eib0915856707

Recuperación de información con *Linked Open Data*

Eder Ávila-Barrientos*

Artículo recibido:
10 de enero de 2022
Artículo aceptado:
22 de abril de 2022

Artículo de investigación

Resumen

El objetivo de este artículo es analizar el proceso de recuperación de información (RI) mediante la aplicación de *Linked Open Data* (LOD). La metodología empleada en este trabajo consta de dos partes: en la primera, se llevó a cabo una revisión de literatura, se apoya en la hermenéutica y análisis del discurso para examinar recursos de información que abordaran modelos, estructuras, estudios de caso, pruebas de concepto y análisis estructurales de la implementación de datos abiertos enlazados en el proceso de RI. Sucesivamente, partiendo del método analítico-sintético, se seleccionó una fuente de datos cuya licencia

* Instituto de Investigaciones Bibliotecológicas y de la Información, Universidad Nacional Autónoma de México, México eder@iibi.unam.mx

abierta admitiera reutilizar y procesar los datos para exponerlos al contexto de recuperación de información, mediante su estructuración en RDF y su respectiva consulta con SPARQL. Los resultados obtenidos permiten acercarse a los principios metodológicos del procesamiento de datos abiertos enlazados y analizar la información adquirida, siendo estos de carácter integrador e intuitivos derivados del uso de representaciones gráficas para hacer posible la consulta y acceso a los datos. La adaptación de LOD en los sistemas para la RI, supone un cambio de paradigma relacionado con la clase de tecnología empleada en sus estructuras. Por ejemplo, transitar de un modelo conceptual sintáctico a otro de tipo semántico.

Palabras clave: Datos abiertos enlazados; Recuperación de información; SPARQL; RDF

Information retrieval with Linked Open Data

Eder Ávila-Barrientos

Abstract

The objective of this article is to analyze the process of information retrieval (IR) through the application of *Linked Open Data* (LOD). The methodology used in this work consists of two parts. In the first one, a literature review was carried out, supported by hermeneutics and discourse analysis to examine information resources that addressed the models, the structures, the case studies, the concept tests and the structural analysis of the implementation of linked open data in the information retrieval process. Successively, based on the analytical-synthetic method, an open-license data source was selected due to the fact that it allowed the data to be reused and processed in order to expose it to the information retrieval context, by structuring it in RDF and its respective query with SPARQL. The results obtained permit an approach to the methodological principles of linked open data processing and an analysis of the information acquired, which has an integrative and intuitive nature derived from the use of graphic representations that enables data query and access. The adaptation of LOD in information retrieval systems represents a paradigm shift

related to the type of technology used in their structures. For example, moving from a syntactic conceptual model to a semantic one.

Keywords: Linked Open Data; Information Retrieval; SPARQL; RDF

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

Introducción

Desde hace más de una década, *Linked Open Data* (LOD) ha propiciado la generación de propuestas para publicar y conectar a los datos de la web de una manera abierta e interoperable. Debido a esto, se ha desarrollado una amplia cantidad de proyectos, vocabularios y herramientas de índole computacional con el objetivo de adaptarse a diferentes tipos de datos y hacerlos compatibles con los principios de LOD. No obstante, uno de los ámbitos que ha sido poco explorado lo conforma la aplicación de LOD en el proceso de la recuperación de información (RI), se estima que la adaptación de ambos elementos en los sistemas de información provoque un cambio de paradigma relacionado con los modelos actuales para estructurar y representar datos. Bajo esta premisa, LOD permitirá desarrollar consultas complejas de información para identificar las vinculaciones de significado entre un dato en particular con las obras, expresiones, manifestaciones y ejemplares que forman parte del universo de información, fomentando con ello la conexión de datos de una manera interna en su contexto, pero con la capacidad de vincularse con fuentes externas disponibles en el entorno digital. Por ello, este documento pretende abordar las siguientes preguntas:

1. ¿Cuáles son los principios que intervienen en la recuperación de información con *Linked Open Data*?
2. ¿Qué características tiene la recuperación de información (RI) mediante el uso de LOD?
3. ¿Cuáles son los alcances y limitaciones de este tipo de recuperación de información?

Así, el propósito de este trabajo consiste en analizar la aplicación de *Linked Open Data* en el proceso de recuperación de información.

Linked Open Data tiene origen en un principio relacionado con la publicación y conexión de datos disponibles en diferentes fuentes de la web: catálogos en línea, repositorios, bases de datos, conjuntos de estos últimos, por ejemplo. Los cuales deben estar libres de restricciones técnicas, económicas y legales, es decir, ser liberados mediante el uso de una licencia abierta que permita su reutilización. Además, LOD forma parte sustancial de un estándar internacional impulsado por el World Wide Web Consortium (W3C) para el desarrollo paulatino de la web semántica.

"El proyecto Linking Open Data tiene como objetivo identificar conjuntos de datos en la web que están disponibles bajo licencias abiertas, para volver a publicarlos en RDF e interconectarlos entre sí" (Bizer *et al*., 2008: 1265). Para ello, LOD emplea herramientas tecnológicas que hacen posible la construcción de un espacio común en donde se puedan consultar, recuperar y visualizar datos en diferentes fuentes de la web. Por ejemplo, *Resource Description Framework* (RDF), *Protocol and RDF Query Language* (SPARQL), *Uniform Resource Identifiers* (URI's), *Hypertext Transfer Protocol* (HTTP). Asimismo, emplea vocabularios para definir el modelo estructural de los datos que serán descritos y representados en un dominio interoperable. En este sentido, "RDF está diseñado para modelar información de manera flexible, su esquema representa objetos de datos como triples en la forma (S, P, O), donde S representa un sujeto, P representa un predicado y O representa un objeto" (Wylot y Sakr, 2019).

El hecho de recuperar y consultar datos disponibles en distintas fuentes plantea la necesidad de establecer interoperabilidad global entre los datos y los sistemas que los contienen. Este fenómeno evidencia la alta complejidad para adaptar los principios de LOD en los diversos sistemas de información. Además, pone de manifiesto la construcción de diversos proyectos, vocabularios y modelos que desean alcanzar este ambicioso propósito, creando un debate reciente acerca de los alcances y limitaciones de la conectividad integral entre sistemas locales y aquellos que están utilizables en el ambiente web. Al respecto, De Faria Cordeiro *et al*. (2011: 83) estiman que la interrelación de datos heterogéneos a través del uso de *Linked Data* supone enfrentar problemas como el reto de ofrecer apoyo a las organizaciones para publicar datos valiosos: aquellos que tienen un significado explícito que favorece su vinculación y acceso por parte del usuario final.

Además, el estudio de la interoperabilidad global de los datos y su implementación en los sistemas de información tiene cercanos antecedentes en el análisis semántico de los datos y la manera en cómo se comparten y comportan al momento de ser almacenados en un determinado sistema.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

De acuerdo con esa idea, Bar-Hillel y Carnap (1953: 146) plantearon, desde un ángulo filosófico apegado a la lógica, algunos de los fundamentos relacionados con la semántica de la información enfocados al análisis de probabilidad de similitud entre el contenido de los datos. Al respecto, (Bar-Hillel y Carnap 1953; Bar-Hillel 1964, como se citó en Floridi, 2016: 20) manifiestan que "si los datos son bien formados y significativos, el resultado es también conocido como contenido semántico". Bajo esta premisa, el análisis de los atributos de los datos otorga la posibilidad de definir su significado mediante la interpretación del contexto y contenido que los representa. Por lo tanto, la recuperación de la información bajo LOD se encarga del análisis de los atributos de los datos, a fin de recuperar aquellos con similitudes en su estructura y contexto.

Estudios previos han abordado la RI del mismo modo. Por ejemplo, Stab *et al.* (2013) realizaron uno en el cual se exponen los resultados de un método de visualización de búsqueda mediante LOD.

Esto supone una adaptación flexible de LOD en las interfaces para la búsqueda de información, en donde los usuarios interactúen de manera intuitiva con los resultados de búsqueda. Por su parte, Musto *et al.* (2017) implementaron los principios de LOD en sistemas de recomendación basados en grafos. En este trabajo, donde se retoma lo anterior, puede observarse una evaluación efectuada a la recuperación de información mediante la utilización de datos correspondientes a obras musicales relacionadas con las preferencias de los usuarios finales, empleando grafos para visualizar las conexiones entre las similitudes de los datos que se han procesado.

Además, pueden localizarse otros trabajos que abordan el estudio semántico y ontológico de los datos abiertos enlazados. Por ejemplo, Campos y De Almeida (2014) han desarrollado la aplicación de una ontología en el contexto de datos abiertos enlazados para describir conceptos y temas que están representados en datos de fuentes heterogéneas, con la particularidad de ofrecer un acercamiento a la interconexión entre una base de datos enlazados abiertos con el proyecto de la *DBpedia*. Esto, aplicado aquí, resulta representativo, pues constata el análisis de las relaciones de significado que deben existir entre los datos para obtener un grado elevado de conectividad, tomando en cuenta el significado y el contexto de los datos que se describen y representan en múltiples entornos.

Por otra parte, Baron Neto *et al.* (2016) han propuesto el desarrollo de una interfaz para visualizar y descubrir conjuntos de datos en tiempo real. En este trabajo, la implementación de LOD en el proceso de RI tiene un efecto en las bases de datos que soportan a los sistemas de información, pues LOD emplea modelos que han sido poco utilizados en el contexto informático de los sistemas, por ejemplo, el uso de RDF.

Al respecto, Wylot *et al.* (2017) realizaron un experimento para identificar cómo las bases de datos en RDF pueden rastrear el origen de los datos, a partir de ejecutar consultas complejas con alcance de procedencia. Es decir, conocer más detalles acerca de los datos que se almacenan en la base, así como la fuente y si han sido utilizados como parte de resultados de investigaciones.

Según Silvello *et al.* (2017: 145), LOD permite abrir datos públicos en formatos legibles por máquina listos para su consumo, para su reutilización y enriquecimiento a través de conexiones semánticas que habilitan la creación de nuevos conocimientos y posibilidades de descubrimiento.

Esto manifiesta la utilización de métodos visuales para recuperar información e identificar patrones de comportamiento complejos, es decir, ver más allá de lo evidente en cuanto a la disponibilidad y vinculación de los datos. De acuerdo con Zhang (2008: 2), la recuperación y la visualización de información tienen una relación natural e inherente. Una presentación visual, aparte de su contenido y forma, está destinada a transmitir información a las personas por un medio visual.

Por consiguiente, los datos abiertos enlazados necesitan de mecanismos relativos a la visión para ser recuperados en un determinado sistema.

Linked Open Data y recuperación de información

El origen de LOD se remonta a un principio que manifiesta el poder establecer vinculaciones de significado entre datos con atributos similares que están disponibles en diferentes fuentes de la web, lo cual supone el establecimiento de un entorno con interoperabilidad global. "Por lo tanto, *Linked Open Data* son datos enlazados que se publican bajo una licencia abierta, que no impide su reutilización de forma gratuita" (Berners-Lee, 2009: § 64).

Desde un punto de vista pragmático:

> los principios de LOD se emplean para establecer hipervínculos entre datos de diferentes fuentes. Estos hipervínculos conectan todos los datos vinculados en un solo grafo de datos global, similar a los hipervínculos en la web clásica que conecta todos los documentos HTML en un solo espacio de información global (Bizer, Vidal y Skaf-Molli, 2018).

La integración de los requerimientos técnicos y de los fundamentos de LOD resultan susceptibles de aplicarse en procesos de RI en concordancia con las características de un sistema de información. Para ello, se hace incapié en comprender cómo se lleva a cabo una implementación holística y sistémica

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

que tiene el propósito de propiciar la formulación de consultas complejas de información. De esta manera, "la recuperación de información es el conjunto de conocimientos que se enfoca en cómo encontrar de manera eficiente información relevante para satisfacer una determinada necesidad de información de un usuario o sistema" (Waitelonis, 2018: 15).

En la actualidad, se requiere realizar consultas de información complejas que permitan descubrir patrones ocultos en la información. Aunado a ello, los principios modernos de la recuperación de información también contemplan el uso de interfaces de usuario inteligentes, comentarios y etiquetado, por nombrar solo algunos. El término *base de datos* en sí se ha extendido a nuevas áreas como bibliotecas digitales y la *world wide web* (Linckels y Meinel, 2011: 81).

La consulta y recuperación de los datos abiertos enlazados necesita de métodos visuales que ayuden a examinarlos, así también sus respectivas vinculaciones. Además, es pertinente considerar que "la visualización de información tiene dos aspectos fundamentalmente relacionados: el modelado estructural y la representación gráfica" (Chen, 2004: 27). De esta manera LOD utiliza RDF *Schema* como modelo estructural de datos y a los grafos para obtener su representación gráfica. La vinculación de los datos quedará representada en el grafo que reúna las consultas y los datos con atributos de significado similares a los que se han recuperado.

> Las visualizaciones bien diseñadas aprovechan las poderosas capacidades del sistema de percepción humano, proporcionando a los usuarios representaciones ricas de datos. Combinados con técnicas de interacción apropiadas, permiten a los usuarios navegar y dar sentido a conjuntos de datos grandes y complejos, ayudan a detectar valores atípicos y anomalías, reconocer patrones e identificar tendencias (Dadzie y Pietriga, 2017: 2).

Además, como expone Chen (2004: 27) el producto visible de todo el proceso de visualización deriva en su representación, que es donde los usuarios interactuarán con la información presentada.

Por lo tanto, la recuperación de información con LOD consituye un proceso que va más allá de la publicación y estructuración de datos, involucra la implementación de métodos visuales que permitan identificar patrones complejos entre datos de atributos similares que remitan a información significativa para el usuario final. "Siempre que visualizamos datos, tomamos valores de datos y los convertimos de forma sistemática y lógica en los elementos visuales que conforman un grafo final" (Wilke, 2019, cap. 2: § 1). De esta manera, los datos se representan en nodos y aristas de un grafo en particular, por ende,

cada dato recuperado tendrá la facultad de remitir a los datos que forman parte de su contexto. Además, la visualización de información requiere de métodos certeros y algoritmos para convertir datos en bruto en representaciones que puedan ser interpretadas y accesibles para el usuario.

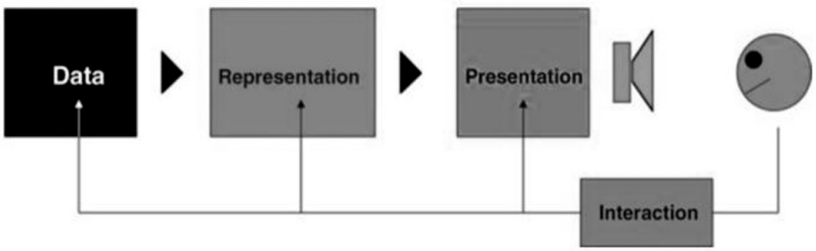

Figura 1. El modelo de referencia del proceso de visualización de información.
Fuente: Spence (2014: ix).

En la *Figura 1* se aprecia el modelo referencial del proceso de visualización de la información propuesto por Spence, en donde los datos conforman el elemento principal que será visualizado por el usuario final. En este proceso la representación y presentación permitirán al usuario interactuar con los datos mediante una serie de comportamientos, entre los cuales destacan sus estrategias de búsqueda y el conocimiento previo que tenga de los datos que se dispone a consultar.

La visualización de la información mediante representaciones gráficas facilita concentrarse en los detalles más específicos que una consulta convencional no puede identificar. Sin embargo, como menciona Mazza (2009: 11) cuando se representan datos de manera visual, se debe lidiar con el problema de su naturaleza.

Metodología

Se desarrolló un proceso de revisión de literatura, apoyado en la hermenéutica y análisis del discurso, para estudiar recursos de información relacionados con el objeto de estudio de la investigación. Se analizaron modelos, estructuras, estudios de caso, pruebas de concepto y analisis estructurales de la implementacion de datos abiertos enlazados en el proceso de recuperacion de informacion. Esos recursos fueron obtenidos en bases de datos especializadas en bibliotecología y estudios de la información y ciencias de la computación,

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

así como de catálogos y descubridores (académicos y especializados). Se implementaron estrategias de búsqueda relacionadas con los siguientes términos: *Linked Data*, *Linked Open Data*, *Information Retrieval* y su equivalente en español. Se seleccionaron obras publicadas de 2015 a 2021, pues al tratarse de un tema tecnológico el factor cronológico fue tratado con rigurosidad. Aquellas anteriores a este periodo fueron elegidas tomando en cuenta su grado de contribución y relevancia con respecto al análisis del objeto de estudio planteado en la investigación.

A su vez, tomando como base el método analítico-sintético, se seleccionó una fuente de datos cuya licencia abierta permitiera reutilizar y procesar los datos para exponerlos a un proceso de recuperación de información, mediante el uso de RDF y SPARQL. "RDF es un formato para la construcción de grafos de datos, por lo tanto, SPARQL es esencialmente un lenguaje para la consulta de estos grafos" (Pérez, Arenas y Gutiérrez, 2009: 2). La fuente elegida ha sido el catálogo en línea de la biblioteca del Instituto de Investigaciones Bibliotecológicas y de la Información de la UNAM, el cual se encuentra disponible (DGBSDI, 2020).

Etapa	Descripción	Procedimiento
Selección de fuente de datos	Identificación y selección de la fuente de los datos que serán adaptados a los principios de LOD	Análisis de formato, licenciamiento y tipo de datos
Obtención de datos	Aplicación de software (Zotero) para la descarga de los datos	Análisis de almacenamiento e integridad de los datos
Procesamiento	Manejo de los datos para su correcto procesamiento	Limpieza y modelado del conjunto de datos (OpenRefine)
Generación del conjunto de datos	Conformación del conjunto de datos	Definición del formato del conjunto de datos y su interoperabilidad con las herramientas para su aplicación
Aplicación del conjunto de datos en GraphDB	Carga y aplicación del conjunto de datos en el gestor de base enfocada a grafos (GraphDB)	Ajuste de parámetros de GraphDB con los atributos del conjunto de datos
Consulta con SPARQL	Análisis de la recuperación de información mediante consultas en SPARQL	Ejecutar consultas mediante SPARQL
Visualización del grafo RDF	Análisis de la visualización e interacción con el grafo	Identificación de propiedades, vinculaciones y atributos similares entre los datos

Tabla 1. Propuesta de metodología para la preparación de conjuntos de datos LOD y su aplicación en el proceso de recuperación de información. Fuente: elaboración propia, 2020.

Del catálogo en línea se analizaron registros que incluyen datos pertenecientes a la colección general de la biblioteca. Los datos seleccionados corresponden a la disciplina de la bibliotecología y los estudios de la información. En la *Tabla 1*, se aprecia la metodología que se empleó para llevar a cabo el procesamiento de los datos que fueron expuestos a la fase sucesiva de recuperación de información. Monteiro y Cabral (2018), Silvello *et al.* (2017) y Jovanovik (2016) han propuesto procedimientos para el sometimiento de datos enlazados a una serie de operaciones programadas y su aplicación en el conjunto de fases de recuperación de la información; se aplicaron a contextos especializados de datos con variables previamente definidas, es decir, se desarrolló el entorno de datos enlazados que permitió recuperarlos y visualizarlos con la lógica de SPARQL y RDF. Para este propósito, se seleccionó el software *GraphDB* debido a su amplia usabilidad y flexibilidad de instalación y manejo, además de tratarse de un software libre que facilita la interoperabilidad y el control de datos de diversa naturaleza. Aunado a ello, se utilizó el software *Open Refine* por su alta funcionalidad para la limpieza y estructuración de datos mediante vocabularios semánticos. Además, el uso de *Zotero* para almacenar y organizar los datos corroboró su vasta interoperabilidad para cosechar datos disponibles en plataformas de bibliotecas.

Análisis de resultados

En la *Tabla 2* se aprecian los datos sobre los que se efectuó el análisis de la recuperación de información con LOD. La tabla fue guardada en formato .csv para efectuar su limpieza correspondiente, que consiste en eliminar puntuaciones, inexactitudes y duplicados. La estructuración de los datos se llevó a cabo de acuerdo con el campo de título, autor, año de publicación, ISBN, editor y lugar de publicación.

Título	Autor	Año	Editor	Lugar
Preserving digital materials in libraries, archives and museums	Martin, Elia	2019	Magnum Publishing	New York
Digital preservation for libraries, archives, and museums	Corrado, Edward M	2019	Rowman & Littlefield	London
XML for catalogers and metadata librarians	Millson Martula, Christopher	2019	Libraries unlimited	Santa Barbara, California

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

Recent developments in the design, construction, and evaluation of digital libraries: case studies	Reese, Terry	2019	Information Scienc Reference	Hershey, Pennsylvania
Robots in academic libraries: advancements in library automation	Angel, Christine	2018	Information Scienc Reference	Hershey, Pennsylvania
Learning from libraries that use WordPress: content-management system best practices and case studies	Cox, Marge	2018	American Library Association	Chicago
Collaboration in libraries and learning environments	Fernandez, Peter	2018	Facet Publishing	London
iPads in the library: using tablet technology to enhance programs for all ages	Jacobs, Brittany	2018	Libraries unlimited	Santa Barbara, California
The survey of the use of tablet computers by academic & special libraries	Krashen, Stephen	2018	Primary Research Group	New York

Tabla 2. Datos en formato csv correspondiente a registros de la colección general de la biblioteca del IIBI. Fuente: elaboración propia, 2020.

Sucesivamente, se llevó a cabo la alineación RDF de los datos. Para ello, se utilizó el vocabulario bibframe 2.0 disponible en *Linked Open Vocabularies* (LOV, sa). "BIBFRAME (Bibliographic Framework) es una iniciativa para evolucionar los estándares de descripción bibliográfica a un modelo de datos vinculados, con el fin de hacer que la información bibliográfica sea más útil tanto dentro como fuera de la comunidad bibliotecaria" (Library of Congress, 2016: § 1).

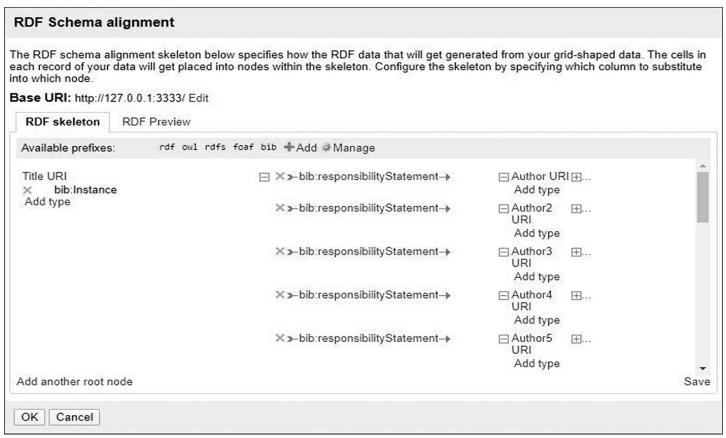

Figura 2. Alineación RDF de los datos mediante BIBFRAME Vocabulary 2.0. Fuente: elaboración propia, 2020.

En la *Figura 2* se advierte el establecimiento de vinculaciones entre los datos, tomando como nodo central la instancia primaria *bib:Instance* que remite a los datos que identifican los títulos de los libros representados en la *Tabla 2*. Para ello, a cada dato le fue asignado un URI que le permitiera establecer conexiones técnicas entre los datos disponibles en el dominio del vocabulario. Cada uno de los datos de la tabla fueron vinculados con una propiedad del vocabulario de BIBFRAME, esto mediante un análisis que sirvió para asignar vinculaciones de manera certera entre los datos.

Número	Sujeto	Predicado	Objeto
1	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:Isbn	http://127.0.0.1:3333/978-1-68095-558-3
2	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:Isbn	http://127.0.0.1:3333/978-1-85604-847-7
3	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:agent	http://127.0.0.1:3333/3g-e-learning
4	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:agent	http://127.0.0.1:3333/facet-publishing
5	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:date	http://127.0.0.1:3333/1997
6	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:date	http://127.0.0.1:3333/2018
7	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:place	http://127.0.0.1:3333/london
8	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:place	http://127.0.0.1:3333/new-york-ny
9	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:responsibilityStatement	http://127.0.0.1:3333/-ford-lyn
10	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:responsibilityStatement	http://127.0.0.1:3333/-haven-kendall
11	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:responsibilityStatement	http://127.0.0.1:3333/norfolk-sherry

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

12	http://127.0.0.1:3333/cloud-computing-for-libraries	bib:responsibilityStatement	http://127.0.0.1:3333/raitt-david-i
13	http://127.0.0.1:3333/cloud-computing-for-libraries	rdf:type	bib:Instance

Tabla 3. Sentencia RDF del título "Cloud Computing for Libraries": http://127.0.0.1:3333/cloud-computing-for-libraries
Fuente: elaboración propia, 2020.

A su vez, en la *Tabla 3* se observa un ejemplo de la representación RDF de los datos que fueron dados de alta en *GraphDB* para realizar las consultas correspondientes con SPARQL. Como se aprecia, la sentencia RDF del título "Cloud Computing for Libraries", se encuentra dividida en tres secciones: sujeto (*subject*), predicado (*predicate*) y objeto (*object*). Es decir, los elementos que permiten definir la tripleta del título e identificar las vinculaciones que tiene asignadas.

Por otra parte, se desarrolló una consulta en SPARQL para identificar y recuperar los datos enlazados pertenecientes a *bib:Instance* y descubrir aquellos con atributos similares. La estructura básica utilizada para ello en SELECT con SPARQL tuvo las siguientes características:

> PREFIX: Declaración de prefijos para abreviar URIs.
> FROM: Definición del conjunto de datos indicando qué grafos RDF se consultaron.
> **SELECT: Selección del resultado, indicando qué información devolverá la consulta.**
> WHERE: Patrón de consulta.

El desarrollo de la consulta quedó definido de la siguiente manera:

> Prefix: onto:<http://www.ontotext.com/>
> From: onto:disable-sameAs
> **Select: s**
> Where: <http://id.loc.gov/ontologies/bibframe/Instance>

Así, la consulta arrojó un total de 965 resultados (véase *Tabla 4*), que pueden consultarse mediante el seguimiento de los URI´s que fueron asignados a cada uno de los datos. La inferencia con SPARQL permite recuperar piezas de datos vinculados semánticamente mediante una sentencia triple conformada por sujeto, predicado y objeto.

De acuerdo con Ali y Qayyum (2019), SELECT es la consulta SPARQL ampliamente utilizada que selecciona todas o algunas de las coincidencias de los datos en forma tabular. En este caso, la inferencia en SELECT ayuda a ubicar todos los títulos que están colocados en el elemento S (sujeto) de las tripletas que contiene el conjunto de datos. Por lo tanto, con la inferencia se pueden deducir cuántos datos representan los títulos recuperados en dicho conjunto.

Por ejemplo, la inferencia en los datos descubre cuáles son aquellos recursos o contenidos informativos con atributos similares, que bien pueden identificarse en un proceso de recuperación, sin conocer la existencia de ellos. Es decir, crea nuevos conocimientos basados en lo existente. La inferencia en el plano semántico de la recuperación de información detecta aquellos datos que, si bien no están representados en un título o tema de un determinado recurso, sí forman parte de su estructura relacional informativa. De esta manera, se observa que la recuperación de información mediante el uso de SPARQL es de carácter integradora, pues no solo muestra los datos referentes a los títulos de los datos que forman parte de *bib:Instance*, si no que busca patrones similares entre los datos para desplegar resultados que se vinculen de manera significativa con la consulta realizada.

La semántica de los datos:

> no sólo añade las definiciones bien hechas y codificadas mediante máquina de vocabularios, conceptos y términos, sino que también explica las interrelaciones entre ellos (y especialmente entre diferentes vocabularios que residen en diferentes documentos o repositorios en la web) en formas declarativas (enunciadas) y condicionales (por ejemplo, formas basadas en reglas o lógicas) (Fox y Hendler, 2009: 162).

Sujeto
http://127.0.0.1:3333/preserving-digital-materials-in-libraries-archives-and-museums
http://127.0.0.1:3333/digital-preservation-for-libraries-archives-and-museums
http://127.0.0.1:3333/xml-for-catalogers-and-metadata-librarians
http://127.0.0.1:3333/recent-developments-in-the-design-construction-and-evaluation-of-digital-libraries-case-studies
http://127.0.0.1:3333/robots-in-academic-libraries-advancements-in-library-automation
http://127.0.0.1:3333/learning-from-libraries-that-use-wordpress-content-management-system-best-practices-and-case-studies
http://127.0.0.1:3333/collaboration-in-libraries-and-learning-environments
http://127.0.0.1:3333/ipads-in-the-library-using-tablet-technology-to-enhance-programs-for-all-ages
http://127.0.0.1:3333/the-survey-of-the-use-of-tablet-computers-by-academic-special-libraries

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

http://127.0.0.1:3333/librarians-guide-to-online-searching
http://127.0.0.1:3333/designing-library-space-for-children
http://127.0.0.1:3333/cloud-computing-for-libraries
http://127.0.0.1:3333/know-it-all-find-it-fast-for-academic-libraries
http://127.0.0.1:3333/lifelong-learning-in-public-libraries-principles-programs-and-people
http://127.0.0.1:3333/libraries-telecentres-cybercafes-and-public-access-to-ict-international-comparisons

Tabla 4. Resultados de la consultada desarrollada con SPARQL.
Fuente: elaboración propia, 2020.

Aunado a ello, SPARQL permite consultar las premisas derivadas de la estructuración de los datos a través de los elementos sujeto, predicado y objeto que son construidas a partir de la ejecución de RDF. En la *Figura 3* se aprecia el grafo RDF derivado de la consulta SPARQL realizada bajo la selección de *bib:Instance*. Se seleccionó el dato referente al título "Cloud Computing for Libraries" para ejemplificar la representación de *Linked Open Data* en un proceso de recuperación de información que resulta visualmente interactivo.

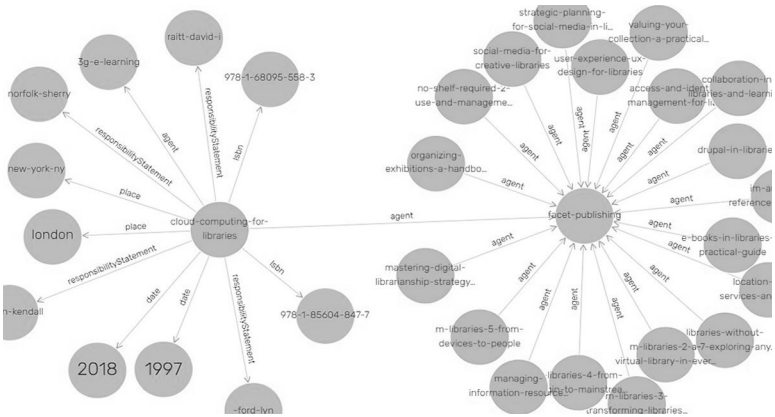

Figura 3. Grafo RDF derivado de la consulta con SPARQL
Fuente: elaboración propia, 2020.

Al seleccionar cualquiera de los nodos del grafo, se ejerce un proceso de interacción intuitiva, la cual fomenta el descubrimiento de datos que en una recuperación textual tradicional pueden pasar desapercibidos. Este modelo de representación logra obtener una RI multifactorial, en la que resulta posible identificar las vinculaciones entre datos disponibles en diversas fuentes, una característica que el modelo relacional no ha abordado en la actualidad.

Además, SPARQL ofrece la posibilidad de realizar inferencias en los procesos de recuperación de información mediante la aplicación de LOD. Por ejemplo, Christodoulou, Paton y Fernandes (2013: 2) señalan que la inferencia aplicada a entornos de datos abiertos enlazados pone en evidencia que, a diferencia de la organización lógica impuesta en bases de datos relacionales, una fuente RDF no se ajusta a ninguna estructura análoga. En el contexto de LOD, un esquema de una fuente RDF constituye una combinación de términos de varios vocabularios que se utilizan para representar los datos, donde la semántica se define en varios vocabularios construidos con el esquema RDF y el lenguaje de ontología web (conocido bajo las siglas en inglés: OWL).

En términos generales, la inferencia en la web semántica se puede caracterizar por descubrir nuevas relaciones, en ella los datos se modelan como un conjunto de conexiones (nombradas mediante predicados) entre recursos. "Inferencia" significa que los procedimientos automáticos generan nuevas relaciones basadas en los datos y en alguna información adicional en forma de vocabulario, por ejemplo, un conjunto de reglas (World Wide Web Consortium, 2015).

De esta manera, la inferencia facilita obtener resultados de recuperación de información a partir de las premisas (sujeto, predicado y objeto) que son construidas con RDF y consultadas con el lenguaje SPARQL. Los tipos más comunes de ésta en dicho contexto son los de subsunción y transitividad.

En la inferencia por subsunción se utiliza un método de razonamiento que permite derivar conocimiento sobre clases enteras de entidades a partir del conocimiento sobre otras, sin necesidad de introducir instancias (Schlegel y Shapiro, 2015: 579). Esto hace posible deducir si un concepto está incluido en otro. Por ejemplo: gatos ⊂ mamíferos ⊂ animales.

Por su parte, en la inferencia por transitividad –mediante un proceso deductivo– se deriva una relación entre elementos que no se han comparado explícitamente antes. Este tipo de inferencia se utiliza a nivel ontológico mediante el uso de *Jena* (un marco semántico de código abierto para Java).

De esta manera si *X* es un hermano de *Y,* y *Y* es el hermano de *Z*, entonces *X* también es el hermano de *Z*. En un nivel ontológico: Juan tiene un hermano Jorge y Jorge tiene un hermano Pablo, entonces vamos a deducir que Juan también tiene un hermano Pablo.

La consulta CONSTRUCT mediante *Jena* quedaría derivada de la siguiente manera (Ali y Qayyum, 2019):

(?x dd:tienehermano?y) + (?y dd:tienehermano?z) → (?xdd: ¿tiene hermano? z)

Entonces, la consulta SPARQL CONSTRUCT sería:

```
CONSTRUCT { ?x dd:TambiénHermanoDe ?z}
 WHERE { ?x dd:hasBrother ?y . ?y dd:tiene hermano ?z}
```

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

En este contexto, el uso de vocabularios semánticos permite representar en un contexto homogéneo los datos que están disponibles en fuentes de datos heterogéneas, causando que el uso de inferencias sea mejor dirigido al descubrimiento de hallazgos basados en lógicas de razonamiento para recuperar información.

Discusión

En los resultados obtenidos pueden observarse las implicaciones de estos aspectos en el despliegue de los datos que están representados en el grafo. Se observa que, a mayor cantidad de datos, es necesario interactuar de manera más intuitiva ejerciendo estrategias de descubrimiento para encontrar patrones de similitud entre los datos. Así, una consulta con SPARQL distribuye los resultados mediante la asignación de elementos de significado que vinculan a los datos en un contexto común.

Estudios previos han planteado la integración de los principios de LOD en el proceso de recuperación de información, sobre todo de SPARQL y RDF. Los más significativos debido a la naturaleza de este trabajo son los realizados por Ichinose *et al.* (2014), Chondrogiannis *et al.* (2015) y Vander Sande *et al.* (2018). En estos, se plantea la implementación de SPARQL y RDF para recuperar datos disponibles en fuentes de diversa naturaleza, sin embargo, ninguno de ellos aborda el comportamiento de la recuperación de datos de índole bibliográfica, pero sí ofrecen aportaciones para construir consultas complejas con SPARQL y la codificación respectiva de los datos con *Resource Description Framework*. Procedimientos que fueron aplicados en este artículo.

Por consiguiente, los alcances de LOD en el proceso de recuperación son:

- Los resultados de la recuperación facilitan consultar a los datos y sus respectivas vinculaciones semánticas.
- Con la navegación en el grafo RDF se descubren datos que no son visiblemente accesibles con una búsqueda simple bajo el método textual.
- La visualización e interacción con el grafo RDF resulta ser de carácter interactiva, es decir, el usuario descubre nuevos datos mediante un proceso intuitivo que es dirigido por la propia lógica del grafo.
- Se trata de una recuperación de información de tipo integradora, ya que a través de una consulta pueden obtenerse datos que forman parte de un mismo contexto y que tienen vinculaciones de significado entre sí.

- La RI con LOD establece inferencias en los datos, dando como resultado la consulta de las vinculaciones entre los triples que forman parte de un determinado conjunto de datos.
- Asimismo, este proceso de RI marca una notable diferencia entre los procesos de recuperación tradicionales y aquellos de orden semántico, pues la RI con LOD faculta la consulta de informaciones a través de bases de datos distribuidas de manera heterogénea, con datos agregados de forma dinámica y con un significado establecido previamente mediante una tarea de análisis.

Por otra parte, las limitaciones identificadas a través de la prueba realizada se exponen de la siguiente manera:

- Es necesario comprender la lógica de SPARQL para efectuar consultas complejas de información.
- Se requiere conocer la naturaleza contextual de los datos que serán sujetos al proceso de recuperación de información, pues ello influye notablemente en los resultados.
- Se necesitan vocabularios interoperables que puedan consultar datos disponibles en diversas fuentes.

El potencial de RDF y SPARQL para recuperar información quedará limitado si no se integran sus principios con un método de visualización gráfica e intuitiva que permita identificar a los datos, sus vinculaciones y patrones ocultos entre ellos. Es preciso señalar que el potencial de LOD recae en la posibilidad de realizar consultas complejas de información en las que se muestran los patrones de interacción entre datos con atributos similares.

Conclusiones

Los principios que intervienen en la recuperación de información mediante *Linked Open Data* están relacionados con la ejecución de un proceso analítico intelectual influenciado por el uso de herramientas tecnológicas y el procesamiento de datos. SPARQL y RDF consituyen dos elementos que permiten realizar consultas complejas de información a través la estructuración semántica de los datos.

Esta clase de consulta se caracteriza por recuperar datos con atributos similares y visualizarlos en una representación gráfica que visibiliza patrones

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

de información que no son evidentes a través de un método de recuperación textual tradicional.

La RI con *Linked Open Data* desarrolla inferencias en los datos a fin de obtener resultados de consultas de carácter intuitivito y descubrir recursos y contenidos informativos con atributos similares. Con la inferencia se puede deducir y conocer de la existencia de vinculaciones que existen entre los datos que han sido codificados con un significado preestablecido a través de un proceso de análisis.

La adaptación de LOD en los sistemas para la RI supone un cambio de paradigma relacionado con el tipo de tecnología empleado en sus estructuras. Por ejemplo, transitar de un modelo conceptual sintáctico a uno de tipo semántico, o utilizar bases de datos enfocadas a grafos en lugar de bases de datos relacionales.

Fomenta la conexión de los datos en lo que concierne a la adaptación de RDF en sus estructuras para realizar consultas de información interna y externa mediante un proceso de hiperconectividad. Además, los sistemas interoperables que funcionen bajo la lógica de LOD deberán habilitar un puerto SPARQL que realice consultas en diversas fuentes de datos con atributos similares, adentro y fuera de su propio contexto, pues este proceso debe ser de carácter universal, de lo contrario, sus principios se reducirían a una plena sistematización de datos con hipervínculos locales.

Referencias

Ali, A. y O. Qayyum. 2019. "Inference New Knowledge Using Sparql Construct Query", en *2nd International Conference on Computing, Mathematics and Engineering Technologies* (iCoMET), 1-4, https://doi.org/10.1109/ICOMET.2019.8673494

Bar-Hillel, Y. y R. Carnap. 1953. "Semantic Information". *The Bristish Journal of the Philosophy of Science* 4 (14): 147-157.

Baron Neto, C., K. Muller, M. Brummer, D. Kontokostas y S. Hellmann. 2016. "LODVader: An Interface to LOD Visualization, Analytics and Discovery in Real-time", en *25th International Conference Companion on World Wide Web*, 11 de abril. Montréal, Québec, Canada. https://doi-org.pbidi.unam.mx:2443/10.1145/2872518.2890545

Berners-Lee, T. 2009. "Linked Data". *World Wide Web Consortium*, 18 de junio. https://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., T. Heath, K. Idehen y T. Berners-Lee. 2008. "Linked Data on the Web". En *WWW '08: Proceedings of the 17th International Conference on World Wide Web*, 21 de abril: 1265–1266.

Bizer, C., M. E. Vidal y H. Skaf-Molli. 2018. "Linked Open Data", en *Encyclopedia of Database Systems*, editada por L. Liu y M. Tamer Özsu. New York: Springer.

Campos, L. M. y M. L. De Almeida. 2014. "Aplicação de dados interligados abertos apoiada por ontologia". *Tendências da Pesquisa Brasileira em Ciência da Informação* 7 (2): 269-288. https://revistas.ancib.org/index.php/tpbci/article/view/316/316

Chen, Chaomei. 2004. *Information visualization: beyond the horizon*. 2nd ed. London: Springer Verlag.

Chondrogiannis, E., V. Andronikou, E. Karanastasis y T. Varvarigou. 2015. "An Advanced Query and Result Rewriting Mechanism for Information Retrieval Purposes from RDF Datasources", en *Knowledge Engineering and Semantic Web*, editado por P. Klinov y D. Mouromtsev, 32-47. Cham: Springer.

Christodoulou, K., N. Paton y A. Fernandes. 2013. "Structure inference for linked data sources using clustering", en *Conference: Proceedings of the Joint EDBT/ICDT Workshops*, 18 de marzo, 60-67. https://doi.org/10.1145/2457317.2457328

Dadzie, A. y E. Pietriga. 2017. "Visualization of linked data reprise". *Semantic web* 8 (1): 1-21.

De Faria Cordeiro, K., F. Firmino de Faria, B. de Oliveira Pereira, A. Freitas, C. Expedito Ribeiro, J. V. Villas Boas Freitas, A. C. Bringuente, et al. 2011. "An approach for managing and semantically enriching the publication of linked open governmental data", en *Proceedings of the 3rd Workshop in Applied Computing for Electronic Government (WCGE)*, 12 de octubre, 82-95.

Dirección General de Bibliotecas y Servicios Digitales de Información (DGBSDI). 2020. LIBRUNAM: Búsqueda básica. Acceso 30 de mayo de 2022. https://librunam.dgb.unam.mx:8443/F/3DFEFXUJUFB7XKSHJKLY883QQP-3VH1IE5FI1NFDAH1VFAQDCXR-19623?func=file&file_name=find-b

Floridi, L. 2016. "Concepciones semánticas de la información", en *Diccionario Interdisciplinar Austral*, editado por C. E. Vanney, I. Silva y J. F. Franck. Argentina: Instituto de Filosofía-Universidad Austral. http://dia.austral.edu.ar/Concepciones_sem%C3%A1nticas_de_la_informaci%C3%B3n

Fox, P. y J. Hendler. 2009. "Semantic e-science: encoding meaning in next generation digitally enhanced science", en *The fourth paradigm: data-intensive scientific Discovery*, editado por T. Hey, S. Tansley y K. Tolle, 147-152. Estados Unidos: Microsoft Research.

Haag, F., S. Lohmann, S. Bold y T. Ertl. 2014. "Visual SPARQL Querying Based on Extended Filter Flow Graphs", en *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, 27 de mayo: 305–312.

Ichinose, S., I. Kobayashi, M. Iwazume y K. Tanaka. 2014. "Ranking the results of DBpedia retrieval with SPARQL query", en *Semantic Technology. JIST 2013, Seoul, South Korea. Revised Selected Papers*, editado por W. Kim, Y. Ding y H. Kim, 306-319. Cham: Springer.

Jovanovich, M. 2016. "Linked data application development methodology". Disertación doctoral. Macedonia: Faculty of Computer Science and Engineering-Ss. Cyril and Methodius University.

Library of Congress. 2016. "Overview of the BIBFRAME 2.0 Model". *Library of Congress*, 21 de abril
https://www.loc.gov/bibframe/docs/bibframe2-model.html

Linked *Open Vocabularies*. (sa) Acceso 30 de mayo de 2022.
https://lov.linkeddata.es/dataset/lov/vocabs.

Linckels, S. y C. Meinel. 2011. "Information Retrieval", en *E-Librarian Service. User-Friendly Semantic Search in Digital Libraryes (X media publishing)*. Berlin: Springer.

Mazza, R. 2009. *Introduction to information visualization*. London: Springer.

Monteiro Cristovão, Henrique y Jorge Cabral Fernandes. (2018). "Information Retrieval in Linked Data: A Model Based on Concept Maps and Complex Networks Analysis". *Transinformação* 30, (2): 193-207.
https://doi.org/10.1590/2318-08892018000200005

Musto, C., G. Semeraro, M. De Gemmis y P. Lops. 2017. "A Hybrid Recommendation Framework Exploiting Linked Open Data and Graph-based Features", en *umap '17: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 375–376. Nueva York: Association for Computing Machinery.
https://doi.org/10.1145/3079628.3079653

Pérez, J., M. Arenas y C. Gutiérrez. 2009. "Semantics and complexity of SPARQL". *ACM Transactions on Database Systems* 34 (3): 1-45.
https://doi.org/10.1145/1567274.1567278

Schlegel, D. y S. Shapiro. 2015. "Inference Graphs: Combining Natural Deduction and Subsumption Inference in a Concurrent Reasoner". *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 1 (29).
https://ojs.aaai.org/index.php/AAAI/article/view/9229

Silvello, G., G. Bordea, N. Ferro, P. Buitelaar y T. Bogers. 2017. "Semantic representation and enrichment of information retrieval experimental data". *International journal on digital libraries* 18: 145-172.
https://doi.org/10.1007/s00799-016-0172-8

Spence, R. 2014. *Information Visualization: An Introduction*. 2a. ed. London: Springer.

Stab, C., D. Burkhartdt, M. Breyer y K. Nazemi. 2013. "Visualizing search results of Linked Open Data", en *Semantic Models for Adaptive Interactive Systems*, editado por T. Hussein, H. Paulheim, S. Lukosch, J. Ziegler y G. Calvary, 133-148. London: Springer.

Vander Sande, M., R. Verborgh, P. Hocstenback y H. Van Sompel. 2018. "Toward sustainable publishing and querying of distributed Linked Data archives". *Journal of Documentation* 1 (74): 195-222.
https://doi.org/10.1108/ JD-03-2017-0040

Waitelonis, J. 2018. "Linked data supported information retrieval". Disertación doctoral. Alemania: Instituto Tecnológico de Karlsruher (KIT).

Wilke, Claus. 2019. *Fundamentals of data visualization*. Estados Unidos de América: O´Reilly.
https://clauswilke.com/dataviz/aesthetic-mapping.html

World Wide Web Consortium. 2015. "Inference". *World Wide Web Consortium*.
https://www.w3.org/standards/semanticweb/inference

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567

Wylot, M. y S. Sakr. 2019. "Native Distributed RDF Systems", en *Encyclopedia of Big Data Technologies*, editada por S. Sakr y A. Zomaya. Cham: Springer.

Wylot, M., P. Cudré-Mauroux, M. Hauswirth y P. Groth. 2017. "Storing, Tracking, and Querying Provenance in Linked Data". *IEEE transactions on knowledge and data engineering* 29 (8): 1751-1764.

Zhang, J. 2008. *Visualization for information retrieval*. Berlin: Springer.

Para citar este texto:

Ávila-Barrientos, Eder. 2022. "Recuperación de información con *Linked Open Data*". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 36 (91): 125-146. http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58567