores proponen recopilar los comentarios de los usuarios al navegar por el recurso.

Existen variantes para representar el mismo concepto y formular la misma consulta

5. Análisis de las interfaces gráficas de consulta semántica

Vistos los problemas en la recuperación, se han analizado interfaces gráficas orientadas a facilitar la formulación de consultas semánticas. El objetivo es determinar en qué medida son una solución a algunos de los problemas discutidos. Se ha restringido el estudio a aquellas interfaces que pueden ser consultadas online sin precisar una instalación local (como es el caso de Allegrograph). También para evitar problemas derivados de la existencia de varias ontologías se ha limitado el estudio a aquellos recursos que operan con DBpedia.

Es necesario analizar los buscadores semánticos en escena-

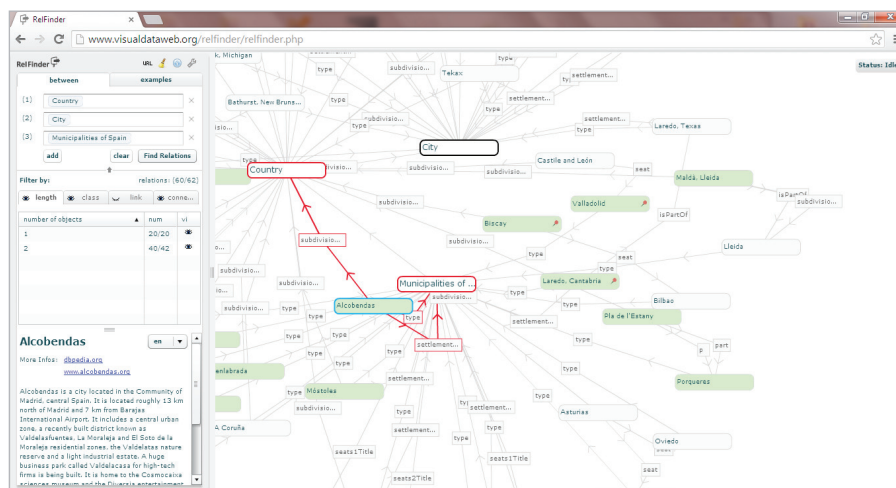


Figura 6. Visual dataweb

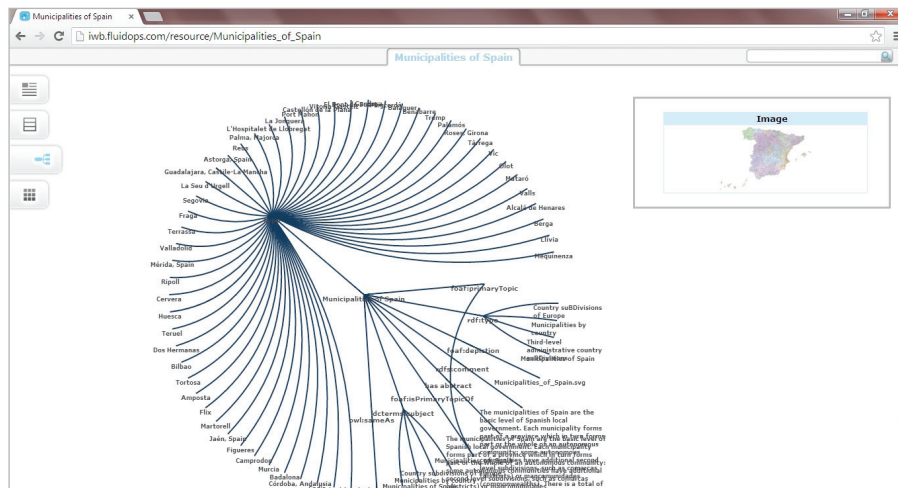


Figura 7. *Linked open data demonstrator*

rios que incorporen elementos como interactividad, filtrado y diversos tipos de consulta. Se ha propuesto una tarea consistente en una ampliación de la consulta de ejemplo sobre municipios españoles, posteriormente refinando a los diez más poblados y accediendo a páginas web relacionadas. Señalando las principales ventajas y desventajas a partir de nuestra experiencia con el buscador. Finalmente se ha realizado una comparativa de algunos aspectos indicados en la tabla 5.

A continuación se analizan algunas de las características de los buscadores:

Oobian

<http://dbpedia.oobian.com>

Aplicación desarrollada por *Maisis*, compañía especializada en gestión del conocimiento y de contenidos. Es una de las que mejores resultados ha proporcionado para probar el escenario propuesto.

- Ventajas: posee una interfaz intuitiva. Permite refinar la consulta y la visualización es rápida, accediendo a recursos web relacionados.
- Desventajas: escasa capacidad de combinar entidades. La consulta *Sparql* generada no es visible.
- Escenario: presentó problemas para combinar entidades, descargar los listados, ordenar o filtrar resultados.

Visual dataweb relfinder

<http://www.visualdataweb.org/relfinder/relfinder.php>

La aplicación se subdivide en: *Relfinder*, especializado en buscar relaciones entre datos *RDF*, permitiendo acceder a nuevas tripletas. Utiliza *Adobe Flex*, lo cual le dota de mayor portabilidad. Por otro lado, *GFacet* permite una exploración a partir de cualquier elemento de la tripleta, con capacidad de filtros y de unir nuevos nodos.

- Ventajas: posibilidad de interrogar por varias tripletas. Muestra mucha información sobre las entidades interrogadas.

-Desventajas: al tener sus funciones divididas entre varias herramientas su uso es más difícil. Además de la ambigüedad de los términos sugeridos, el abarrotamiento de términos, la dificultad de filtrado y tiempos excesivos complican su uso.

-Escenario: aunque ha sido una de las más completas y permite interrogar por varias tripletas y elegir caminos alternativos, presenta problemas para filtrar o establecer un ranking sobre cuál es el camino más acertado.

Linked open data demonstrator

<http://iwb.fluidops.com>

Es una plataforma para *linked data*, que integra varias fuentes (*DBpedia*, *Twitter* o *Facebook*). Permite integrarlo con otros recursos como *big data*.

- Ventajas: fácil de utilizar, y gráficos novedosos. Contiene recursos asociados a los términos. Fácil de consultar en lenguaje natural, pero no para todos los tipos de consultas. Es sencillo obtener un listado con las distintas tripletas a partir de un recurso.
- Desventajas: la visualización no es intuitiva, incluso la fuente llega a ser ilegible. Las consultas relacionales son complicadas de realizar.
- Escenario: las dificultades en interrogación simultánea por varios conceptos, obtención de listados y la gran densidad de conceptos dificultan su uso.

Wikistalker

<http://sepans.com/wikistalker/#>

Muestra los enlaces de cada artículo de la *Wikipedia*.

- Ventajas: visualización en un solo vistazo de todos los enlaces del artículo, es de las pocas que posee un ranking visual de resultados.
- Desventajas: la visualización se resiente del elevado número de enlaces, no permite búsquedas más allá de las de un artículo.

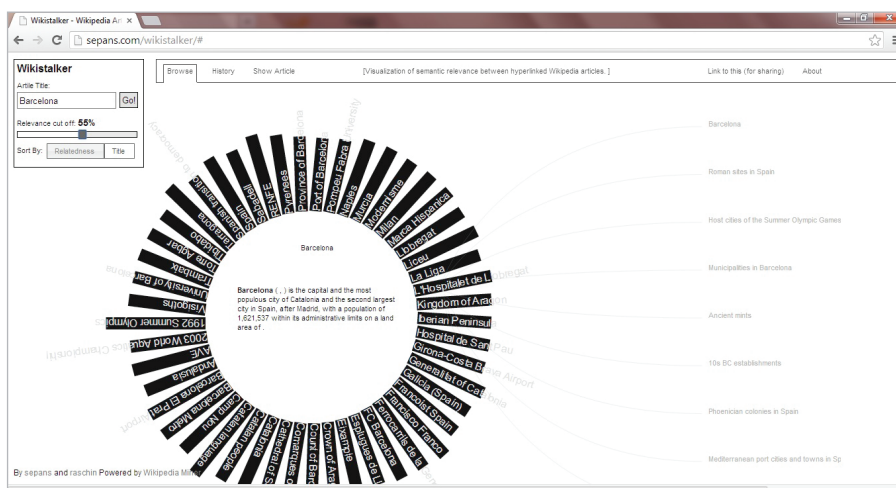


Figura 8. *Wikistalker*

- Escenario: aunque indica enlaces relacionados, el acceso al documento o la descarga de los resultados de la búsqueda no es posible. Como en otros casos, el acceso a varias tripletas de forma simultánea no forma parte de las funciones del sistema.

Lodlive
<http://en.lodlive.it>

Se trata de una interfaz usable y funcional. Se basa en tecnologías de gran difusión y libres como *JavaScript*, *jQuery*, *html* y *JSON*.

- Ventajas: facilidad para navegar entre resultados relacionados semánticamente.
- Desventajas: dificultad para consultar varias tripletas al mismo tiempo y acceder a los documentos de resultados.
- Escenario: la imposibilidad de combinar varias tripletas, descargar listados y documentos asociados y volumen de datos mostrados sin limitar el espacio de búsqueda hacen imposible la ejecución de las tareas.

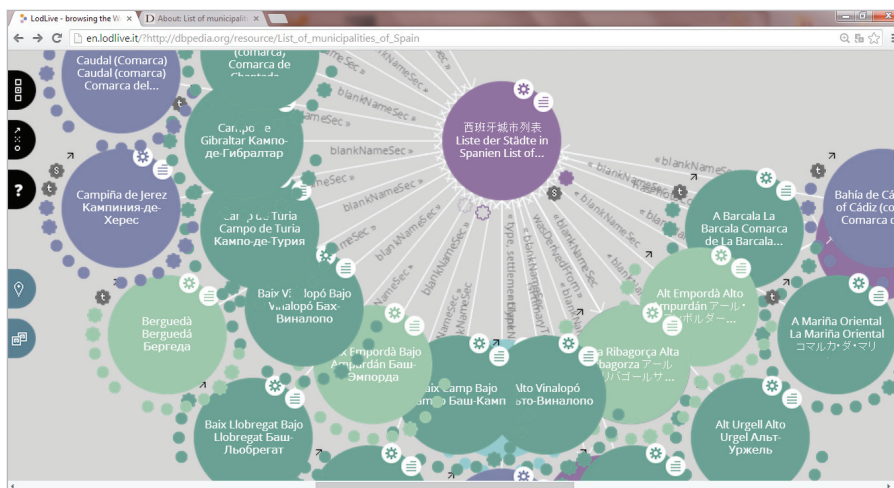


Figura 9. Lodlive

Las interfaces gráficas muestran grandes volúmenes de datos pero escasas posibilidades de filtrar y combinar consultas

En la tabla 6, basada en la tabla 5, se analizan aspectos relacionados con los tipos de consulta, viendo la posibilidad gráfica de expansión de la consulta a elementos relacionados semánticamente, acceso documental a páginas web relacionadas y posibilidad de descargar elementos de las tripletas resultantes (como listas de entidades). El refinamiento (Uren et al., 2007; Mangold, 2007) se ha analizado observando la capacidad de buscar instancias en las que se solapan varios conceptos (ejemplo: solapamiento de conceptos población y municipio) y la posibilidad de filtrado (ejemplo: municipios con más población). Elementos como portabilidad (Uren et al., 2007), multilingüismo y desambiguación (Morato et al., 2013) y expresión de búsqueda (Strasunskas; Tomassen, 2010) se estudiaron mediante la posibilidad de desambiguación de homógrafos y visualización de la consulta generada en lenguaje *Sparql* o similar. Casi todas las herramientas permiten comenzar la consulta en lenguaje libre, pero con

unitérminos, mostrando listas de conceptos relacionados. Por último, se vio la posibilidad de realizar un ranking bien entre consultas alternativas o con los resultados, como propone Uren et al. (2007).

En algunos datos en la comparativa de la tabla 6, la implementación es parcial, así en Lodlive la consulta *Sparql* se puede introducir, pero no se puede obtener la conseguida con la navegación. Como conclusión se puede observar que la ausencia de ranking, de descarga de tripletas, de búsqueda de elementos concurrentes e integración con *Sparql* limitan estas herramientas. Por lo tanto el escenario propuesto no ha podido completarse dado el escaso soporte para filtrados y consultas a múltiples conceptos (sólo *Visual dataweb* lo soporta).

Conclusiones

El uso de los recursos semánticos depende de la disponibilidad de buscadores orientados a los usuarios. Para ello, los motores de la web semántica deben optar por simplificar la interfaz de recuperación y representar los recursos y/o los resultados de forma gráfica, con la presunción de que una visualización de los recursos interrelacionados mejora la interpretación de los datos.

Tras probar la ejecución en diferentes buscadores semánticos se ha comprobado que no es factible a día de hoy pres-

Tabla 6. Comparación de herramientas para algunas funciones

	Oograph	Visual dataweb	LOD demonstrator	Wikistalker	Lodlive
Desambiguación homógrafos en la consulta	Sí	Parcial	Sí	Sí	Sí
Multilingüismo	Sí	Sí	Sí	Sí	Parcial
Búsqueda con expansión a relacionados	Sí	Sí	Sí	Sí	Sí
Búsqueda por varios conceptos concurrentes	No	Sí	No	No	No
Filtrado de resultados	Parcial	No	No	No	No
Ranking consultas o resultados	No	No	No	Sí	No
Acceso documentos resultantes	Sí	Sí	No	No	Sí
Descarga tripletas resultantes (ejemplo: listas)	No	No	No	No	No
Visualización de la consulta <i>Sparql</i>	No	No	No	No	Parcial

cindir del lenguaje de interrogación tipo *Sparql*. En cuanto a cómo se representan los datos, se obtienen resultados diferentes condicionados a la formulación de la consulta, que requiere conocer además los vocabularios de metadatos por los que está representado el recurso que se desea interrogar. Conocer la representación de los datos en los recursos también es una tarea ardua, porque aunque existe control de vocabulario en cada uno de los recursos mediante los vocabularios de metadatos, son muchos los conceptos similares pero representados de forma diferente. Existe por tanto control de vocabulario a nivel metadato, pero no al fusionarse con otros vocabularios, por lo que es necesario conocer previamente cómo están representados los datos antes de poder interrogarlos.

« Aun se requieren elementos que faciliten la recuperación y comprensión de los datos y recursos recuperados »

La consecuencia más inmediata es la complicación para usar el motor de búsqueda. En comparación con los buscadores de la web tradicional, quedan patentes las diferencias en cuanto al tipo de consulta (que está lejos de parecerse a una consulta por palabras clave), a los requisitos necesarios para formular la interrogación, y al tipo de recurso que se va a recuperar.

Estos buscadores difícilmente pueden ser evaluados sólo mediante métricas de precisión y exhaustividad. Creemos que deben añadirse características cualitativas asociadas a la usabilidad y funcionalidad. La capacidad de los buscadores para localizar determinada información dentro del gran volumen de documentos en la web semántica podría ser mejorada con algoritmos de posicionamiento para determinar los *datasets* o vocabularios más fiables.

La función de los buscadores visuales actuales se limita a mostrar los nodos directamente relacionados, y con la ejecución de las consultas complejas no se obtienen los resultados esperados. La visualización de los datos no es clara en aquellas consultas cuyos resultados son muy abundantes y con muchas relaciones. La navegación es compleja puesto que no existen directrices que orienten al usuario. Consideramos que aún se requieren elementos que faciliten la recuperación y comprensión de los datos y recursos recuperados.

Agradecimientos

Agradecemos la financiación recibida por el proyecto *Alma-histo* (HAR2011-27540) del *Mincin*.

Bibliografía

Andago, Martin O.; Phoebe, The P. L.; Thanoun, Bassam A. M. (2010). "Evaluation of a semantic search engine against a keyword search engine using first 20 precision". *International journal for the advancement of science & arts*, v. 1, n. 2, pp. 55-63.
<http://www.ucsiuniversity.edu.my/cervie/pdf/ijasa/paperV1N2IT3.pdf>

Baeza-Yates, Ricardo; Rivera-Loaiza, Cuauhtémoc; Velasco, Javier (2004). "Arquitectura de la información y usabilidad en la web". *El profesional de la información*, mayo-junio, v. 13, n. 3, pp. 168-178.

<http://www.elprofesionaldelainformacion.com/contenidos/2004/mayo/1.pdf>

Bauer, Florian; Kaltenböck, Martin (2011). *Linked open data: the essentials*. Viena: Semantic Web Company. ISBN: 978 3 902796 05 9

<http://www.semantic-web.at/LOD-TheEssentials.pdf>

Bizer, Christian; Heath, Tom; Berners-Lee, Tim (2009). "Linked data—the story so far". *International journal on semantic web and information systems*, v. 5, n.3, pp. 1-22.

<http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>

<http://dx.doi.org/10.4018/jswis.2009081901>

Broader, Andrei (2002). "A taxonomy of web search". En: *Sigir Forum*, v. 36, n. 2, pp. 3-10.

<http://www.cis.upenn.edu/~nenkova/Courses/cis430/p3-broder.pdf>

<http://dx.doi.org/10.1145/792550.792552>

Di-Martino, Beniamino (2010). "An approach to semantic information retrieval based on natural language query understanding". En: *ICWE 2010 Workshop*, pp. 211-222.

http://dx.doi.org/10.1007/978-3-642-16985-4_19

Esmaili, Kyumars-Sheykh; Abolhassani, Hassan (2006). "A categorization scheme for semantic web search engines". En: *Aiccsa*, pp. 171-178.

<http://dx.doi.org/10.1109/AICCSA.2006.205086>

Fazzinga, Bettina; Lukasiewicz, Thomas (2010). "Semantic search on the web". *Semantic web: interoperability, usability, aplicability*, v. 1, n. 1-2, pp. 89-96.

<http://www.semantic-web-journal.net/sites/default/files/swj51.pdf>

<http://dx.doi.org/10.3233/SW-2010-0023>

Fuentes-Lorenzo, Damaris; Morato, Jorge; Gómez, Juan-Miguel (2009). "Knowledge management in biomedical libraries: a semantic web approach". *Information systems frontiers*, v. 11, n. 4, pp. 471-480.

http://eprints.networks.imdea.org/111/1/Knowledge_management_in_biomedical_libraries_a_semantic_web_approach.pdf

<http://dx.doi.org/10.1007/s10796-009-9159-y>

Hirsch, Christian; Hosking, John; Grundy, John (2009). "Interactive visualization tools for exploring the semantic graph of large knowledge spaces". En: *Workshop on visual interfaces to the social and the semantic web* (Vissw2009).

<http://ceur-ws.org/Vol-443/paper6.pdf>

Lopez, Vanessa; Uren, Victoria; Sabou, Marta; Motta, Enrico (2011). "Is question answering fit for the semantic web? A survey". *Semantic web: interoperability, usability, aplicability*, n. 2, pp. 125-155.

<http://www.semantic-web-journal.net/content/question-answering-fit-semantic-web-survey>

<http://dx.doi.org/10.3233/SW-2011-0041>

Mangold, Christoph (2007). "A survey and classification of

semantic search approaches". *International journal of meta-data, semantics and ontologies*, v. 2, n. 1, pp. 23-34.

<http://dee.srv1.eu/refs/files/1/65.pdf>

<http://dx.doi.org/10.1504/IJMSO.2007.015073>

Mejía-Sánchez-Bermejo, Antonio (2013). *Similitud semántica entre conceptos de Wikipedia*. Trabajo fin de grado. Universidad Carlos III de Madrid.

<http://hdl.handle.net/10016/17170>

Méndez, Eva; Greenberg, Jane (2012). "Linked data for open vocabularies and HIVE's global framework". *El profesional de la información*, mayo-junio, v. 21, n. 3, pp. 236-244.

http://www.elprofesionaldelainformacion.com/contenidos/2012/mayo/03_eng.pdf

<http://dx.doi.org/10.3145/epi.2012.may.03>

Miličić, Vuk (2011). "Introducing hypernotation an alternative to linked data". *Bew citnames*.

<http://milicicvuk.com/blog/2011/11/17/introducing-hypernotation-an-alternative-to-linked-data>

Morato, Jorge; Sánchez-Cuadrado, Sonia; Dimou, Christos; Yadav, Divakar; Palacios, Vicente (2013). "Evaluation of semantic retrieval systems on the semantic web". *Library hi tech*, v. 31, n. 4, pp. 638-656.

<http://dx.doi.org/10.1108/LHT-03-2013-0026>

Strasunskas, Darijus; Tomassen, Stein (2010). "On variety of semantic search systems and their evaluation methods".

En: *Procs of the Intl conf on information management and evaluation*, pp. 380-387.

Tümer, Duygu; Shah, Mohammad-Ahmed; Bitirim, Yiltan (2009). "An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, Yahoo, MSN and Hakia".

En: *4th Intl conf on internet monitoring and protection (Icimp'09)*, pp. 51-55.

<http://dx.doi.org/10.1109/ICIMP.2009.16>

Uren, Victoria; Lei, Yuangui; Lopez, Vanessa; Liu, Haiming; Motta, Enrico; Giordanino, Marina (2007). "The usability of semantic search tools: a review". *The knowledge engineering review*, v. 22, n. 4, pp. 361-377.

<http://oro.open.ac.uk/23501/1/download.pdf>

<http://dx.doi.org/10.1017/S0269888907001233>

Wei, Wang; Barnaghi, Payam; Bargiela, Andrzej (2008). "Search with meanings: an overview of semantic search systems". *International journal communications of SIWN*, n. 3, pp. 76-82.

<http://personal.ee.surrey.ac.uk/Personal/P.Barnaghi/doc/siwn2008Survey.pdf>

<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.324.2897>

iralis®

International Registry for Authors:
Links to Identify Scientists

es:

- una **guía** para los autores hispanos para que firmen sus trabajos en el formato internacional usual
- una **base de datos** que registra las variantes de firma usadas por cada autor en diferentes épocas
- un **buscador** que usa automáticamente todas las variantes registradas

¡Ni un día más sin decidir tu firma!

Inscríbete en <http://iralis.org>

**El formato de
firma internacional**

Con el apoyo de:

El profesional de la
información

