Recuperación de información con datos abiertos enlazados

Eder Ávila Barrientos

Recuperación de información con datos abiertos enlazados

COLECCIÓN
TECNOLOGÍAS DE LA INFORMACIÓN
Instituto de Investigaciones Bibliotecológicas y de la Información

Recuperación de información con datos abiertos enlazados

Eder Ávila Barrientos

Universidad Nacional Autónoma de México
2022

Z666.73 Ávila Barrientos, Eder.
L56A95 Recuperación de información con datos abiertos enlazados / Eder Ávila Barrientos. - México : UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2022.
xiii, 162 p. - (Tecnologías de la información)
ISBN: 978-607-30-6348-7

1. Datos vinculados.
2. Recuperación de información - Modelos teóricos.
3. Web semántica. I. Título. II. ser.

Diseño de la portada: Wendy Chávez
Primera edición: Agosto de 2022

Instituto de Investigaciones Bibliotecológicas y de la Información
Circuito Interior s/n, Torre II de Humanidades,
pisos 11, 12 y 13, Ciudad Universitaria, C. P. 04510,
Alcaldía Coyoacán, Ciudad de México

ISBN: 978-607-30-6348-7

Publicación dictaminada

Impreso y hecho en México

Contenido

Introducción

En la literatura especializada en Bibliotecología y Estudios de la Información, se identifican hallazgos teóricos que proponen el estudio de los datos abiertos enlazados y su presencia en el contexto de la recuperación de información. En estos trabajos se hace énfasis en la necesidad de adaptar las normas bibliotecarias a los principios de dichos datos.

Bajo esta premisa, los estudios efectuados por Tillet (1987) y Taylor (1999) ponen de manifiesto los principales fundamentos teóricos del establecimiento de relaciones entre los datos bibliográficos que están disponibles en los recursos de información de las bibliotecas. No obstante, estos trabajos carecen de una fundamentación apegada al contexto digital actual que permita validar su sistematización y aplicación pragmática.

En este sentido, es necesario formular un aparato teórico que sustente la sistematización de los datos abiertos enlazados para aprovechar los datos abiertos que están disponibles en los medios y recursos digitales con el propósito de optimizar su recuperación. Por lo tanto, esta obra se propone responder tres preguntas elementales:

- ¿Cuáles son los componentes que constituyen un modelo teórico que sustente la recuperación de información mediante datos abiertos enlazados?

- ¿Qué patrones de comportamiento manifiestan los datos abiertos enlazados en el proceso de recuperación de información?
- ¿Cuáles son los factores que intervienen en la sistematización y recuperación de información mediante el uso de datos abiertos enlazados?

De esta manera, la obra se encuentra divida en tres capítulos elementales. El primer capítulo, denominado "Datos abiertos enlazados", aborda los principios teóricos, conceptos, la normatividad y los desarrollos destacados de los datos abiertos enlazados; en este capítulo se ofrece un panorama general de los atributos y componentes de los datos en cuestión.

El segundo capítulo, titulado "Recuperación de información", aborda la conceptualización que forma parte de este proceso; se presentan los tipos de recuperación de información existentes y se analizan las características de los sistemas para la recuperación de información. Además se lleva a cabo un análisis de la interacción existente entre los datos abiertos enlazados y la recuperación de información.

En el tercer capítulo de esta obra, denominado "Modelo para la recuperación de información con datos abiertos enlazados", se presentan los elementos que forman parte de su fundamentación y estructura. También se aborda la metodología utilizada para el procesamiento de los datos ocupados en esta investigación. Mención especial requiere el análisis de las consultas SPARQL presentadas en este capítulo, pues son la base para comprender la manera de recuperar información mediante los datos abiertos enlazados. Finalmente se presentan los principios que abordan la sistematización e integración de los componentes mencionados en el modelo.

El objetivo de esta obra consiste en presentar la formulación de un modelo teórico para la recuperación de información con datos abiertos enlazados que permita la sistematización, vinculación y accesibilidad de los datos disponibles en los medios y recursos presentes en el entorno digital.

Bajo esta premisa, si se desarrolla un modelo teórico que sustente la sistematización de los datos abiertos enlazados de índole

bibliográfica, se generarán aportes para la óptima recuperación de los datos bibliográficos que están disponibles en los diversos medios y recursos del entorno digital.

En este sentido, la web semántica es un entorno digital en donde conviven datos estructurados que han sido codificados bajo normas y estándares que les permiten vincularse significativamente entre sí. Berners-Lee, Hendler y Lassila (2001, 2) consideran que "[...] la web semántica no es una web separada, sino una extensión de la web actual, en la que la información recibe un significado bien definido".

Los elementos tecnológicos principales de la web semántica son la norma Resource Description Framework, el lenguaje de marcado extensible (XML) y el identificador de recurso uniforme (URI). Estos elementos hacen posible la construcción de estructuras de datos capaces de vincularse semánticamente mediante el análisis de sus atributos. Russo (2015, 38) agrega al respecto que "[...] la web semántica es un sistema informático que permite detectar el contenido de los datos. Consiste en un sistema de metadatos que se basa en Resource Description Framework a través del cual se puede crear un sistema único de recolección de datos utilizando URIs".

El término *datos enlazados* se refiere a un "[...] conjunto de las mejores prácticas para publicar y conectar datos estructurados en la Web usando estándares internacionales del World Wide Web Consortium (por su acrónimo W3C)" (Wood *et al.* 2016, 4). Para que los datos enlazados alcancen un potencial que impacte en el desarrollo de la web semántica, necesitan ser liberados bajo una licencia de acceso abierto. Los datos abiertos enlazados son estructuras de datos codificadas con los principios del W3C y liberados bajo el uso de una licencia de acceso abierto. Estas características les permiten vincularse semánticamente con otras fuentes de datos mediante un proceso interoperable.

En la web semántica proliferan enormes cantidades de datos, los cuales al momento de ser organizados se convierten en conjuntos estructurados de datos. Los datos disponibles en la web semántica son de diversa tipología y temática. Así, los datos que se

encuentran en las bibliotecas tienen el potencial de integrarse en la web semántica.

Los datos de las bibliotecas son de tipo bibliográfico, de autoridad y temáticos. Estos datos se encuentran registrados en formatos especializados como MARC21 y están disponibles en los catálogos de las bibliotecas. Estos datos remiten a una amplia gama de recursos de información documental.

Por lo tanto, los datos abiertos enlazados de bibliotecas son estructuras semánticas de datos que forman parte de los recursos de información documental. La vinculación entre estos datos puede ser bibliográfica, de autoridad y temática. Además, estos datos deben sujetarse a una política de acceso abierto que permita su libre consulta y acceso en el entorno web.

> El uso de tecnologías semánticas y los principios de Linked Data proporcionan características potenciales para la descripción de recursos bibliográficos, tales como normalización semántica de elementos bibliográficos y etiquetado de recursos web con el uso de vocabularios estandarizados (Cabrera *et al.* 2018, 14).

Se estima que los modelos conceptuales pueden contribuir a la conformación de los datos abiertos enlazados en bibliotecas. Estos modelos proporcionan estructuras de datos que son similares a las utilizadas en los principios de Linked Open Data. Sin embargo, es necesario estudiar la interoperabilidad de las estructuras con las normas y estándares bibliotecarios.

El informe final del grupo incubadora de datos enlazados del W3C (2011) identificó las siguientes áreas de investigación que permitirán el desarrollo de datos enlazados en las bibliotecas:

- Área de preparación de los datos. Que se enfoca a la creación de herramientas que permitan transformar, almacenar y vincular los datos de las bibliotecas.
- Área de definición de normas. Que se encarga de abordar la construcción de normas que permitan controlar y uniformar el proceso de creación de datos enlazados.

- Área de desarrollo de interfaces. Encargada del diseño de interfaces de búsqueda y recuperación de información. La interfaz de cualquier sistema de información digital es de suma relevancia para que el usuario remoto pueda tener acceso a la información en el entorno digital.
  El campo del desarrollo de la web semántica y su interacción con la biblioteca es creciente y está abierto a futuras investigaciones. Si bien se han realizado algunos trabajos preliminares para convertir los registros MARC a RDF y crear proyectos prototipo de muestra, aún queda la oportunidad de seguir estudiando y creando prototipos (Lapolla 2013, 135).

El estudio teórico de los datos enlazados y su interacción con las bibliotecas reafirmarán los futuros prototipos metodológicos para la implementación de los principios de Linked Open Data en las bibliotecas, pues no se trata solo de sistematizar los datos y vincularlos con otras fuentes disponibles de la web. La visión de los datos abiertos enlazados y su aplicación en las bibliotecas responde a la necesidad de identificar el comportamiento de los datos en el entorno digital y su latente utilización para descubrir vinculaciones de conocimiento que han sido plasmadas en los recursos de información documental.

La recuperación de la información es un proceso fundamentado en etapas específicas. En este proceso se ponen de manifiesto aspectos como la organización, la búsqueda y el acceso a la información contenida en los recursos. Los sistemas para la búsqueda y recuperación de la información utilizan estrategias y métodos que permiten acceder a los datos y recursos que están representados en un determinado contexto.

Para Lancaster y Warner (1993, 11), recuperación de información, tal y como se utiliza habitualmente, es sinónimo de búsqueda de literatura; es el proceso de buscar en una colección de documentos (utilizando el término *documento* en su más amplio sentido) para identificar aquellos que tratan de un determinado tema.

La figura de los datos en el proceso de recuperación de información resulta significativa, pues los datos representan los atributos

bibliográficos y temáticos de los recursos que son almacenados en un determinado sistema y que pertenecen a un contexto previamente definido.

Ungvarsky (2017) es más puntual y plantea que la recuperación de la información es el proceso de búsqueda para localizar, identificar, refinar y presentar información relevante sobre un tema en particular. La relevancia de la información recuperada dependerá de la exactitud de la representación de los datos que son registrados en las herramientas para la búsqueda y el acceso a la información.

Entonces, la recuperación de la información es un proceso que nace a partir de la definición de una necesidad informativa. En la actualidad es posible recuperar información utilizando diversas herramientas como catálogos en línea, repositorios institucionales, bases de datos, buscadores web, descubridores de información y directorios.

La integración de los datos abiertos enlazados en la recuperación de información no consiste solamente en una estrategia para obtener información de manera eficaz y eficiente. Se trata de implementar un método que permita organizar, buscar, recuperar, acceder y visualizar las vinculaciones entre los datos que forman parte de los recursos y que están disponibles en los medios digitales de la web.

Por lo tanto, se requiere de investigaciones teóricas que sustenten la interacción de los datos abiertos enlazados bibliográficos en el proceso de recuperación de información. Pues las investigaciones actuales versan principalmente en la labor pragmática de la integración de ambos elementos.

La presente obra se enmarca en un contexto en donde los modelos de datos y conceptuales intentan alcanzar la interoperabilidad de los datos disponibles en las bibliotecas con otras fuentes disponibles en el entorno de la web. Ejemplos como BIBFRAME y LRM de IFLA así lo constatan. Si bien se pueden consultar algunas investigaciones que intentan abordar la aplicación de los datos enlazados en las bibliotecas, son escasas aquellas que formulen modelos teóricos que permitan sentar las bases metodológicas para un modelo funcional y aplicable para las unidades informativas.

Esta obra aspira a contribuir a la teoría de los datos abiertos enlazados para sugerir la creación de modelos metodológicos que sean susceptibles de aplicarse en las bibliotecas. Se considera que los datos abiertos enlazados pueden contribuir a la identificación de patrones que expliquen el comportamiento de la información que es plasmada en los recursos de información documental, lo que permitirá descubrir datos que se encuentren ocultos en el universo de información disponible en la actualidad.

Eder Ávila Barrientos

Datos abiertos enlazados

Principios teóricos

Los datos abiertos enlazados son una propuesta que emana del argumento de establecer un entorno digital mejor organizado mediante la publicación y conexión de datos a través de un vínculo semántico entre ellos. Es decir, fomentar una web de mayor significado que contribuya a contar con un entorno digital armónico y mejor representado.

Desde hace más de una década, los datos abiertos enlazados han sido motivo de estudio en diferentes campos disciplinarios. Los más significativos debido a sus postulados y generación de iniciativas son las ciencias de la computación, la informática y recientemente en la bibliotecología y los estudios de la información.

Los principios teóricos de los datos abiertos enlazados aparecieron por primera vez en un trabajo publicado por Bizer, Heath y Berners-Lee (2009, 2) en el cual manifiestan que el uso de estos datos en el ambiente web permitirá crear enlaces escritos entre datos de diferentes fuentes. Estos datos pueden ser tan diversos como bases de datos mantenidas por dos organizaciones en diferentes ubicaciones geográficas, o simplemente sistemas heterogéneos dentro de una organización que, históricamente, no ha interoperado fácilmente a nivel de datos.

Técnicamente, los datos enlazados se refieren a los datos publicados en la web de tal manera que son legibles por máquina; su significado está definido explícitamente; está vinculado a otros conjuntos de datos externos y, a su vez, pueden vincularse desde conjuntos de datos externos. Si los datos se vinculan semánticamente, se otorga la posibilidad de conocer patrones de comportamiento entre datos disponibles en diversas fuentes del ambiente digital.

En los últimos años ha cobrado especial interés el concepto de web semántica o web de los datos, lo que constituye una evolución natural de la web tradicional. La web semántica no constituye una web totalmente nueva, sino que constituye una extensión de la web tradicional en la que los datos poseen un significado comprensible por sistemas informáticos (Hidalgo *et al.* 2013, 78). En este sentido, los sistemas para la recuperación de información también han recibido una notable influencia de los cambios de paradigma asociados a la aplicación de tecnologías semánticas en sus estructuras. Desde la adopción de nuevos esquemas para organizar los datos, hasta la posibilidad de incursionar en gestores de bases de datos más sofisticados.

De esta forma, los datos abiertos enlazados no son sólo un factor determinante para aprovechar los datos que están disponibles en el ambiente digital, sino que representan todo un conglomerado de normas, estándares y aplicaciones que hacen posible la interacción de los datos que están disponibles en diversas fuentes. Por lo tanto, la evolución de los datos abiertos enlazados se encuentra muy influenciada por el progreso de las aplicaciones, estándares y normas que se utilizan para la conformación de un ambiente digital de datos con un significado previamente establecido.

Al respecto, Méndez y Greenberg (2012, 238) realizaron un breve recorrido relacionado con los avances que han permitido la conformación de LOD de 1996 al 2012.

En la figura 1 puede apreciarse la conformación de estándares como la norma Resource Description Framework (RDF) que permite la generación de estructuras semánticas entre los datos. RDF, conocido en español como el Marco de Descripción de Recursos,

fue diseñado originalmente como un modelo de datos para el desarrollo de metadatos.

La idea principal de RDF es representar "objetos" mediante URIs. La información se proporciona mediante declaraciones sobre los objetos y las declaraciones se expresan como triples. Estos triples consisten en un sujeto, un predicado y un objeto, y expresan que el sujeto está en cierta relación (identificado por el predicado) con el objeto (Gottron y Staab 2018, 2036).

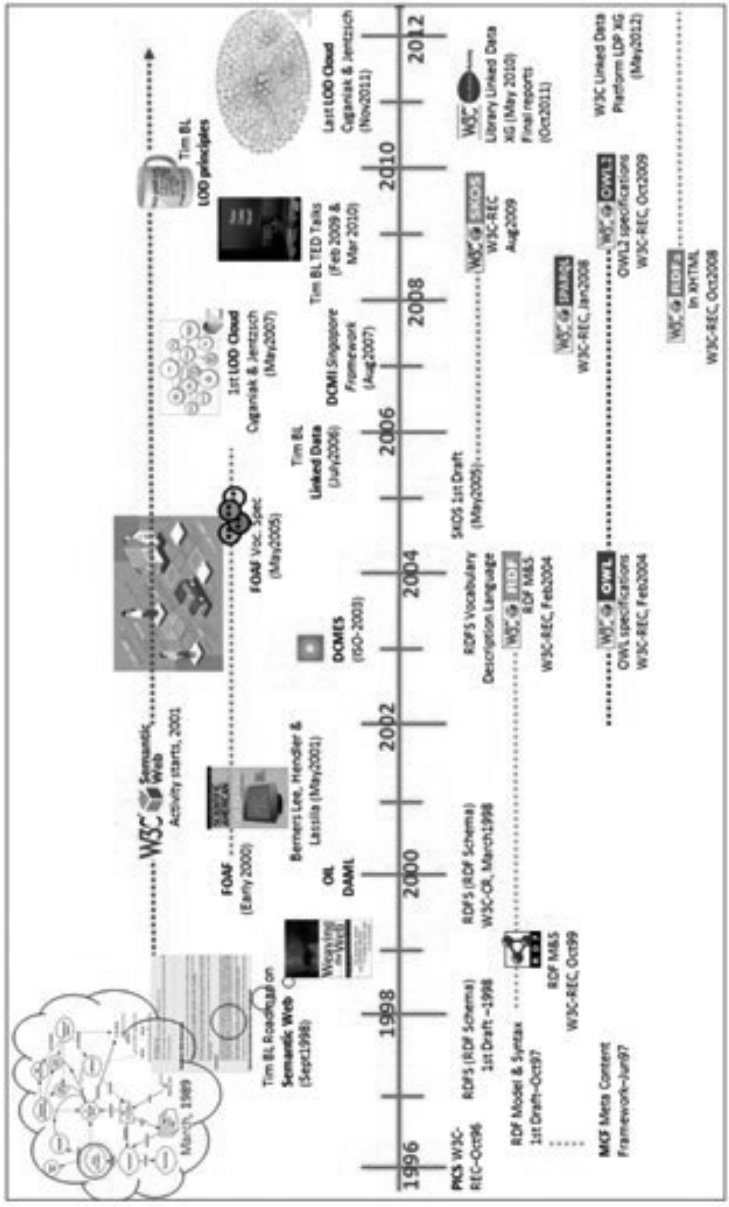
Las relaciones entre los objetos se pueden interpretar y representar en un formato de grafo, donde los sujetos y los objetos son los nodos del grafo y el predicado es un borde etiquetado entre los nodos. Los grafos RDF son representaciones de datos que pueden ser elementales o complejas dependiendo del tipo y la cantidad de datos que se desea representar.

La complejidad en el comportamiento de la información ha dado la pauta para generar representaciones que permitan interpretar los datos que forman parte de un dominio concreto. No obstante, los datos abiertos enlazados han sido propuestos para conectar e interpretar a los datos que forman parte de diversos dominios.

Esta singularidad puede apreciarse en la vinculación semántica de los datos. En este sentido, el análisis visual de datos, facilitado por interfaces interactivas, permite la detección y validación de resultados esperados y a la vez descubrimientos inesperados en la ciencia; permite la validación de nuevos modelos teóricos y ofrece una comparación entre modelos y conjuntos de datos; permite la consulta cuantitativa y cualitativa, mejora la interpretación de datos y facilita la toma de decisiones (Hansen *et al.* 2009, 165).

Actualmente, los vocabularios y las ontologías juegan un papel crucial en la construcción de datos abiertos enlazados, pues son elementos que permiten construir el aditivo semántico que los datos requieren para poder conectarse de una manera significativa. En este sentido, cada dato que forma parte del ambiente digital tiene un contexto y atributos que son trascendentales para construir su significado.

Figura 1. Evolución de la web semántica en datos vinculados



Fuente: Méndez 2012. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthJhneGhM2HcjC5Y52Q?e=avXmbG.

De esta manera, los datos abiertos enlazados reúnen una serie de principios que permiten definir su naturaleza y aplicación en el entorno digital. Dichos principios contemplan lo siguiente:

- Cada dato es unívoco e irrepetible.
- Los datos representan atributos de objetos disponibles en el ambiente digital.
- La similitud de atributos entre los datos permite construir vinculaciones de significado entre ellos.
- El aditivo semántico de los datos permite realizar consultas complejas de información.
- Las consultas complejas de información permiten analizar comportamientos dinámicos que manifiestan los datos.
- Los datos abiertos enlazados fomentan el descubrimiento y la recuperación de piezas de información complejas por naturaleza.

Además de esto, los datos abiertos enlazados utilizan licencias abiertas que fomentan su uso libre de restricciones económicas, legales y técnicas. De acuerdo con Bizer, Vidal y Skaf-Molli (2018, 2097), la idea detrás de estos principios es, por un lado, utilizar estándares para la representación y el acceso a los datos en la web. Por otro, los principios se propagan para establecer hipervínculos entre datos de diferentes fuentes. Estos hipervínculos conectan a todos los datos vinculados en un solo grafo de datos globales, similar a los hipervínculos en la web clásica que conectan a todos los documentos HTML en un solo espacio de información digital.

Una representación gráfica que ayuda a comprender de mejor manera estos principios la podemos obtener a partir de la revisión de la nube de datos abiertos enlazados, también conocida como "The Linked Open Data Cloud" (https://lod-cloud.net/). Actualmente, esta nube contiene 1255 conjuntos de datos con 16174 enlaces. Los tipos de datos que pueden localizarse en esta representación son de índole geográfica, gubernamental, disciplinaria (ciencias de la vida, lingüística), multimedia, de publicaciones, generados en redes sociales y por usuarios de diferentes dominios.

En el grafo de publicaciones (véase figura 2) pueden observarse las conexiones entre datos de tipo bibliográfico correspondientes a proyectos de datos enlazados desarrollados en bibliotecas, editoriales, museos y archivos que han liberado y conectado una amplia variedad de datos que forman parte de un dominio documental.

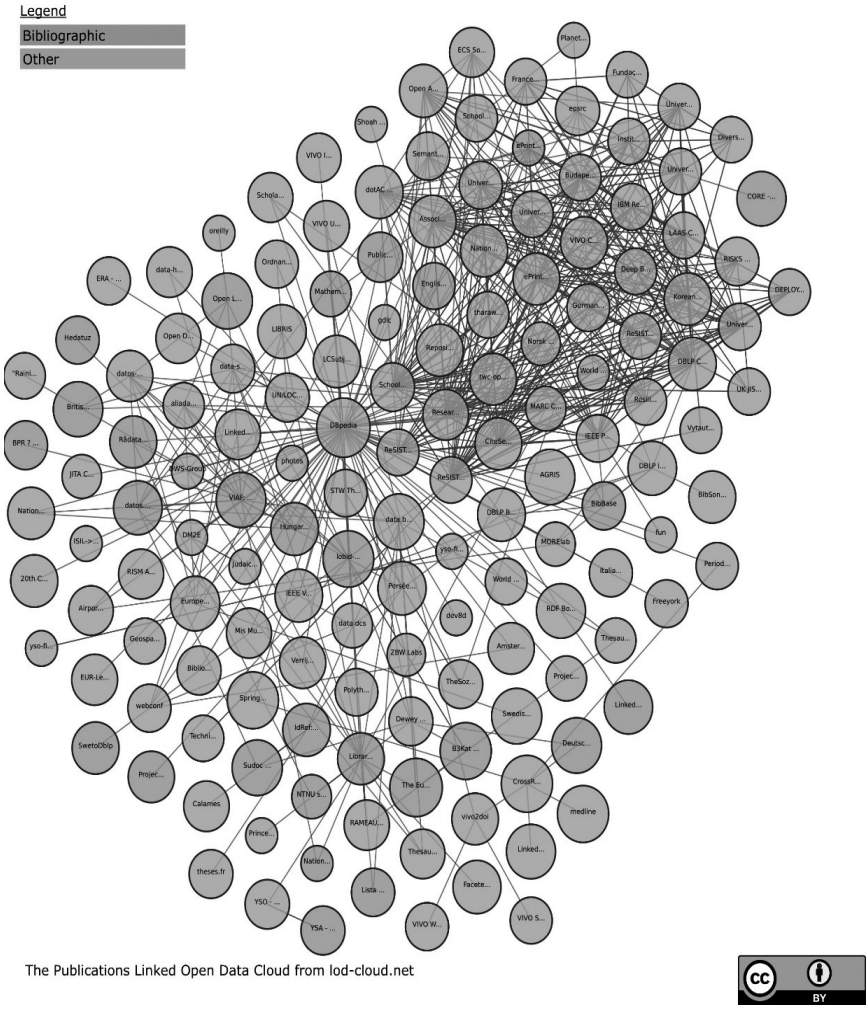
En este grafo se pueden apreciar proyectos como listas de encabezamientos de materia, vocabularios y metadatos que han sido liberados en formatos RDF con la capacidad de conectarse entre sí mediante atributos similares en sus estructuras. La gran mayoría de estos datos son compatibles con la iniciativa del sistema simple para la organización del conocimiento (conocido como SKOS por sus siglas en inglés). SKOS es un desarrollo del World Wide Web Consortium en forma de aplicación de RDF que proporciona un modelo para representar la estructura básica y el contenido de esquemas conceptuales como listas encabezamientos de materia, taxonomías, esquemas de clasificación, tesauros y cualquier tipo de vocabulario controlado (Sánchez 2020).

Si bien en la actualidad la implementación de SKOS en los sistemas para la recuperación de información sigue siendo un reto, existen desarrollos que ponen de manifiesto los esfuerzos por alcanzar la conformación de un ecosistema interoperable de datos que permita representar de una mejor manera los datos que representan diversos contextos disciplinarios y de conocimiento.

Las bibliotecas nacionales de diversas partes del mundo han optado por generar proyectos relacionados con la integración de los principios de LOD en sus estructuras. Sin embargo, la falta de consenso en la manera de tratar los datos de índole documental, ha retrasado la adopción de modelos y estructuras uniformes que permitan consolidar un entorno interoperable de datos enlazados en bibliotecas.

La interoperabilidad global de los datos es un tema que cobra vital relevancia al momento de concebir un contexto de datos con atributos multifactoriales capaces de conectarse mediante vinculaciones de atributos similares existentes entre sí. Dicha interoperabilidad no es sólo un aspecto técnico de la sistematización de los datos en diferentes ubicaciones digitales, ya que establece la

Figura 2. Grafo de publicaciones disponible en The Linked Open Data Cloud Diagrama



Fuente: https://lod-cloud.net/clouds/publications-lod.svg. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthJhkX6MkLWIOTBe7Bg?e=kyLXvl.

capacidad de múltiples sistemas para funcionar de manera homogénea respecto al flujo de datos e informaciones que forman parte de su contexto.

Las problemáticas de la interoperabilidad, también llamadas heterogeneidades, se pueden distiguir de la siguiente manera (Haslhofer y Neuhold 2011, 4):

- Heterogeneidades técnicas: denota todas las diferencias de protocolo de intercambio y plataforma del sistema que impiden que las aplicaciones envíen y reciban objetos de información.
- Heterogeneidades estructurales y sintácticas: ocurren cuando las unidades de datos en los objetos de información se representan utilizando diferentes estructuras y convenciones sintácticas.
- Heterogeneidades semánticas: son conflictos que ocurren debido a las diferencias en la semántica de las unidades de datos.

Las problemáticas de interoperabiblidad que aquejan a los datos han dado la pauta para la aparición de métodos informáticos que ayuden a resolver sus inconsistencias y fomentar una mayor calidad de los datos. Por ejemplo, la técnica de comprensión basada en RDF ayuda a eliminar espectros de redundancia que se presentan en los datos y permiten estructurarlos de una mejor manera. En este sentido, la redundancia simbólica se debe a la repetición de símbolos y signos que impiden una mejor codificación del significado de los datos. Los principales contribuyentes de esta importante fuente de redundancia son los URI grandes y altamente repetitivos que se utilizan para nombrar a los datos en un determinado sistema.

En la práctica, un conjunto de datos RDF comprende muchos URI diferentes, pero generalmente se definen a partir de un pequeño grupo de dominios y tienden a tener prefijos largos comunes. Por otra parte, la redundancia sintáctica se refiere a la repetición innecesaria de palabras o conceptos que están ya expresados con

otras palabras o que se sobreentienden; aunque en numerosos casos es una estrategia ampliamente usada para evitar malentendidos o errores de decodificación entre los datos.

A su vez, la redundancia semántica, a diferencia de las anteriores, aparece en el nivel lógico de la implementación de los datos. Surge cuando se pueden usar menos datos para proporcionar y explicar el mismo conocimiento sobre ellos. Por lo tanto, la redundancia no depende de cómo se codifiquen los datos, sino del conocimiento que aportan. Esta redundancia no se puede eliminar utilizando enfoques de compresión tradicionales. En este caso, los compresores RDF específicos deben diseñarse desde cero para obtener el subconjunto mínimo de "triples canónicos" que permitan representar eficazmente el conocimiento original de la fuente donde fueron obtenidos los datos.

De esta manera, los datos son una entidad que puede ser estudiada a partir de sus atributos y capacidades para interarctuar entre sí y con otros dominios semejantes. Pues, a mayor cantidad de datos en un contexto específico, será muy favorable contar con estrategias y mecanismos que contribuyan a mejorar su organización y capacidad de interconectarse.

Una idea importante que surge sobre el tratamiento de los datos heterogéneos nos remite a conocer lo que los datos "significan"; de esta manera será más fácil utilizarlos. Conforme crecen el volumen, la complejidad y la heterogeneidad de los datos, los científicos necesitan cada vez nuevas competencias basadas en recientes enfoques "semánticos"; por ejemplo, en la forma de ontologías, codificaciones de términos, conceptos y relaciones entre ellos mediante la interacción del usuario con los datos (Fox y Hendler 2014, 160).

La interacción entre los datos y los usuarios pone de manifiesto un comportamiento por parte del usuario al momento de utilizarlos para resolver su demanda informativa. Este comportamiento está caracterizado por el uso de sistemas de información que permiten acceder a los datos a través de consultas mediante una interfaz.

En la figura 3 puede apreciarse el proceso de interacción que se ejerce entre los usuarios y los datos en un sistema de información

común. En este sentido, la demanda informativa del usuario es la que motiva el desarrollo de una determinada consulta en el sistema de información.

Entonces, cuando se ejerce esta búsqueda en la base de DAE, se recuperan datos que se relacionan con diferentes contenidos y recursos disponibles en el universo de información. La creación de aplicaciones para recuperar información que satisfaga eficazmente las necesidades de información de los usuarios suele ser problemática por diversas razones.

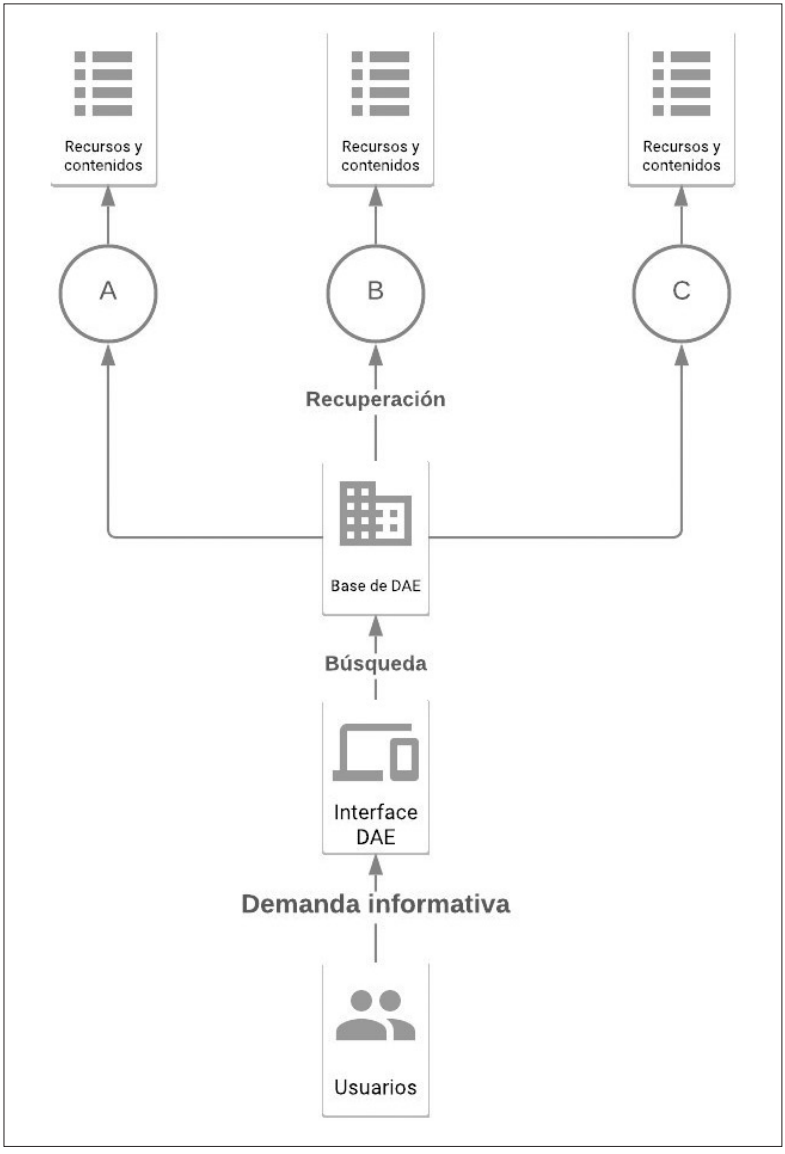
De acuerdo con Peters, Braschler y Clough (2011, 89), las personas pueden tener dificultades para articular sus necesidades y traducirlas en una representación adecuada para un sistema de búsqueda específico; pueden tener dificultades para encontrar los términos de consulta adecuados; pueden sentirse abrumados con demasiados resultados de búsqueda; puede que no obtengan suficientes resultados o resultados en absoluto (cero aciertos), y pueden tener dificultades para interpretar listas de resultados desorganizadas o con el uso de sintaxis de consulta especializada.

En una interfaz para recuperar información mediante DAE, será necesario establecer con claridad las variables de búsqueda del usuario, que le permitan trasladar su necesidad informativa en el sistema que almacena a los datos. De esta manera, la recuperación de la información con DAE se ve fuertemente complementada por los métodos de visualización de información que permiten consultar grandes cantidades de datos acompañados de sus respectivas vinculaciones.

Actualmente, los motores de búsqueda web y los hipervínculos son las formas básicas y de uso común que el usuario utiliza para buscar cosas en la web y navegar por el contenido, pero no permiten un análisis detallado de las vinculaciones entre datos, recursos y contenidos. Para mejorar esto, una cuestión clave es la necesidad de estandarización y su uso interoperable, además de utilizar las herramientas que puedan respaldar la integridad de los datos en diferentes dominios.

Los datos abiertos enlazados al momento de sistematizarse deberán fomentar la interoperabilidad global de los datos, haciéndolos compatibles con los recursos y contenidos que contengan

Figura 3. Proceso de interacción entre los datos y usuarios



Fuente: elaboración propia 2021.

atributos similares en sus estructuras. Además, un sistema de estas características deberá integrar vocabularios y ontologías para representar conceptos y atributos entre los conjuntos de datos disponibles en diversas fuentes.

Los datos están en el centro de los desarrollos actuales en la web semántica, donde se encuentran diferentes tipos de datos que son publicados, intercambiados, descritos y consultados todos los días. En particular, existe un esfuerzo creciente por publicar y utilizar una serie de repositorios, catálogos y registros para respaldar el descubrimiento y la reutilización de datos de la web semántica (Morato *et al.* 2014). A medida que los datos estructurados se vinculan de manera más enriquecedora, el usuario puede notar capacidades mejoradas para descubrir y usar datos. De esta manera, los datos abiertos enlazados no se utilizan para crear una web diferente, sino para mejorar la web mediante la adición de datos estructurados.

Estos datos estructurados, expresados mediante tecnologías como RDF en atributos (RDFa) y microdatos, desempeñan un papel muy importante en los algoritmos de rastreo y diversas técnicas presentes en los motores de búsqueda, ya que proporcionarán una forma para que los datos mejoren su visibilidad a través de la optimización de las consultas efectuadas por el usuario.

En este sentido, los datos estructurados incrustados en las páginas HTML también facilitarán la reutilización de los datos en servicios especializados para los buscadores de información: por ejemplo, la gestión de citas se puede hacer tan simple como cortar y pegar URI. De esta manera, los datos enlazados favorecerán la investigación interdisciplinaria al enriquecer el conocimiento mediante la vinculación entre múltiples bases de conocimiento específicas de dominio.

Al publicar datos en la Web de acuerdo con los principios de datos vinculados, los proveedores de datos agregan sus datos a un espacio de datos global, lo que permite que sean descubiertos y utilizados por varias aplicaciones.

La publicación de un conjunto de datos como datos vinculados en la web implica las siguientes etapas (Bizer *et al.* 2009):

1. Asignar URIs a las entidades descritas por el conjunto de datos y proporcionar la desreferenciación de estos URIs a través del protocolo HTTP en representaciones RDF.
2. Establecer vínculos RDF a otras fuentes de datos en la web, de modo que los usuarios puedan navegar por la web de datos en su totalidad siguiendo los vínculos RDF.
3. Proporcionar metadatos sobre los datos publicados para que los usuarios puedan evaluar la calidad de los datos publicados y elegir entre diferentes medios de acceso.

De esta manera, los datos abiertos enlazados utilizan el modelo de datos Resource Description Framework y otros estándares relacionados con RDF, al igual que HTTP. Los datos enlazados se basan en RDF pero no son lo mismo que RDF. Los datos abiertos enlazados siguen los siguientes cuatro principios:

- Utilizar URIs como nombres para las cosas.
- Utilizar HTTP para que los usuarios puedan buscar esos nombres.
- Cuando alguien busque un URI, proporcionar información útil utilizando los estándares RDF y SPARQL.
- Incluir enlaces a otros URI para que el usuario pueda descubrir más cosas.

Estos principios están fuertemente influenciados por los elementos que permiten la consulta de información en el ambiente web, pues DAE es un componente extensible de la propia World Wide Web, y ha sido tomado en cuenta como uno de los principales estandartes para desarrollar una web con mayor significado.

Concepto

Actualmente no existe un consenso que permita identificar con claridad la conceptualización de DAE. Esto es debido en parte a la

gran complejidad que existe para identificar epistemológicamente la naturaleza de los datos en un contexto unificador, pues la presencia de los datos en múltiples contextos dificulta la generación de una definición integradora para englobar todos los atributos que tienen presencia en los datos abiertos enlazados.

Como parte de la revisión de la literatura, ha sido posible identificar los elementos que caracterizan a DAE. Para ello, primeramente debe comprenderse el concepto de datos abiertos que tiene una fuerte influencia sobre DAE. De acuerdo con Davies y Calderón (2020, 1):

> [...] los datos abiertos se pueden describir en términos de tres propiedades clave: datos que son accesibles (generalmente en línea), datos que son legibles por máquina (en formatos de datos estructurados que se pueden explorar sin software propietario) y datos que están disponibles para una amplia reutilización (generalmente señalado por declaraciones de licencias abiertas).

Tomando como base la noción de libre apertura y uso de los datos, Tim Berners Lee publicó una serie de buenas prácticas para el desarrollo y la publicación de datos abiertos enlazados en el ambiente de la web. Estas recomendaciones se conocen como las cinco estrellas de Linked Open Data y expresan lo siguiente (Berners Lee 2009):

> * Publicar los datos en la web (bajo cualquier formato) y utilizar una licencia abierta.
> ** Publicarlos como datos estructurados, por ejemplo en Excel, en lugar de una tabla escaneada.
> *** Utilizar formatos no propietarios para codificarlos, por ejemplo, .csv en lugar de .xls.
> **** Utilizar RDF y URIs para nombrar cosas, así los usuarios podrán encontrarlas.
> ***** Enlazar tus datos con datos de otros usuarios para proporcionar contexto.

En la figura 4 se aprecia el sistema acumulativo de cinco estrellas de DAE, que se ha explicado con anterioridad. Cada estrella adicional supone que sus datos cumplen con los criterios de los pasos anteriores. El Consorcio World Wide Web (W3C) define los estándares para la Web, incluido un modelo de datos abiertos y varios formatos para ese modelo.

A medida que evolucionan las aplicaciones basadas en Internet, se está desarrollando e implementando una nueva gama de especificaciones. La idea es que los usuarios de la web puedan desarrollar una nueva funcionalidad 'semántica' con la misma facilidad o de manera similar al uso y la evolución de HTML, como la base de la WWW. Es decir, dotar a los datos de la capacidad de vincularse abiertamente, tal y como sucede con los documentos representados en lenguajes de marcado.

La idea básica de los datos abiertos enlazados es aplicar la arquitectura general de la World Wide Web a la tarea de compartir datos estructurados a escala global. Similar a la Web de documentos clásica, la Web de datos abiertos enlazados se basa en un pequeño conjunto de estándares y la idea de utilizar enlaces para conectar contenido de diferentes fuentes (Bizer *et al.* 2018).

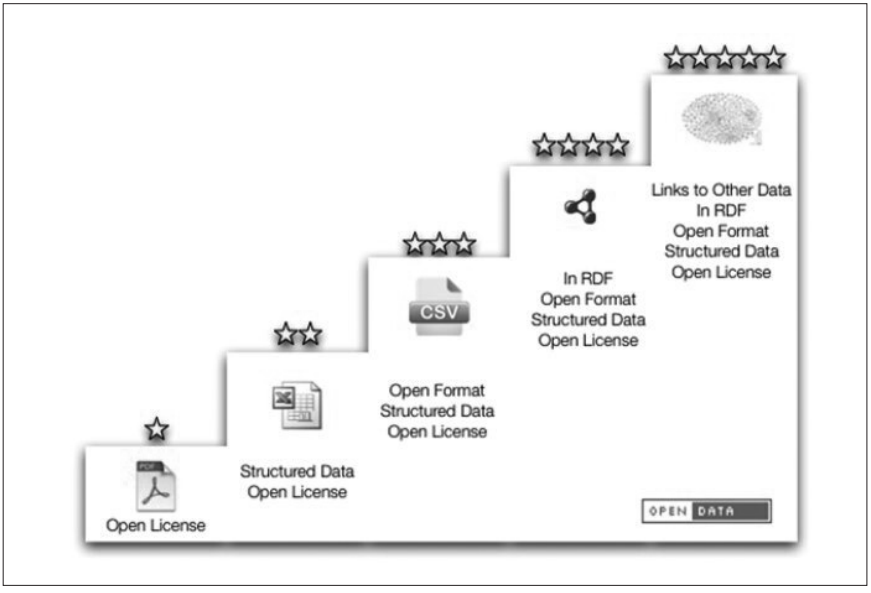
De esta manera, los datos abiertos enlazados son datos publicados de manera abierta mediante el uso de una licencia, tienen el propósito de vincular semánticamente los datos que forman parte de diferentes fuentes disponibles en el ambiente digital. Cuando son aplicados al contexto de la recuperación de información, propician el descubrimiento de obras, manifestaciones y expresiones que rodean un recurso o contenido de información en particular.

Literalmente cualquier cosa puede ser descrita con los datos abiertos enlazados. Los datos vinculados en la World Wide Web se pueden encontrar, compartir y combinar con los datos de otros usuarios. A diferencia de los sistemas tradicionales de gestión de datos, DAE libera información de contenedores propietarios para que cualquiera pueda usarla. No obstante, será responsabilidad del usuario valorar, calificar y definir la utilidad de los datos que utilice, pues DAE no resuelve por sí solo los problemas de veracidad y calidad que puedan aquejar a los datos.

Quizás valga la pena cuestionarse si los datos abiertos enlazados son demasiado buenos para ser verdad. No lo son. Los datos abiertos enlazados se crean en la web y tienen los mismos beneficios y problemas que la web. Los datos abiertos enlazados no son una fórmula mágica, ya que no subsanan los problemas de calidad de datos o de fallas en un determinado servicio de consulta. Nada inherente a los datos abiertos enlazados mejora la eficiencia de las consultas distribuidas.

Sin embargo, DAE proporciona nuevas formas de gestionar estos desafíos existentes de gestión de datos. La calidad de los datos es un problema en todos los sistemas de gestión de datos. Los datos sucios en una base de datos relacional o en un sitio web pueden convertirse muy rápidamente en datos enlazados sucios.

Figura 4. Las 5 estrellas de Linked Open Data



Fuente: Wood, Zaidman y Ruth 2014. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKIvpX0suthJ0yDYMA0qmbvCX4Xg?e=WR6zu9

Se pone de manifiesto que los datos vinculados se publican con mayor frecuencia de formas que no son específicas de una aplicación en particular. Eso puede exponer problemas de limpieza de datos que antes no eran transparentes y, al mismo tiempo, hacerlos más fáciles de detectar.

Hace muchos años lo único que realmente importaba era la información dentro de los contenedores, se creaban instituciones dedicadas a la preservación de la forma, no a la información. De hecho, las primeras bibliotecas protegían pergaminos, rollos y libros, no las palabras que contenían.

En este sentido, las palabras, al ser vistas como datos de índole textual, tienen la capacidad de conectar semánticamente millones de datos desde un punto de vista cualitativo. Esto sin duda es altamente relevante para establecer vinculaciones entre recursos y contenidos que forman parte de múltiples dominios contextuales.

Para la conformación de un entorno interoperable de datos abiertos enlazados, será necesario utilizar normatividad que contribuya al progreso de este propósito, pues los estándares propiciarán una comunicación efectiva entre los datos que forman parte de diferentes dominios. La normativdad es altamente relevante porque a nivel técnico fomenta la conectividad y transferencia de los datos, es decir, establece la posibilidad de conectar diferentes sistemas de información y bases de datos con atributos similares pero con características diversas.

Estandarización

Actualmente, las bases de datos todavía se consideran "silos", y los usuarios a menudo no quieren que otros toquen la base de datos de las que son responsables. Esta forma de pensar se basa en algunas suposiciones de la década de 1970: que sólo un puñado de expertos puede trabajar con bases de datos y que sólo el círculo interno del departamento de cómputo es capaz de comprender el esquema y el significado de los datos. Para DAE esta idea resulta obsoleta, pues en la era actual de Internet, millones de desarrolladores pueden crear aplicaciones valiosas siempre que obtengan datos interesantes.

En segundo lugar, los datos todavía están bloqueados en ciertas aplicaciones. El problema técnico con la arquitectura de información más común de hoy en día es que los metadatos y la información no están bien separados de la lógica de la aplicación. Los datos no se pueden reutilizar tan fácilmente como deberían.

Si alguien diseña una base de datos, a menudo conoce cierta aplicación para ser construida en la parte superior de una determinada estructura de trabajo. Si dejamos de enfatizar qué aplicaciones usarán nuestros datos y nos enfocamos en una descripción significativa de los datos en sí, se ganará más impulso a largo plazo. En esencia, DAE requiere que los datos se encuentren abiertos a cualquier tipo de aplicación y esto se puede lograr si se utilizan estándares abiertos como RDF para describir los metadatos que los acompañan.

Resource Description Framework

RDF son las siglas en inglés de Resource Description Framework, conocido en español como Marco para la Descripción de Recursos. Se trata de un modelo estándar para el intercambio de datos en la web. RDF tiene características que facilitan la fusión de datos incluso si los esquemas de datos subyacentes difieren, y apoya específicamente la evolución de los esquemas a lo largo del tiempo sin requerir que se modifiquen todos los consumidores de datos.

De acuerdo con Gottron y Staab (2018, 2036), "RDF proporciona un modelo para representar datos. Su trasfondo se establece históricamente en un entorno web donde se utiliza para representar información de una manera independiente del dispositivo y la plataforma". RDF representa los datos disponibles en la web a manera de grafo. Un grafo es una representación gráfica que consta de nodos (datos) y aristas (vinculaciones), en el cual se pueden apreciar las relaciones que los datos obtienen mediante su procesamiento. El grafo básico de RDF se denomina triple, estos triples están conformados por un sujeto, un predicado y un objeto y expresan que el sujeto está en cierta relación (identificado por el predicado) con el objeto. Las vinculaciones entre los datos se pueden interpretar

y representar a manera de grafo, donde los sujetos y los objetos son los nodos y el predicado es la vinculación que explica la relación que se ejerce entre los datos.

En la figura 5 se puede apreciar una ejemplificación de un grafo RDF básico. Se han tomado datos relativos al dominio documental, los cuales se refieren a las relaciones entre los autores y sus respectivas obras. Como es sabido, Eric Arthur Blair, conocido bajo el seudónimo de George Orwell, es el autor del libro *1984*.

Con este tipo de representaciones se otorga la posibilidad de descubrir las vinculaciones de datos que se conforman en el universo bibliográfico, pues este grafo básico puede extenderse según las relaciones de los datos que se establece en el contexto real de una determinada obra, lo que fomenta el descubrimiento de datos que caracterizan a las obras, expresiones y manifestaciones de tipo documental.

Existen tres tipos de nodos en un grafo RDF: IRIs, literales y nodos en blanco. Cualquier IRI y literal remite a una cosa que está disponible en el mundo. Estas cosas se llaman recursos. Cualquier cosa puede ser un recurso, incluidos los objetos físicos, documentos, conceptos abstractos, números y cadenas de texto; el término *recurso* es sinónimo de "entidad", tal y como se usa en la especificación RDF Semantics [RDF11-MT].

El recurso denotado por un IRI se denomina su referente y el recurso denotado por un literal se llama su valor literal. Los literales tienen tipos de datos que definen el rango de valores posibles, como cadenas textuales, números y fechas. A su vez, los tipos especiales de literales, cadenas con etiquetas de idioma, denotan cadenas de texto sin formato en un lenguaje natural.

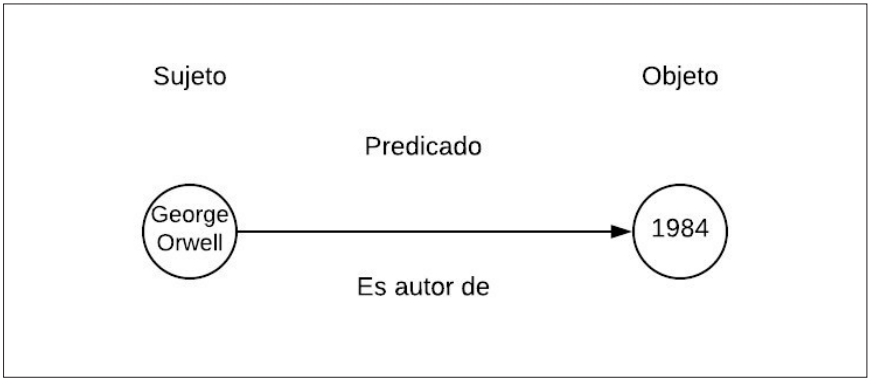
De esta manera, un triple RDF establece una vinculación que es indicada por el predicado y que cumple la función de unir al sujeto con el objeto. El predicado en sí es un IRI y denota una propiedad, es decir, un recurso que se puede considerar como una relación binaria. A diferencia de los IRI y los literales, los nodos en blanco no identifican recursos específicos. Las declaraciones que involucran nodos en blanco dicen que existe algo con las relaciones dadas, sin nombrarlo explícitamente.

RDF ayuda a proporcionar información descriptiva sobre los recursos que se encuentran en la Web, lo que permite el intercambio de información a través de diferentes aplicaciones sin que los datos pierdan su significado, lo que facilita la reutilización de los recursos. Además, RDF está basado en la idea de identificar cosas usando URIs (que son cadenas que identifican algún recurso —como imágenes, documentos, archivos, etcétera— para hacerlo disponible bajo una gran cantidad de esquemas como HTTP o FTP). RDF provee una forma para expresar enunciados simples acerca de los recursos usando propiedades y valores.

La capacidad que tiene RDF para procesar metadatos facilita la interoperabilidad entre diversas aplicaciones, lo que proporciona un mecanismo de perfecto intercambio de información a través de la Web. Tal y como se afirma en la Recomendación W3C, RDF tiene distintas áreas de aplicación; como la recuperación de recursos (que proporciona mejores prestaciones a los motores de búsqueda) y la catalogación en bibliotecas digitales (especificando también las relaciones de contenido disponibles en un sitio web determinado) (Senso 2003).

Sin embargo, la comunidad de usuarios de RDF se percató de que se necesitaba indicar algunas veces que lo que estaban

Figura 5. Grafo RDF básico



Fuente: elaboración propia, 2021.

describiendo eran tipos o clases específicas de recursos. RDF por sí sólo no proporciona tal vocabulario, por lo que las clases y propiedades se describen en RDF Vocabulary (también conocido como RDF schema. Su definición se encuentra en RDFs).

RDF Schema

Una de las características más importantes de RDF respecto a otros modelos de datos consiste en que posibilita la combinación de distintas fuentes de datos, aunque los esquemas subyacentes sean distintos. Por otro lado, es fácilmente extensible, ya que la evolución de los esquemas no requiere que los consumidores de datos sean adaptados.

El esquema RDF (RDFs) no proporciona ningún vocabulario específico, sino que ha sido creado para diseñar dichos vocabularios, por lo que, en cierta manera, se puede considerar como un sistema de tipado para RDF similar al que tienen algunos lenguajes de programación orientados a objetos. Sirve como una extensión semántica para construir ontologías de diferentes dominios temáticos.

El esquema contiene reglas e instrucciones para establecer las relaciones semánticas entre los datos y que ésta pueda ser comprensible por las aplicaciones informáticas. De esta manera, los usuarios de RDFs pueden definir las relaciones entre los datos con una semántica preestablecida.

RDFs sigue los principios de interoperabilidad, escalabilidad y descentralización. En particular, es posible interconectar de manera extensible descripciones de recursos (superponiendo declaraciones diferentes utilizando los mismos URI de recursos) o espacios de nombres de esquema (reutilizando o refinando las definiciones de propiedades y clases existentes) independientemente de su ubicación física en la web (Christophides 2009).

Actualmente el esquema RDF se encuentra en su versión 1.1, la cual fue publicada en el 2014. El esquema se conforma de los siguientes elementos (Rodríguez-Cruz 2018, 4):

- Recursos: cualquier objeto web identificable unívocamente por un URI, es decir, un identificador uniforme de recursos como un URI. Un recurso puede ser un documento HTML; una parte de una página web como por ejemplo un elemento HTML o XML dentro de un documento fuente, una colección de páginas, un sitio web completo; y en síntesis, cualquier recurso entendido como objeto de información.
- Propiedades: son aspectos específicos, características, atributos o relaciones utilizadas para describir recursos. Cada tipo de propiedad tiene sus valores específicos, define los valores permitidos, los tipos de recursos que puede describir y las relaciones que existen entre las distintas propiedades.
- Descripciones: Son el conjunto de un recurso, un nombre de propiedad y el valor de esa propiedad —sujeto, predicado y objeto, respectivamente.

De esta manera el esquema RDF permite la construcción de vocabularios interoperables semánticamente, lo cual fomenta la consolidación paulatina de un entorno global de datos que sean capaces de conectarse semánticamente y establecer relaciones de significado entre los datos, recursos y contenidos de información que están disponibles en el ambiente digital.

SPARQL Query Language for RDF

La manera de acceder a los datos que han sido estructurados mediante RDF se basa en la utilización de un lenguaje de consulta, el cual se denomina SPARQL. Este lenguaje de consulta puede utilizarse para expresar consultas que permiten interrogar diversas fuentes de datos, si los datos se almacenan de forma nativa como RDF o son definidos mediante vistas RDF a través de algún sistema. SPARQL contiene las capacidades para la consulta de los patrones obligatorios y opcionales de grafo, junto con sus conjunciones y disyunciones. SPARQL también soporta la ampliación o las restricciones del ámbito de las consultas indicando los grafos sobre los que se opera. Los resultados de las consultas SPARQL pueden ser

conjuntos de resultados o grafos RDF (W3C, 2013). Cada base de datos necesita un lenguaje de consulta. SPARQL es para los datos RDF como SQL es para una base de datos relacional. Por lo tanto, SPARQL es el lenguaje de consulta para datos vinculados. El propósito principal de SPARQL es proporcionar un lenguaje formal en el que se puedan formular preguntas significativas y generar consultas complejas de información.

Las consultas SPARQL pueden adoptar distintas formas. La más común es una consulta SELECT, que selecciona información según las restricciones que han sido configuradas en un determinado entorno de datos. Este tipo de consulta es muy similar en forma a las consultas de selección de SQL. Cada consulta SPARQL SELECT está organizada de la siguiente manera:

1. PREFIX (Prefijos de espacio de nombres).
2. SELECT (Define lo que desea recuperar).
3. FROM (Especifica el conjunto de datos del cual extraer los resultados).
4. WHERE (Describe los criterios sobre los cuales se basa la selección. Esta descripción tiene la forma de un patrón triple de consulta).
5. ORDER BY y LIMIT (son modificadores que afectan el resultado deseado).

Las plataformas de datos abiertos enlazados que están disponibles actualmente en la web a menudo exponen un punto final SPARQL. Un punto final SPARQL es un servicio de consulta accesible desde la web que acepta el lenguaje de consulta SPARQL. Un HTTP GET en un punto final SPARQL generalmente devuelve un formulario de consulta HTML. En la figura 6 puede apreciarse uno de los formularios de consulta SPARQL más avanzados en la actualidad; se trata del servicio de consulta de Wikidata (disponible en https://query.wikidata.org/). Mediante este servicio de consulta se ha ejemplificado una consulta básica en donde se han solicitado datos referentes a personas que nacieron en Nueva York y que se encuentran registradas en uno de los conjuntos de datos que forman parte de Wikidata.

A su vez, los resultados de esta consulta se pueden apreciar en la figura 7, en donde a manera de tabla se presentan los nombres de las personas que están registradas en los conjuntos de datos de Wikidata. Se considera que este tipo de recuperación de información es de carácter semántico, pues permite obtener acceso a enlaces que se relacionan con el nombre de la persona y tener un acercamiento a la descripción que representa a cada persona.

Además de la consulta SELECT, SPARQL permite realizar las consultas DESCRIBE, las cuales proporcionan una forma rápida de preguntar: "¿Qué se sabe acerca de un recurso en particular?" A su vez, la consulta ASK permite determinar si una consulta en particular devolvería resultados, y la consulta CONSTRUCT da la posibilidad de construir nuevos grafos RDF a partir de los resultados de la consulta SPARQL. De esta manera el lenguaje de consulta SPARQL es el principal elemento que permite el establecimiento de interfaces de recuperación de información de tipo semántico. De acuerdo con Polleres (2014, 1961), "[...] la existencia de un lenguaje de consulta estándar de este tipo ha contribuido significativamente a la creciente adopción de RDF como formato de datos básico en la web durante los últimos años". Será trascendental perfeccionar el uso de SPARQL para adoptar su estructura en la latente generación de nuevos buscadores de información de carácter semántico.

Vocabularios

Los vocabularios son elementos imprescindibles para el funcionamiento correcto de los datos abiertos enlazados en un ambiente de recuperación de información. Una de las principales barreras para la visualización de datos abiertos enlazados es la dificultad que tienen los editores y generadores de datos para determinar qué vocabularios utilizar para describir su semántica. Estos vocabularios proporcionan un "aditamento semántico" que permite que los datos simples se conviertan en "datos significativos".

De esta manera, un vocabulario consta de clases, propiedades y tipos de datos que definen el significado de los datos.

> Los vocabularios se expresan en sí mismos como datos enlazados. Cuando un vocabulario no está publicado o no está disponible ya, los humanos y las máquinas no tienen acceso a la definición de los términos utilizados para calificar los datos. Esto rompe la interoperabilidad semántica, uno de los fundamentos de la web semántica (Vandenbussche *et al.* 2014, 2).

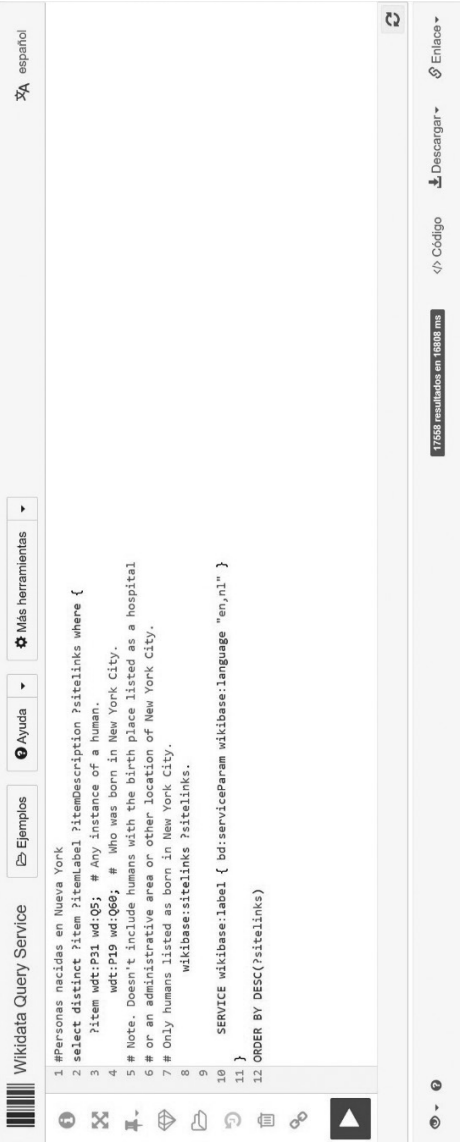
Así, los vocabularios ayudan a las computadoras a comprender el significado que se asigna a los datos, lo que facilita la construcción de un entorno interoperable de datos. Se puede afirmar que los vocabularios son el lenguaje que otorga la semántica a los datos.

Los vocabularios enlazados ayudan a la adquisición de conocimiento a través de un control estricto y de una contextualización de los datos (conceptos, objetos, etc.). Este enfoque hace posible los procesos habituales de metadatos. A la vez, permite enlazar vocabularios con los registros de datos en sí mismos, lo que proporciona una infraestructura que mejora la eficacia del uso y recuperación de la información (Mendez y Greenberg 2012, 3).

Así pues, los tipos de objetos que a menudo se describen primero en Datos vinculados son personas, proyectos, recursos web, publicaciones y direcciones. También es posible que se desee migrar una taxonomía existente u otro esquema de clasificación a un vocabulario RDF. Los vocabularios que definen estos términos a veces se denominan vocabularios RDF básicos. De esta manera, en la tabla 1 se muestran algunos vocabularios RDF básicos y de uso común en diversas aplicaciones de datos.

En la tabla 1 pueden apreciarse vocabularios de DAE que forman parte de diferentes dominios temáticos. Estos vocabularios puden complementarse para generar entronos de datos interoprables entre sí, mediante una semántica definida previamente. En el proyecto Linked Open Vocabularies (disponible en https://lov.linkeddata.es/dataset/lov/) se puede tener un acercamiento más específico en el tema de la generación de vocabularios de datos abiertos enlazados. Este proyecto reúne vocabularios y ontologías que pueden ser utilizados para describir datos de diferentes dominios temáticos, con la singularidad de que estos vocabularios tienen compatibilidad con la norma RDF.

Figura 6. Wikidata Query Service. Consulta a personas nacidas en Nueva York



Fuente: https://query.wikidata.org/, 2021. Esta figura es para efectos ilustrativos; para verla de forma detallada consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthJ1DPRI2gPHozB83Hw?e=inrq6m.

Figura 7. **Resultados de la consulta SPARQL relativos a personas nacidas en Nueva York**

item	itemLabel	itemDescription	sitelinks
wd:Q19848	Lady Gaga	American singer, songwriter, actress, and activist	178
wd:Q180589	Boris Johnson	Prime Minister of the United Kingdom since 2019	123
wd:Q36949	Robert De Niro	Italian-American actor, director and producer	122
wd:Q4985	Herman Melville	American novelist, short story writer, essayist, and poet	121
wd:Q102124	Sigourney Weaver	American actress	115
wd:Q16390	Humphrey Bogart	American actor (1899-1957)	106
wd:Q79904	J. D. Salinger	American writer (1919–2010)	102
wd:Q80596	Arthur Miller	American playwright	101
wd:Q181667	Washington Irving	American writer, historian and diplomat (1783-1859)	100
wd:Q47899	Paris Hilton	American socialite and media personality	99
wd:Q132537	Robert Oppenheimer	American theoretical physicist, known as "father of the atomic bomb"	98
wd:Q41142	Jane Fonda	American actress and activist	97
wd:Q42745	Rita Hayworth	American actress, dancer and director (1918-1987)	97
wd:Q93157	Eugene O'Neill	American playwright, and Nobel laureate in Literature	95

Fuente: https://query.wikidata.org/, 2021. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKIvpX0suthJ1ECG4w8aUfJQ58WA?e=2rJq3U.

Ontologías

Las ontologías aplicadas en el contexto de los datos abiertos enlazados y su interacción con el proceso de recuperación de información son representaciones de datos que definen los tipos, las propiedades y las relaciones que existen entre las entidades de un determinado dominio temático o de conocimiento. De acuerdo con Lozano Tello (2001), las ontologías tienen los siguientes componentes que servirán para representar el conocimiento de algún dominio:

- Conceptos: son las ideas básicas que se intentan formalizar. Los conceptos pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento, etcétera.
- Relaciones: representan la interacción y el enlace entre los conceptos del dominio. Suelen formar la taxonomía del dominio. Por ejemplo: subclase-de, parte-de, parte-exhaustiva-de, conectado-a, etcétera.
- Funciones: son un tipo concreto de relación donde se identifica un elemento mediante el cálculo de una función que considera varios elementos de la ontología. Por ejemplo, pueden aparecer funciones como categorizar-clase, asignar fecha, etcétera.
- Instancias: se utilizan para representar objetos determinados de un concepto.
- Axiomas: son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología. Por ejemplo: "Si A y B son de la clase C, entonces A no es subclase de B", "Para todo A que cumpla la condición C1, A es B", etcétera.

Originalmente, la ontología es el estudio filosófico de la naturaleza de la existencia, así como de las categorías básicas del ser y sus relaciones. La ontología se ocupa de cuestiones relativas a qué entidades existen o se puede decir que existen y cómo dichas entidades pueden agruparse, relacionarse dentro de una jerarquía y subdividirse de acuerdo con similitudes y diferencias (Wang 2013).

Tabla 1. Vocabularios RDF básicos y de uso común

Nombre	Prefijo	Namespace URI	Descripción
Airport Ontology	air:	http://www.daml.org/2001/10/html/airport-ont#	Aeropuertos más cercanos
BIBO	bibo:	http://purl.org/ontology/bibo/	Datos bibliográficos
Bibframe	bf	https://www.loc.gov/bibframe/docs/	Datos bibliográficos
Creative Commons Rights Expression Language	cc:	http://creativecommons.org/ns#	Licencias
Dublin Core Elements	dc:	http://purl.org/dc/elements/1.1/	Descripción de recursos de información
FOAF	foaf:	http://xmlns.com/foaf/spec/	Descripción de funciones de personas
Geo	pos:	http://www.w3.org/2003/01/geo/wgs84_pos#	Posiciones geográficas
Geonames	gn:	http://www.geonames.org/ontology#	Ubicaciones geográficas
Object Reuse and Exchange	ore:	http://www.openarchives.org/ore/terms/	Material cartográfico
RDF	rdf:	http://www.w3.org/1999/02/22-rdfsyntax-ns#	Marco núcleo
RDFS	rdfs:	http://www.w3.org/2000/01/rdfschema#	Vocabularios RDF
SIOC	sioc:	http://rdfs.org/sioc/ns#	Comunidades en línea
SKOS	skos:	http://www.w3.org/2004/02/skos/core#	Vocabularios controlados
Web Ontology Language (OWL)	owl:	http://www.w3.org/2002/07/owl#	Ontologías
XML Schema Datatypes	xsd:	http://www.w3.org/2001/XMLSchema#	Tipos de datos
ISBD	isbd:	http://metadataregistry.org/schema/show/id/25.html	Descripción bibliográfica de recursos de información

Nombre	Prefijo	Namespace URI	Descripción
British Library Terms	bl:	http://www.bl.uk/schemas/bibliographic/bl-terms	Esquema bibliográfico
Event ontology	event:	http://motools.sourceforge.net/event/event.html#term_time	Descripción de eventos y sus participantes

Fuente: elaboración propia con datos de Wood, Zaidman y Ruth (2014).

Existen dos propuestas de ontologías muy significativas para el contexto de los datos abiertos enlazados. La primera de ellas es el lenguaje de ontologías web (por sus siglas en inglés OWL). OWL está diseñado para ser usado en aplicaciones que necesitan procesar el contenido de la información en lugar de únicamente representar información para los humanos.

OWL posibilita un mejor mecanismo de interpretabilidad de contenido Web que los mecanismos admitidos por XML, RDF y el esquema RDF (RDF-S) al proporcionar un vocabulario adicional junto con una semántica formal. OWL tiene tres sublenguajes con un nivel de expresividad creciente: OWL Lite, OWL DL y OWL Full (W3C, 2004).

La versión actual de OWL, también conocida como "OWL 2", fue desarrollada por el W3C OWL Working Group y publicada en 2009, con una segunda edición publicada en 2012. Una gran cantidad de estudios de caso abordan la aplicación de ontologías como fundamento para conocer las variables del uso de DAE en el contexto de la recuperación de información. Dichos estudios son altamente especializados acorde a la temática y naturaleza de los datos que representa a los proyectos.

Por otra parte, SKOS (siglas de *Simple Knowledge Organization System*) es una iniciativa del W3C en forma de aplicación de RDF que proporciona un modelo para representar la estructura básica y el contenido de esquemas conceptuales como listas de encabezamientos de materia, taxonomías, esquemas de clasificación, tesauros y cualquier tipo de vocabulario controlado (Sánchez 2021).

El modelo de datos SKOS es una ontología definida con OWL Full. Al estar basado en RDF, SKOS estructura los datos en forma de tripletas que pueden ser codificadas en cualquier sintaxis válida para RDF. De esta manera, SKOS puede ser utilizado conjuntamente con L para expresar formalmente estructuras de conocimiento sobre un dominio concreto, ya que SKOS no puede realizar esta función al no tratarse de un lenguaje para la representación de conocimiento formal.

El modelo SKOS contempla el establecimiento de enlaces entre conceptos denominados relaciones semánticas. Estas relaciones pueden ser jerárquicas o asociativas, y contemplan la posibilidad de ampliar la tipología de relaciones. Los conceptos también pueden agruparse en colecciones que a su vez pueden etiquetarse y ordenarse. SKOS se complementa con la posibilidad de que conceptos de diferentes esquemas se pueden mapear entre sí empleando relaciones jerárquicas, asociativas o de equivalencia exacta.

Las ontologías constituyen una buena alternativa para representar el conocimiento compartido sobre un dominio. Dejando de lado las características accidentales, las ontologías esperan representar un punto de vista objetivo de una parte de la realidad, por lo que su aplicación en el contexto de los datos abiertos enlazados permitirá generar un entorno de mayor interoperabilidad semántica.

Metadatos

Los metadatos son elementos descriptivos que permiten representar los atributos de los recursos de información y facilitan generar registros de datos correspondientes a dichos recursos. Los metadatos técnicos, descriptivos y administrativos son los más comunes dentro de los sistemas para la recuperación de información.

De acuerdo con Gartner (2016, 53), se considera que los metadatos tienen los siguientes tres componentes fundamentales, algunos de los cuales (pero no necesariamente todos) se definen en un estándar determinado:

- Semántica: los significados de los campos o elementos en los que se colocan los metadatos.

- Sintaxis: la forma en que se codifican los metadatos, tal vez en una hoja de cálculo, tabla de base de datos o un formato más genérico como XML (*eXtensible Markup Language*, del que hablaremos más adelante en este capítulo).
- Reglas de contenido: las reglas, si las hay, que rigen el contenido de los metadatos en sí, qué se registra, qué forma deben tomar y qué debe excluirse.

Uno de los estándares de metadatos que más se ha utilizado para la representación de los datos que forman parte del dominio documental es Dublin Core. Este esquema ha adaptado sus alcances en una versión que fomenta la construcción de entornos de datos enlazados conocida como DCMI RDF (disponible en https://www.dublincore.org/schemas/rdfs/). En la figura 8, puede apreciarse un grafo básico de RDF, en el cual se especifican las vinculaciones entre los datos mediante los elementos descriptivos del esquema de metadatos de Dublin Core.

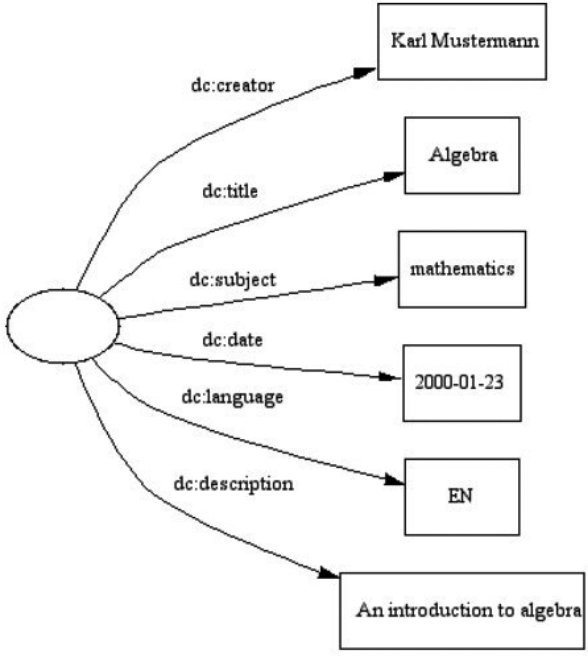
En este caso, los metadatos permiten establecer las relaciones semánticas entre los datos que forman parte de un determinado recurso de información. En el ámbito de las bibliotecas, este tipo de representaciones pueden ser bastante representativas, pues se fomentaría la conformación de un entorno interoperabe de datos documentales que permitiera descubrir las relaciones entre obras, expresiones, manifestaciones y cada uno sus de sus creadores. Por lo tanto, los metadatos para DAE son un mecanismo que permite representar los atributos y las relaciones existentes entre los datos que forman parte de un determinado recurso informativo.

Catálogos de datos

Los catálogos de datos se han convertido en un pilar importante en el ciclo de vida de la gestión de datos. De hecho, casi todos los pasos del ciclo de vida de los datos se describen en los campos de metadatos o se puede acceder a ellos a través de la interfaz en línea del catálogo de datos.

De acuerdo con Quimbert, Jeffery, Martens Martin y Zhao (2020, 141), los catálogos de datos existen para recopilar, crear y

Figura 8. Grafo RDF-DCMI




```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:dc="http://purl.org/dc/elements/1.1/">

  <rdf:Description>
    <dc:creator>Karl Mustermann</dc:creator>
    <dc:title>Algebra</dc:title>
    <dc:subject>mathematics</dc:subject>
    <dc:date>2000-01-23</dc:date>
    <dc:language>EN</dc:language>
    <dc:description>An introduction to algebra</dc:description>
  </rdf:Description>
</rdf:RDF>
```

Fuente: https://www.dublincore.org/specifications/dublin-core/dcq-rdf-xml/.

mantener metadatos. Estos registros están indexados en una base de datos y los usuarios finales deben acceder a la información a través de una interfaz fácil de usar. Esta interfaz debe ofrecer funcionalidades de búsqueda de datos comunes que permitan a los usuarios limitar su búsqueda de acuerdo con diferentes criterios: palabras clave (vocabularios controlados), ubicación geográfica, resolución temporal y espacial, y fuentes de datos.

En este sentido, el World Wide Web Consortium (W3C) se ha dado a la tarea de generar un vocabulario que permita describir este tipo de catálogos. Por ejemplo, DCAT es un vocabulario RDF diseñado para facilitar la interoperabilidad entre catálogos de datos publicados en la Web.

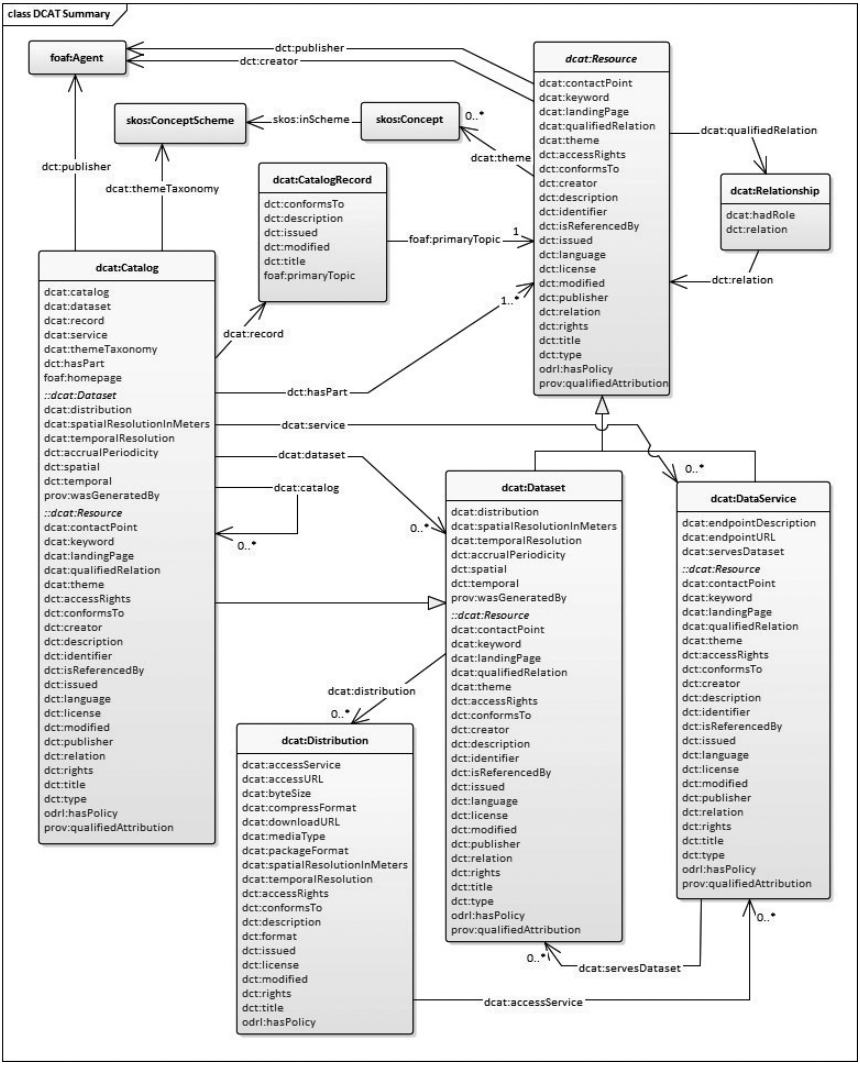
DCAT permite a un editor describir conjuntos y servicios de datos en un catálogo utilizando un modelo estándar y un vocabulario que facilite el consumo y la agregación de metadatos de varios catálogos (W3C 2020). Esto permite aumentar la capacidad de descubrimiento de conjuntos de datos, a la vez que hace posible tener un enfoque descentralizado para publicar catálogos de datos y hace posible la búsqueda federada de conjuntos de datos en catálogos de varios sitios utilizando el mismo mecanismo y estructura de consulta.

El espacio de nombres para los términos DCAT es http://www.w3.org/ns/dcat#. El prefijo sugerido para el espacio de nombres DCAT es dcat. A su vez, en la figura 9 puede apreciarse un esquema que representa el modelo DCAT en el cual se muestran las clases de recursos que pueden ser miembros de un catálogo y las relaciones entre ellos.

En el modelo DCAT, un conjunto de datos es una entidad conceptual y se puede representar mediante una o más distribuciones que serían el conjunto de datos para su transferencia. Las distribuciones de un conjunto de datos se pueden proporcionar a través de los denominados servicios de datos.

Se estima que el incremento de los catálogos de datos ayude en gran manera a mejorar la interoperabilidad de los datos respecto al proceso de su implementación en los sistemas para la recuperación de información, pues estos sistemas deberán contar con la

Figura 9. Descripción general del modelo DCAT, que muestra las clases de recursos que pueden ser miembros de un catálogo y las relaciones entre ellos



Fuente: https://www.w3.org/TR/vocab-dcat-2/. Esta figura es para efectos ilustrativos, para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKIvpX0suthJ1_cNlCWTyBVQElWA?e=9V0uqj.

capacidad para concretar datos de atributos similares disponibles en diversas fuentes del ambiente digital. Algunas de estas fuentes se enmarcan en los desarrollos actuales de la implementación de los datos abiertos enlazados.

Desarrollos

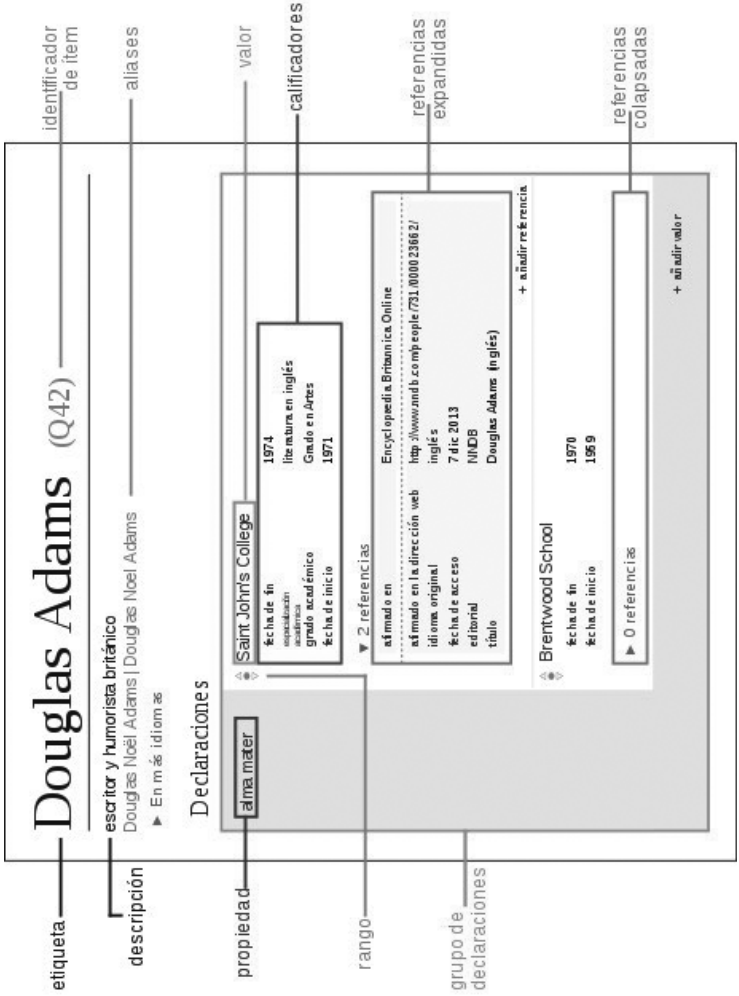
La implementación de los datos abiertos enlazados en el ambiente digital es un tema que se caracteriza por la generación de sistemas y plataformas que permiten la consulta de información mediante la lógica de descubrimiento que se establece a través de DAE. Actualmente, estos sistemas se encuentran en constante desarrollo generando conjuntos de datos y enlazándolos con un objetivo en particular, el cual puede ser desde proporcionar a los usuarios un nuevo mecanismo para la búsqueda y recuperación de información, hasta fomentar la interoperabilidad de los datos mediante la implementación de redes colaborativas de trabajo. En este apartado, se abordan los casos más representativos de la adopción de DAE en la generación de paltaformas digitales, lo cual se ha considerado un paso para la conformación de los llamados sistemas complejos para la recuperación de información.

Wikidata

Wikidata es una base de datos libre, colaborativa y multilingüe que sirve como base de datos secundaria y que recopila datos estructurados para dar soporte a Wikipedia, Wikimedia Commons, así como a otras wikis del movimiento Wikimedia y a cualquier usuario en el mundo (Wikidata 2021). El repositorio de Wikidata es el núcleo central de este proyecto, pues este sistema almacena los datos mediante elementos, propiedades y valores que representan declaraciones. Estas declaraciones describen los atributos de cada dato. En la figura 10 puede apreciarse un elemento de Wikidata en donde se describen sus elementos principales.

De acuerdo con Wikidata, para el caso de una persona se puede agregar una propiedad para indicar dónde se educó, especificando

Figura 10. Descripción de un elemento en Wikidata



Fuente: https://www.wikidata.org/wiki/Wikidata:Introduction/es#/media/File:Datamodel_in_Wikidata_es.svg. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKIvpX0suthJ4POFWaZAQSXxEPiA?e=58GXIF.

el valor del centro educativo. Para edificios, es posible asignar propiedades de coordenadas geográficas indicando los valores de longitud y latitud. Las propiedades también pueden enlazarse a bases de datos externas. "Una propiedad que enlaza un elemento a una base de datos externa, como a una base de datos de control de autoridades utilizada por bibliotecas y archivos, se llama un identificador" (Wikidata 2021, s.p.).

Wikidata es un proyecto que se encuentra en constante desarrollo, a la par de los proyectos subyacentes que conforman su estructura y fuentes de alimentación de datos. Además, este proyecto fue uno de los pioneros en implementar puntos SPARQL para la consulta de datos enlazados directamente en plataforma.

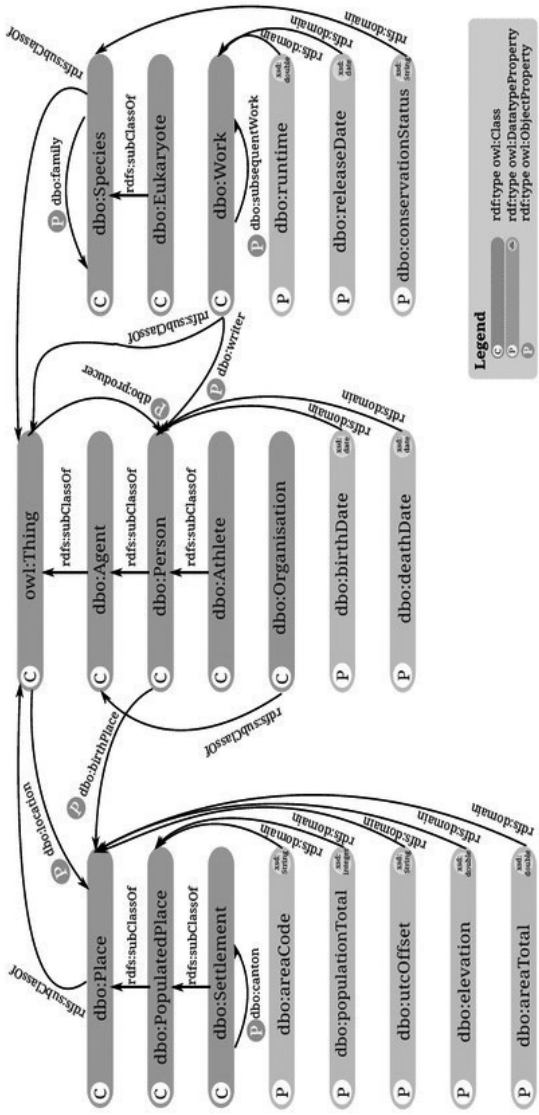
DBpedia

La Universidad Libre de Berlín y la compañía OpenLink Software han generado un proyecto de extracción de datos de Wikipedia para proponer una versión semántica de esta plataforma digital. Este proyecto ha sido denominado DBpedia. Un sistema de recomendación como DBpedia deberá contar con la capacidad de ayudar a los usuarios a encontrar lo que es relevante para ellos en una amplia gama de posibilidades; de hecho, los datos abiertos enlazados aplicados en el contexto de los sistemas de información deberán propiciar la generación de diversas variables para que el usuario seleccione la que más le convenga para satisfacer su necesidad informativa.

"La ontología de DBpedia se basa en OWL y forma parte de la columna vertebral de DBpedia" (Morsey *et al.* 2012, 169). Esta ontología describe clases; por ejemplo, escritor, obra musical y libro. También describe propiedades, como lugar de nacimiento, tiempo de ejecución y autores. La ontología DBpedia es de carácter superficial mediante el uso de dominios cruzados que se han creado manualmente en función de las consultas de información más utilizadas en Wikipedia. La ontología actualmente cubre 685 clases que están descritas por 2.795 propiedades diferentes.

En la figura 11 puede apreciarse una representación de una parte de la ontología de DBpedia en la cual se observan las clases

Figura 11. Parte de la ontología de DBpedia



Fuente: Petrovic y Arsic 2013. Disponible en https://www.researchgate.net/figure/Part-of-DBpedia-ontology_fig1_273461604. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKIvpX0suthJ4Uo60FVS8Nuacxyg?e=17w3BD.

y propiedades que son utilizadas para describir los datos que forman parte de Wikipedia. En color azul, pueden apreciarse las clases principales utilizadas en el Ontology Web Language. Por otra parte, en color naranja se aprecian las propiedades que describen los atributos de los datos que están colocados en el dominio de representación. Finalmente, el nodo naranja, con la letra "P", representa a las propiedades de los objetos que son redirigidos en la DBpedia.

El progreso de la DBpedia depende esencialmente de la integridad de uso entre las clases y propiedades que son colocadas en la ontología pues, en los últimos años, se perciben redudandancias que afectan el uso irrestricto y armónico de las clases para plasmar la jerarquía principal de los datos y las propiedades para representar a sus atributos.

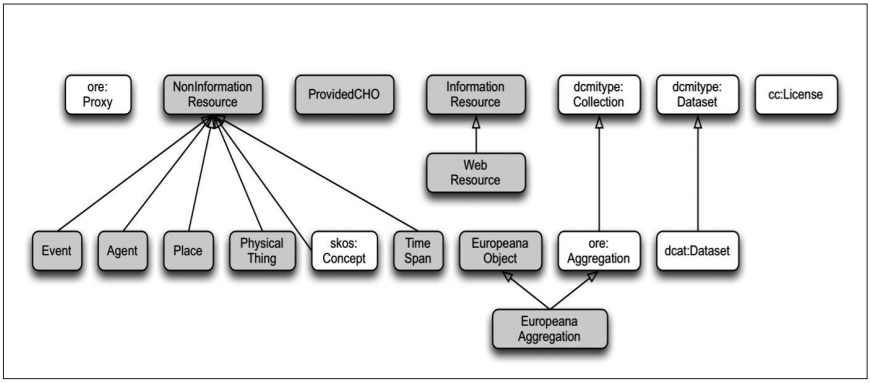
Europeana Data Model

La biblioteca digital europea, también conocida como Europeana, cuenta con un modelo de datos compatible con los principios de los datos abiertos enlazados. El Europeana Data Model (EDM) es la especificación formal de las clases y propiedades que pueden usarse en Europeana. En la figura 12 puede apreciarse una representación de la jerarquía de clases que es utilizada en el modelo.

Las clases incluidas en EDM se muestran en los rectángulos de color azul. A su vez, las clases de los rectángulos blancos son reutilizadas por otros esquemas de datos interoperables con EDM. Cada una de las clases puede trasladarse para la vinculación de datos en el contexto de los recursos de información, tal y como se aprecia en la figura 13.

Por ejemplo, la clase providedCHO de EDM permite identificar objetos reales mediante la relación entre los datos que remiten a personas (edm:agent) y conceptos que representan el contenido temático de dicho objeto (skos:concept). En esta relación puede apreciarse la interoperabilidad de EDM con la ontología SKOS, lo que fomenta la vinculación de datos disponibles en diferentes fuentes digitales y vocabularios semánticos.

Figura 12. Jerarquía de clases de EDM

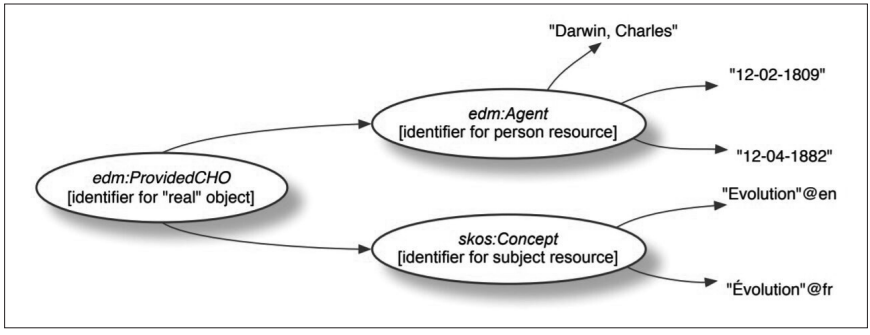


Fuente: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf.

De esta manera, EDM proporciona un esquema para la definición de los datos que pueden representarse bajo el contexto de Europeana y una serie de directrices que permiten la manipulación de los datos que está orientado a la descripción de las operaciones para manejar los datos en diversos marcos contextuales.

El objetivo de EDM es, claramente, permitir que los usuarios realicen mejores consultas y obtener mejores resultados mediante

Figura 13. EDM aplicado en el contexto de recursos de información



Fuente: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf.

el uso de un sofisticado y cada vez más ontológico modelo de metadatos sobre el cual dejar que funcione el motor de búsqueda. Europeana ahora está trabajando en mejorar la calidad de las respuestas a las consultas de los usuarios, pero aún queda mucho por hacer (Peron *et al.* 229).

The British National Bibliography

La plataforma de datos enlazados de la Bibliografía Nacional Británica (BNB) proporciona acceso a la bibliografía nacional publicada como datos abiertos enlazados, la cual está disponible para los usuarios a través de los servicios SPARQL. En este sentido, la plataforma proporciona dos interfaces diferentes: un editor SPARQL y un nodo SPARQL de servicio para realizar consultas remotas.

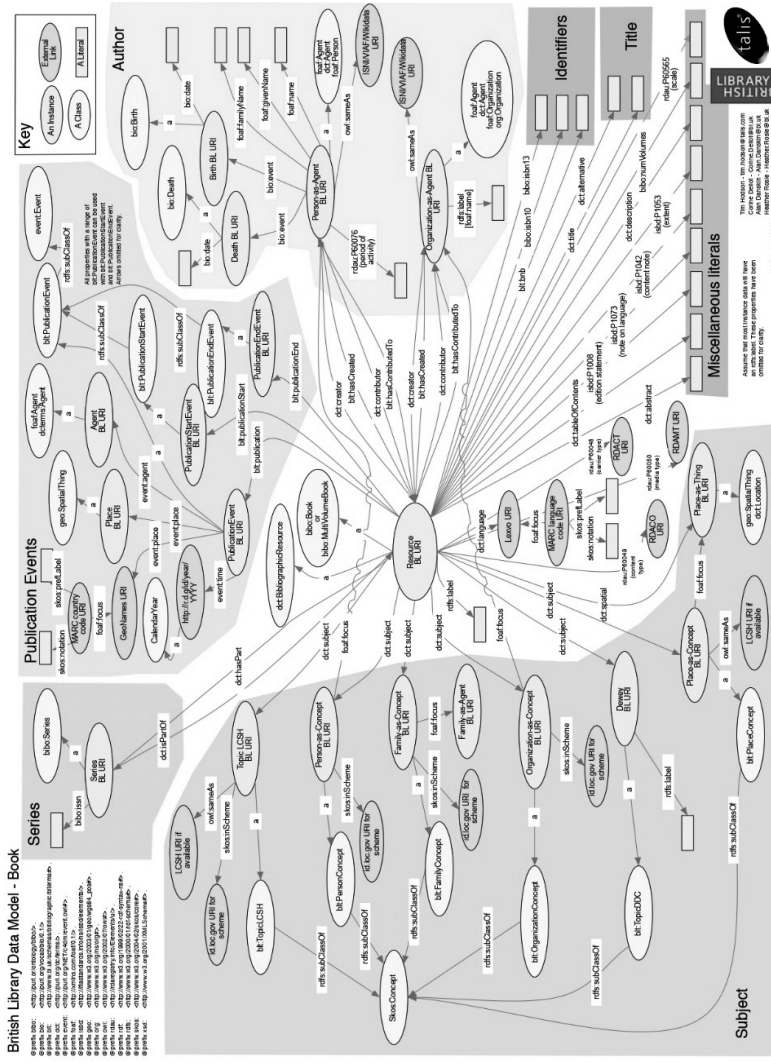
La BNB contiene conjuntos de datos que incluyen registros a libros publicados (incluidas monografías publicadas a lo largo del tiempo), publicaciones seriadas y libros nuevos y futuros, que representan aproximadamente 4,4 millones de registros. El conjunto de datos está disponible bajo una licencia Creative Commons CC0 1.0 Universal Public Domain Dedication.

En la figura 14 puede apreciarse una ejemplificación del modelo de datos generado por la Biblioteca Británica para representar en DAE los registros de libros que forman parte de la BNB. En este ejemplo puede apreciarse el uso de diversos vocabularios semánticos y ontologías, por ejemplo bibo, bio, bit, dct, owl, rdf y skos.

Este modelo tiene la intención de enlazar los atributos de los libros de la BNB de acuerdo a su temática, autoridad, eventualidad y seriación. De acuerdo con Deliot (2014, 1), el propósito de este modelo es romper con los formatos específicos de las bibliotecas y utilizar más estándares basados en XML de dominios cruzados para llegar a audiencias más allá del mundo de las bibliotecas. Es decir, conectar los datos de las bibliotecas con otras fuentes disponibles en el entorno web.

De esta manera, la BNB representa uno de los mayores esfuerzos realizados por las bibliotecas con el objetivo de establecer un entorno interoperable de datos mediante la reutilización de metadatos y vocabularios semánticos. Se trata de un modelo de

Figura 14. Modelo de datos de la Biblioteca Británica para libros



Fuente: http://vll-minos.bl.uk/bibliographic/pdfs/publishing_bnb_as_lod.pdf. Esta figura es para efectos ilustrativos; para verla de forma detallada consultarla, en: https://1drv.ms/u/s!AkMKlvpX0suthJ4jmvAJU2DGQxeej?e=Rhpy06.

datos de biblioteca que puede ser compatible con otros modelos de índole semántica mediante el uso de la norma RDF.

A su vez, el modelo pone de manifiesto un cambio de paradigma relacionado con los tradicionales formatos de codificación utilizados por las bibliotecas, en el sentido de incursionar en el aprovechamiento y la explotación de formatos más flexibles y adaptables a un entorno de datos cada vez más global e interconectado.

Share VDE

Los proyectos relacionados con la implementación de los datos abiertos enlazados en el ambiente de las bibliotecas son cada vez más frecuentes y visibles en el ambiente de la web. En este sentido, la generación de proyectos de DAE en el ambiente de las bibliotecas requiere tener presente que no serían posibles sin el desarrollo de vocabularios y ontologías para el tratamiento de los datos de índole documental.

Share-VDE es una iniciativa impulsada por bibliotecas que reúne los catálogos bibliográficos y los archivos de autoridad de una comunidad de bibliotecas en un entorno de descubrimiento compartido basado en datos vinculados. Share-VDE incrementó su alcance para abarcar una comunidad más amplia de instituciones también de los dominios del arte y la música al crear la familia Share (Share-VDE 2021). La base central de Share-VDE es el vocabulario Bibframe 2.0. Esta plataforma permite tener acceso a datos enlazados disponibles en diferentes paltaformas digitales diseñdas por bibliotecas, por ejemplo con el catálogo de la Biblioteca del Congreso de los Estados Unidos, con World Cat de OCLC y el portal de datos de la Biblioteca Nacional de Francia.

Además, Share VDE enlaza contenidos disponibles en la Wikipedia con los datos de autoridades disponibles en el fichero de autoridades virtual internacional (por sus siglas en inglés, VIAF). Share VDE permite realizar búsquedas de información a través de personas, obras, editores y temas.

En la figura 15 se puede apreciar una imagen que ejemplifica el resultado de una búsqueda de información realizada en la

plataforma Share-VDE. El proyecto Share-VDE incluye una plataforma de descubrimiento virtual (http://www.share-vde.org) con una adaptación del modelo de datos BIBFRAME que se desarrolló para proporcionar una opción de descubrimiento de datos abiertos enlazados. Desde el inicio, en octubre de 2016, y a diferencia de otros esfuerzos en la comunidad BIBFRAME que se centraron en la creación de nuevos metadatos en BIBFRAME, el proyecto Share-VDE comenzó con un enfoque en la conversión de metadatos de MARC a RDF utilizando el vocabulario BIBFRAME y otras ontologías adicionales.

De acuerdo con Casalini (2017), Share-VDE también espera ayudar a revelar una riqueza dentro de los datos de las colecciones, a menudo ocultas o no expresadas en un catálogo tradicional. Esto puede llevarse a cabo a través de la experimentación y configuración de las opciones para la futura creación de datos, mejora y puesta en común de todo tipo de recursos con la comunidad bibliotecaria.

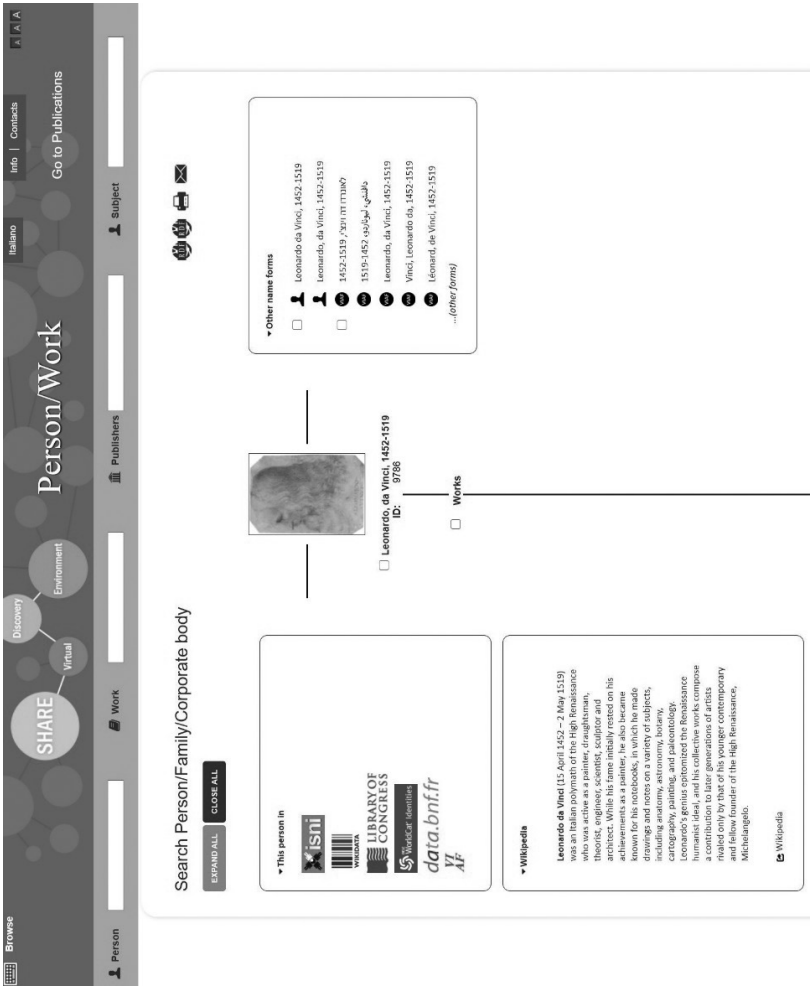
Además de esto, la plataforma fomenta el descubrimiento de información con base en la lógica del establecimieinto de datos abiertos enlazados, pues la implementación de DAE en los sistemas para la recuperación de información propicia la interacción entre recursos de información mediante el descubrimiento de obras, expresiones y manifestaciones que forman parte del universo de información y en donde los datos son piezas fundamentales para motivar el descubrimiento de nuevos hallazgos informativos.

OpenCitations

OpenCitations es una organización independiente dedicada a la publicación de datos bibliográficos y citas mediante el uso de datos abiertos enlazados. También se dedica a la promoción de las citas abiertas, en particular como miembro fundador de la Iniciativa para las Citas Abiertas (I4OC).

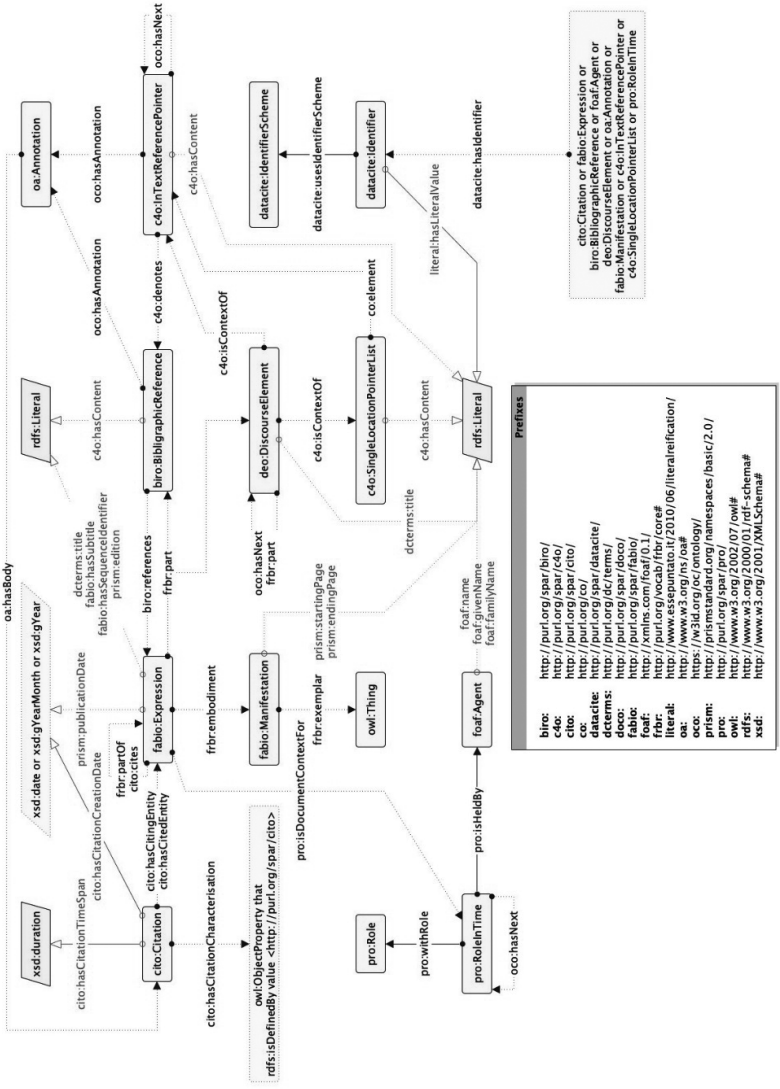
OpenCitations es un proyecto que se adhiere a los principios fundamentales de la ciencia abierta. Cumple con los principios de datos FAIR, los cuales manifiestan que los datos deben ser buscables,

Figura 15. Resultados de búsqueda en Share VDE



Fuente: https://share-vde.org/sharevde/searchNames?n_cluster_id=9786. Esta figura es para efectos ilustrativos;, para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKIvpX0suthJ4lZIhfCBHye5FzVA?e=IL9CZz.

Figura 16. Modelo de datos de OpenCitations



Fuente: https://opencitations.net/model. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKIvpX0suthJ4pSnnE7gNHdGGmaw?e=2Wf1m6.

accesibles, interoperables y reutilizables (OpenCitations 2021). Recientemente, en este proyecto se ha publicado una definición formal de una cita abierta y ha lanzado un sistema para identificadores globales únicos y persistentes para citas bibliográficas, el cual ha sido denominado Open Citation Identifiers (OCI).

En la figura 16 se aprecia el modelo de datos que es utilizado por OpenCitations. El Open Citation Data Model (OCDM) se utiliza para modelar todas las entidades bibliográficas y de citas (es decir, los rectángulos amarillos en la figura 16, que definen las clases de objetos que el modelo de datos permite describir), sus atributos (es decir, las flechas verdes) y las relaciones con otras entidades (es decir, las flechas azules).

Todos estos elementos se exponen en cualquier conjunto de datos de OpenCitation codificado en RDF. Esto permite la publicación de datos bibliográficos y de citas como datos abiertos enlazados, lo cual da legibilidad e interoperabilidad a los datos en el contexto de la web. El OCDM también puede ser empleado por terceros, ya sea para su propio uso o para estructurar sus datos para su envío y publicación por OpenCitations.

De acuerdo con Daquino y colegas (2020, 449), en los últimos años OpenCitations ha desarrollado otros conjuntos de datos, mientras que OCDM ha sido adoptado por proyectos externos y se ha ampliado para adaptarse a estos cambios. Recientemente, hemos ampliado aún más el modelo de datos de OpenCitations para acomodar los requisitos de metadatos extendidos del proyecto Open Biomedical Citations in Context Corpus (CCC).

De esta manera, los desarrollos presentados en este capítulo dan fe del avance lento pero consistente de la adopción de los principios de los datos abiertos enlazados en diversos contextos y proyectos informativos. En el siguiente capítulo, se abordarán los principios teóricos de la recuperación de información, con el propósito de obtener un marco general para la aplicación de los datos abiertos enlazados en este proceso.

Recuperación de información

CONCEPTUALIZACIÓN

El concepto de recuperación de información es muy amplio y forma parte de diferentes campos disciplinarios. Las disciplinas que mayormente se han dedicado al estudio de la RI son la informática, las ciencias de la computación y la bibliotecología y los estudios de la información. Históricamente el concepto de RI ha estado ligado al surgimiento de sistemas para la recuperación de información (SRI). Los SRI son estructuras integrales que permiten desde almacenar información, hasta propiciar su acceso por parte de los usuarios que interactúan con ellos.

Por lo tanto, no es posible entender el concepto de SRI si no se contempla su adaptación y funcionalidad en los sistemas que hacen posible su funcionalidad en diversos contextos de información. En este sentido, aunque la RI tradicionalmente se limitaba a la recuperación de documentos escritos, el término se redefinió para incorporar la creciente aparición de recursos digitales.

Así, los nuevos motores de búsqueda en Internet, que originalmente buscaban textos, expandieron su actividad a imágenes, audiovisuales y una amplia gama de recursos digitales. De esta forma términos como recuperación de textos, recuperación documental y recuperación de información son utilizados como equivalentes.

Entonces, las tecnologías de la información han tenido un notable impacto en la RI, al grado de adaptar su conceptualización acorde a las características del contexto en donde se implementa; a su vez, la fundamentación teórica de la RI ha permitido conocer sus alcances y limitaciones, lo cual ha sido factor para identificar su evolución en determinados periodos de tiempo o épocas.

En la actualidad la recuperación de información "convencional" significa la búsqueda *online* en bases de datos electrónicas, de forma interactiva y en tiempo real. Normalmente, esto implica que el usuario construya una estrategia de búsqueda usando términos con distintas relaciones lógicas (booleanas) y que el programa de búsqueda simplemente divida la base de datos en dos conjuntos: elementos recuperados y elementos no recuperados (Lancaster 2001, 213).

Este método convencional para recuperar información está influenciado por la demanda informativa del usuario, el cual manifiesta un comportamiento mediante una serie de interrogantes efectuadas al SRI con el propósito de obtener la información que desea y que le ayude a satisfacer su demanda informativa. Además de eso, el proceso de recuperación de información se encuentra fuertemente influenciado por la organización de la información que es aplicada a los recursos, pues es preciso señalar que esta organización se realiza tomando como base los atributos informativos, con la intención de generar registros que puedan ser almacenados en un determinado SRI.

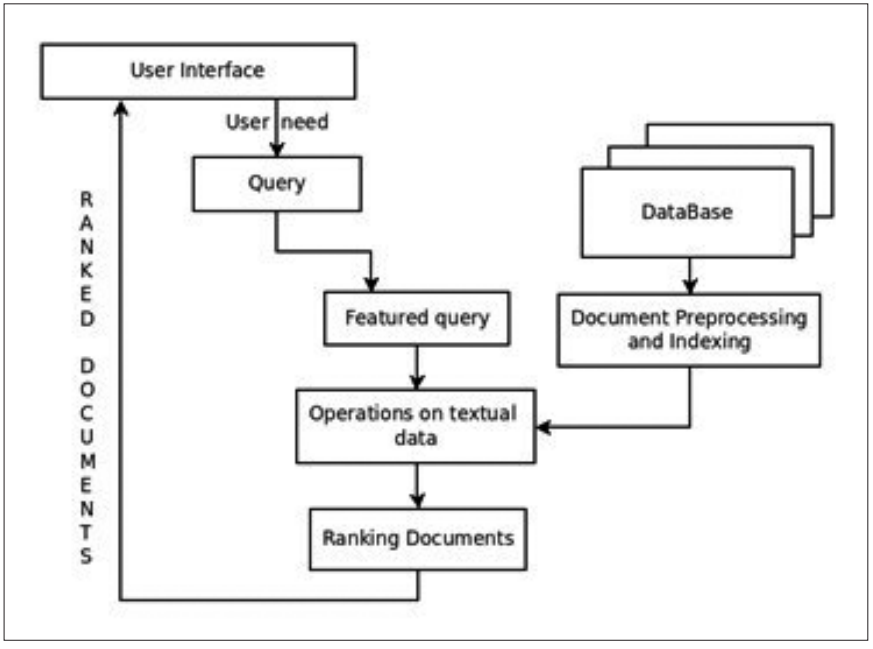
En la figura 14, puede apreciarse el proceso de recuperación de información propuesto por Gunjal (2016), el cual da comienzo con la interacción del usuario en la interfaz del SRI. Esta interacción está motivada por la necesidad informativa del usuario. En este sentido, Calva González (2011, 8) manifiesta que:

> [...] al surgir la necesidad de información en el sujeto es cuando ésta se transforma en un comportamiento, es decir presenta una acción que puede ser preguntarle a otra persona, acudir a la biblioteca más cercana, buscar en Google la información, etcétera; existen muchas manifestaciones o

comportamientos en el sujeto cuando se propone buscar la información que necesita para responder a la incógnita que se planteó.

Por lo tanto, el usuario realizará una serie de consultas en el SRI que contiene recursos de información almacenados y organizados. La consulta por excelencia dentro de estos sistemas se desarrolla a través de una operación textal, relacionada con el uso de un determinado lenguaje. En este sentido, el procesamiento del lenguaje natural (PLN) es una subdisciplina de la inteligencia artificial y la lingüística que tiene el propósito de estudiar las problemáticas derivadas de la generación y comprensión automática del lenguaje natural.

Figura 17. Proceso de recuperación de información



Fuente: https://amitgunjal.wordpress.com/2016/11/21/the-process-of-information-retrieval-from-scratch/.

En los últimos años, el PLN ha sido utilizado en el proceso de recuperación de la información tanto para facilitar la descripción del contenido de los documentos, como para representar la consulta formulada por el usuario; ello con el objetivo de comparar ambas descripciones y presentar al usuario aquellos documentos que satisfagan en mayor grado su necesidad de información.

De esta manera, un sistema de recuperación de información que emplea consultas textuales lleva a cabo las siguientes tareas para responder a las consultas del usuario (Vallez y Pedrazza 2007):

1. Indexación de la colección de documentos: en esta fase, mediante la aplicación de técnicas de NLP, se genera un índice que contiene las descripciones de los documentos. Normalmente, cada documento es descrito mediante el conjunto de términos que, hipotéticamente, mejor representa su contenido.
2. Cuando un usuario formula una consulta el sistema la analiza, y si es necesario la transforma con el fin de representar la necesidad de información del usuario del mismo modo que el contenido de los documentos.
3. El sistema compara la descripción de cada documento con la descripción de la consulta y presenta al usuario aquellos documentos cuyas descripciones más se asemejan a la descripción de su consulta.
4. Los resultados suelen ser mostrados en función de su relevancia, es decir, ordenados en función del grado de similitud entre las descripciones de los documentos y de la consulta.

De esta manera, el proceso de recuperación de información tendrá una gran relación con las habilidades informativas que el usuario posea y que puedan ayudarle a encontrar información de una manera más eficiente. Además, dichas habilidades estarán relacionadas con las descripciones de los recursos de información contenidas en el SRI, pues un usuario con habilidades informativas "[...] es capaz de reconocer sus necesidades de información, sabe

cómo localizar, acceder, recuperar, evaluar, organizar y utilizar la información" (Lau 2007, 12).

Por lo tanto, el proceso de RI no puede entenderse sin la presencia e interacción entre el usuario y el sistema que utiliza para satisfacer su demanda informativa. Además, es preciso tener en cuenta que las descripciones y los registros contenidos en dichos sistemas son el puente entre el usuario y el recurso de información que desea consultar.

En este momento, es conveniente valorar si la presencia y frecuencia de las palabras del propio recurso describen su contenido de una manera eficiente y asimilable por el propio usuario, pues se estima que en la RI debe existir un emparejamiento y compatibilidad entre los términos empleados por el usuario para recuperar información y por el autor al momento de crear un determinado recurso.

El usuario de un sistema de recuperación debe traducir su necesidad de información en una consulta en el idioma proporcionado por el sistema. Con un sistema de recuperación de información, esto normalmente implica especificar un conjunto de palabras que transmiten la semántica de la necesidad informativa.

Hace unos años, en los sistemas tradicionales de recuperación de información el proceso de búsqueda era intuitivo. La redundancia de resultados generados en una búsqueda específica realizada por el usuario se tomó como base para que el proceso de recuperación pudiera repetirse utilizando patrones de búsqueda repetitivos generados por el usuario al momento de interactuar con el SRI. Sin embargo, dado que los usuarios pueden manifestar diversos comportamientos informativos, los especialistas en el estudio de la RI comenzaron a centrarse en el análisis del contexto de la RI.

El contexto de un proceso de RI se puede inferir de muchas formas diferentes, esto en dependencia a la subjetividad del propio contexto y de sus usuarios. Pues "[...] la información no es sin el sujeto que la utiliza: su contenido existe hasta que alguien la lee, analiza, sintetiza, asimila y lo incorpora a su acervo de conocimientos" (Hernández 2017, s.p.).

De esta manera, la recuperación de la información no puede concebirse sin el usuario que hace posible este proceso dentro de los SRI. Pues en esencia, el usuario determina la efectividad de la RI en cierto contexto y como puente para satisfacer una demanda informativa.

El objetivo de la RI es proporcionarle información relevante al usuario para satisfacer su necesidad informativa. La RI ha evolucionado acorde a los desarrollos tecnológicos, sobre todo aquellos que tienen que ver con la informática y las ciencias de la computación, trayendo consigo la aparición de nuevos métodos para recuperar información.

Al respecto, Baeza y Ribeiro (2011, 2) afirman que:

> [...] en los últimos 20 años, el área de recuperación de información ha crecido mucho más allá de sus objetivos principales de indexar texto y buscar documentos útiles en una colección. Hoy en día, la investigación en RI incluye modelado, clasificación de documentos y categorización, arquitectura de sistemas, interfaces de usuario, visualización de datos, filtrado, idiomas, etc.

Por lo tanto, la RI, más que un concepto, es susceptible de convertirse en un fenómeno de investigación que manifiesta diferentes comportamientos derivados del avance de la tecnología y del surgimiento de nuevas necesidades informativas por parte del usuario. De hecho, la propia demanda informativa del usuario provocará el surgimiento de nuevas alternativas para satisfacer sus necesidades. Es en este ámbito en donde la RI tiene una fuerte injerencia al momento de formular nuevos modelos para la obtención de información.

De esta manera, la recuperación de información es un proceso que tiene el propósito de brindar información a un determinado usuario. La eficacia de este proceso dependerá de las habilidades del propio usuario para interactuar de una manera eficiente a través de un sistema que permita realizar consultas de información.

Entonces, la recuperación de información es un proceso que es aplicado dentro un determinado sistema con el fin de obtener

información. La recuperación incluye todos los procedimientos utilizados para identificar, buscar, encontrar y localizar la información que es útil para resolver la demanda informativa del usuario.

De acuerdo con Sammut y Web (2017, 671), la recuperación de información engloba un conjunto de técnicas que extraen de una colección de documentos aquellas que son relevantes para una consulta determinada. Inicialmente atendiendo las necesidades de bibliotecarios y especialistas, el campo ha evolucionado dramáticamente con la llegada de la Word Wide Web.

La RI es un proceso más general que la recuperación de datos, cuyo propósito es determinar qué documentos contienen apariciones de las palabras clave que componen una consulta. Mientras que la sintaxis y la semántica de los marcos de recuperación de datos está estrictamente definida con consultas expresadas en un lenguaje totalmente formalizado, las palabras de un lenguaje natural sin estructura o con una estructura limitada son el medio de comunicación para obtener información de un determinado sistema.

En este sentido, una tarea crucial para un sistema de RI es indexar la colección de documentos para que su contenido sea accesible de manera eficiente. Los documentos recuperados por el sistema generalmente se clasifican según la relevancia esperada, y el usuario que examina algunos de ellos podría proporcionar retroalimentación para que la consulta pueda ser reformulada y los resultados mejorados.

La recuperación de información es un medio por el que los usuarios de un sistema o servicio de información pueden encontrar los documentos, registros, imágenes gráficas o registros de sonido que satisfagan sus necesidades o intereses (Meadows *et al.* 1999, 2). Entonces la RI implica encontrar información en un sistema, catálogo o base de datos. No obstante, esta recuperación ha cambiado y evolucionado acorde a las características de las tecnologías de la información actuales. El elemento más significativo de este cambio han sido la aparición de Internet y el desarrollo de la web.

En este sentido, la naturaleza del entorno digital actual sugiere una transformación de los fenómenos que impactan en la RI. Esto

da como resultado no solo la aparición de nuevos problemas técnicos relacionados con la eficiencia computacional de implementaciones de sistemas de recuperación, sino también en la necesidad de desarrollar nuevos métodos de recuperación destinados a mejorar la calidad de los resultados de búsqueda.

Por ejemplo, actualmente se requieren interfaces que sean capaces de realizar consultas complejas de información adaptando estrategias y métodos de visualización para identificar patrones particulares existentes entre abismales cantidades de información. De acuerdo con Mirel (1999, 1), para analizar visualmente la información mediante grafos interactivos, los usuarios requieren de nuevas habilidades, pues tienen que adaptarse a nuevas convenciones y declaraciones de consulta escritas en un nuevo paradigma que presenta características interactivas poco convencionales.

Así pues, los usuarios seleccionan datos e informaciones directamente de grafos interactivos que se encuentran vinculados dinámicamente. Por lo tanto, en este nuevo paradigma, la búsqueda, la recuperación y el análisis ocurren casi simultáneamente. De esta manera, una de las áreas más importantes de competencia del usuario en la consulta visual es saber seleccionar (buscar y recuperar) los datos e informaciones correctas para el propósito correcto.

Por consiguiente, en los modernos procesos de RI se pueden utilizar técnicas de visualización de información para consultar y acceder a los atributos complejos de la información, lo que permite a los buscadores ver vistas previas o descripciones generales de la nueva información que podrían obtener al ver determinados resultados de búsqueda.

Por ejemplo, es posible medir el nuevo conocimiento en cada uno de los resultados de la búsqueda y presentarlos visualmente dentro de la lista de los resultados de búsqueda; o cuando se emiten consultas complejas, se pueden utilizar técnicas de visualización para mostrar las relaciones entre la búsqueda, resultados y diferentes aspectos de la consulta.

Un elemento central de los sistemas modernos de recuperación de información es el índice de documentos. El índice es un conjunto de estructuras de datos que se construyen a partir de una

colección de documentos fuente con el objetivo de permitir que un sistema de recuperación de información proporcione una respuesta oportuna y eficiente a las consultas de búsqueda.

El proceso de creación de índices generalmente implica leer y procesar la colección de documentos de origen, analizar el texto en cada documento individual y extraer las características necesarias para recuperar y clasificar ese documento en respuesta a una consulta específica.

Además, los sistemas de indexación a menudo utilizan la reducción de dimensiones, la compresión y otras técnicas relacionadas para reducir drásticamente los rastros de almacenamiento de la colección de origen en su forma indexada. Los índices de documentos se almacenan con frecuencia en un conjunto de estructuras de archivos que permiten una rápida recuperación y clasificación por parte de un sistema de recuperación de información en respuesta a una determinada consulta.

De acuerdo con Amati (2018), la información recuperada de los sistemas de RI puede variar de una lista ordenada de elementos textuales relevantes de cualquier tipo, como documentos completos o sus extractos, o puede ordenarse en formas más elaboradas, como resúmenes de documentos o respuestas específicas a preguntas.

Sin embargo, toda consulta desarrollada en un proceso de RI remitirá a un recurso de información documental, el cual puede ser textual, multimedia o digital. Los datos contenidos en estos recursos son representados en registros descriptivos que tienen la capacidad de integrar la consulta del usuario con los atributos representados en dichos recursos.

Además, en la RI un usuario con una necesidad particular genera una consulta simple y espontánea y desea una respuesta rápida y que responda a su demanda informativa. Por el contrario, en las modernas opciones de filtrado de información disponibles actualmente en algunos sistemas, debido a que el tema de la búsqueda es permanente, el usuario tiene tiempo para elaborar y afinar la consulta.

De esta manera, la recuperación de información se refiere a un proceso de búsqueda, exploración y descubrimiento de información

en diferentes SRI organizados para satisfacer las necesidades de información de los usuarios. En este sentido, Zang (2008, 4) afirma que "[...] la recuperación de información contiene dos componentes fundamentales: recuperación de información y organización de la información. Dependen el uno del otro como las dos caras de una moneda. Simplemente no se puede hablar de un término e ignorar el otro".

Desde la perspectiva de los usuarios, la organización de la información en un sistema de recuperación de información es un proceso interno. Aunque la organización de la información es esencial para su recuperación, puede ser invisible y no transparente para los usuarios. No obstante, desde la perspectiva del sistema, la organización de la información es indispensable, el método de organización de la información afecta y determina la forma y el método de su recuperación.

Además, la necesidad de desarrollar novedosos sistemas para recuperar información se produce cuando una colección alcanza un tamaño en el que las técnicas tradicionales de catalogación ya no pueden hacer frente. Debido a esto es latente la evolución de los principios para organizar y representar la información que se encuentra disponible en el contexto de las unidades de información, pues si se pretende que la información y los datos de estas unidades puedan conectarse con el ciberespacio, será necesario dotar de flexibilidad estos principios.

Esto nos lleva a considerar que las aplicaciones de búsqueda y recuperación de información evolucionarán a medida que cambia el universo de información. Además, los dispositivos que hacen posible recuperar información, también han ido evolucionado acorde a las características de dicho universo. El ejemplo inmediato de este tipo de cambio es el rápido crecimiento de los dispositivos móviles. De hecho, en la actualidad los SRI se desarrollan contemplado la adaptación de las búsquedas de información mediante el uso de un dispositivo móvil.

Por otra parte, la ubicuidad en la búsqueda de información a través de la web ha hecho posible acceder instantáneamente a cientos de terabytes de páginas web, videoclips, noticias, imágenes,

redes sociales, libros escaneados, trabajos académicos, música, televisión, programas y películas, casi siempre a través de motores de búsqueda. En los últimos años, el acceso también ha sido posible desde un teléfono móvil.

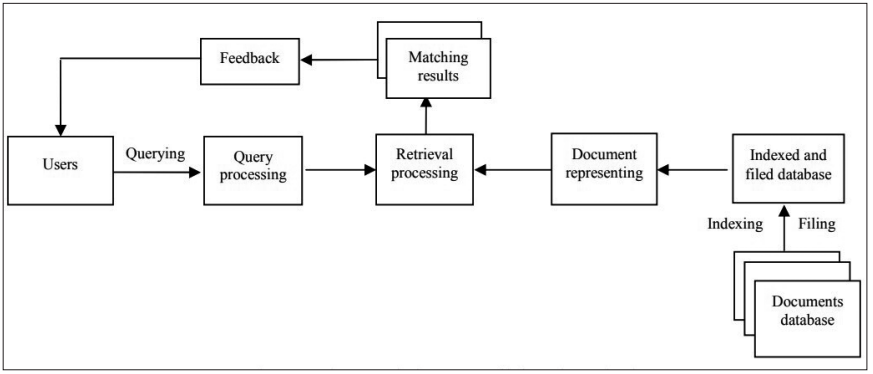
Por lo tanto, la conceptualización de la recuperación de información en tiempos modernos no puede concebirse sin la presencia de un entorno digital complejo, caracterizado por la exacerbada acumulación de datos e información de múltiples y variadas características y tipologías.

Bajo esta premisa, la recuperación de información es un proceso complejo que consiste en una serie de etapas. Surge a partir de una demanda informativa por parte del usuario, el cual canaliza su necesidad a un SRI mediante un proceso de interacción a través de una interfaz de búsqueda mediante un dispositivo computacional o móvil. El puente que conecta la necesidad informativa del usuario con el SRI es el lenguaje de consulta que está basado en palabras que representan la información contenida en dicho sistema y en los recursos de información que lo conforman. En la figura 18 se ejemplifica la estructura lógica de un SRI, en el cual intervienen las etapas básicas del proceso de RI y la respectiva interacción del usuario.

De esta manera, el usuario recupera una serie de registros que representan los recursos de información, sobre todo porque estos registros contienen las palabras o términos que el usuario utilizó durante el proceso de búsqueda. No obstante, este proceso de interacción entre el usuario, el registro y el recurso no asegura la satisfacción de la necesidad informativa del individuo.

Este fenómeno ha despertado el interés por investigar los diferentes tipos de recuperación de información, con un notable énfasis en los modelos disponibles para llevar a cabo este proceso. En el contexto de la RI existen múltiples modelos y, en la mayoría de los casos, estos modelos remiten a diferentes tipos de SRI, pues "[...] diferentes métodos de representación de documentos provocarán el diseño de diferentes procesamientos de consultas y la coincidencia de similitudes producirán diferentes modelos de recuperación de información" (Hua 2009, 441). Por lo tanto, en el

Figura 18. Estructura lógica del sistema de recuperación de información



Fuente: Jing Hua 2009,144.

siguiente apartado se abordan los modelos de SRI más comunes en el contexto de la información documental.

MODELOS

La recuperación de la información vista como un campo de investigación ha dado la pauta para el desarrollo de diferentes modelos para recuperar información a través del tiempo. Recordemos que estos modelos son parte del propio desarrollo de la RI en diferentes sistemas, contextos y épocas en donde la información ha presentado determinados comportamientos y características. En este sentido:

> Los sistemas de recuperación de la información también han ido evolucionando con el fin de adaptarse a este nuevo entorno; de hecho, se han desarrollado algunos de los sistemas más innovadores, al mismo tiempo que extensos, por no hablar de su popularidad, aunque aún no disponemos de metodologías suficientemente consolidadas que evalúen su efectividad (Martínez 2004, 7).

Los modelos para recuperar información han sido desarrollados con el objetivo de medir la efectividad de los sistemas para recuperar información; sin embargo, muchos de éstos no contemplan en sus variables la demanda informativa del usuario, lo cual sigue siendo un reto para medir la satisfacción del usuario en función de un determinado sistema utilizado.

La historia de la recuperación de información ha evolucionado en paralelo al desarrollo de modelos de RI. En general, un modelo de RI se identifica principalmente con su función de recuperación que se emplea para clasificar documentos, pues lo que importa en los modelos de RI es la eficacia de la recuperación.

En los primeros días, la indexación automática fue el foco principal de la investigación acerca de la recuperación de información con el objetivo de ayudar a la clasificación manual realizada por los bibliotecarios. Los primeros trabajos de recuperación de información se referían a la construcción de métodos efectivos para la selección de palabras clave que ayudarán a representar sucintamente los recursos de información.

De esta manera, surgieron estudios que abordaban la coincidencia entre dichos recursos y las consultas que realizaban los usuarios de manera más simple y mediante una búsqueda booleana dentro de los sistemas para la recuperación de información. Así pues, dentro de la literatura especializada se mencionan los modelos de recuperación de información comúnmente adoptados en los SRI.

En la tabla 2 puede apreciarse una clasificación de modelos de recuperación de información desarrollada desde una perspectiva clásica acorde a las tendencias de la RI de finales del siglo XX. Notablemente los modelos propuestos por Dominich han evolucionado en la actualidad, debido en parte a nuevos comportamientos en la información y la manera de buscar información por parte del usuario.

Los modelos basados en inteligencia artificial son ampliamente utilizados actualmente en respuesta a las nuevas características del universo de información, pues tratan de representar las complejidades de la información acorde a la interpretación de

las necesidades del usuario. La sistematización de ambos elementos provocará la aparición de nuevas metodologías para acercar al usuario a la información que realmente ayude a satisfacer su demanda informativa.

Sin embargo, de acuerdo con Sparck Jones:

> La presunción de que un sistema de inteligencia artificial con una base de conocimientos integral sería superior a un archivo de documentos pasa por alto el punto de que los textos de los documentos individuales tienen su propio valor como relatos de la información que sus autores quieren transmitir. Por lo tanto, desmenuzar documentos para construir bases de conocimiento y, en el proceso, purificar su lenguaje, pierde el elemento crítico de quién dijo qué y cómo lo dijo (Jones 1999, 260).

Los modelos de RI fundamentados en los principios de la inteligencia artificial deberán proporcionar una estructura de conocimiento que permita interpretar las necesidades informativas de

Tabla 2. **Clasificación de los Modelos de Recuperación de Información según Dominich**

Modelo	Descripción
Modelos clásicos	Incluye los tres más comúnmente citados: booleano, espacio vectorial y probabilístico.
Modelos alternativos	Están basados en la lógica difusa.
Modelos lógicos	Basados en la lógica formal. La recuperación de información es un proceso inferencial.
Modelos basados en la interactividad	Incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la relevancia de los documentos recuperados.
Modelos basados en la inteligencia artificial	Dominios de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Fuente: Dominich, S. (2000).

los individuos de una manera integral y acorde al contexto informativo en donde la necesidad se manifieste.

El siglo XX y principios del XXI fueron transformadores en la forma en que las personas deseaban acceder a la información. A principios del siglo XX, una persona con necesidad de información probablemente iría a una biblioteca local y utilizaría un catálogo de fichas, localizaría libros o documentos que, con suerte, respondieran a esa necesidad.

Debido a la relativa inconveniencia de acceder a la información de esa manera, lo más probable es que esa persona sólo busque responder a una pequeña cantidad de preguntas. El alcance de la información disponible para las personas estaría limitado por el tamaño de su biblioteca; para un número reducido de necesidades muy importantes, se podría haber concertado un préstamo entre bibliotecas.

Puesto que los sistemas que son accesibles hoy en día son tan fáciles de usar, es tentador pensar que la tecnología detrás de ellos es igualmente sencilla de construir. No obstante, el camino hacia la creación de sistemas de recuperación de información exitosos requirió mucha innovación y reflexión durante un largo periodo de tiempo. Al considerar posibles direcciones futuras, es preciso concebir que los modelos de RI seguirán evolucionando a la par del desarrollo de las tecnologías de la información, sobre todo aquellas que se relacionan con la generación de sistemas, aplicaciones y algoritmos.

El descubrimiento intensivo de datos sobre todo en el área científica ha propiciado la generación de modernos sistemas para recuperar información, los cuales han adoptado diversos modelos de RI, sobre todo aquellos que hacen énfasis en la visualización de enormes cantidades de datos en un espacio común y concentrado. Este tipo de modelos fomentan en mayor grado la consulta de patrones de comportamiento entre los datos y las relaciones que se establecen entre ellos. Por lo tanto, en este contexto también ha sido necesario definir nuevas metodologías y herramientas computacionales que respalden el desarrollo de los modelos para recuperar información.

Por ejemplo, la infraestructura de tecnología de la información de hace 25 años era adecuada para la cultura científica de esa época. Los volúmenes de datos eran relativamente pequeños y, por lo tanto, cada conjunto de datos era muy preciso y sobre todo estructurado. Los sistemas de TI eran relativamente caros y accesibles solo para expertos.

En este sentido, el flujo de trabajo fundamental se basaba en una recopilación de datos por parte del sistema (por ejemplo, un laboratorio o un sensor de campo); de esta manera, el sistema almacenaba datos, los procesaba y los dejaba listos para su posterior análisis. En la actualidad, los modelos que sirven de estructura para estos sistemas también deben contar con la capacidad para analizar a los datos, visualizarlos y hacerlos accesibles a las comunidades de usuarios que requieran buscarlos y recuperarlos. Por lo tanto, los modelos de RI deben desarrollarse de manera integral y cubrir todas las fases del ciclo de vida de la información.

Además de eso, la visualización de información es un aspecto central que cualquier modelo de RI contemporáneo debe adoptar en sus estructuras, sobre todo cuando estos modelos son adoptados en sistemas de contextos científicos, pues el análisis visual de los datos, facilitado por interfaces interactivas, permite la detección y validación de los resultados esperados, al tiempo que posibilita descubrimientos científicos inesperados.

Esto permite la validación de nuevos modelos teóricos, proporciona una metodología para la comparación entre modelos y conjuntos de datos, y permite consultas cuantitativas y cualitativas con el propósito de mejorar la interpretación de los datos y facilitar la toma de decisiones.

De acuerdo con Rodríguez y Pinto (2018, 53), la importancia de la información en la toma de decisiones viene dada porque "[...] una organización usa información estratégicamente para percibir los cambios de su ambiente, crear nuevo conocimiento para innovar y tomar decisiones acerca de sus cursos de acción". Por lo tanto, la información es un elemento fundamental, pues dentro de cualquier organización es necesario contar con sistemas que permitan a los individuos realizar consultas efectivas de información de una manera integral y en respuesta a su demanda informativa.

De esta manera, en la actualidad podemos diferenciar modelos para la recuperación de información, los cuales responden a necesidades contextuales específicas y a tipos de información muy diversos.

En la figura 19 pueden apreciarse los modelos de RI de orden contemporáneo, los cuales están caracterizados por recibir una fuerte influencia de la teoría matemática y lógica, pues recordemos que la recuperación de la información tiene en su haber una fuerte incidencia de los campos lógico-matemáticos que le permite su implementación en un determinado sistema de recuperación de información.

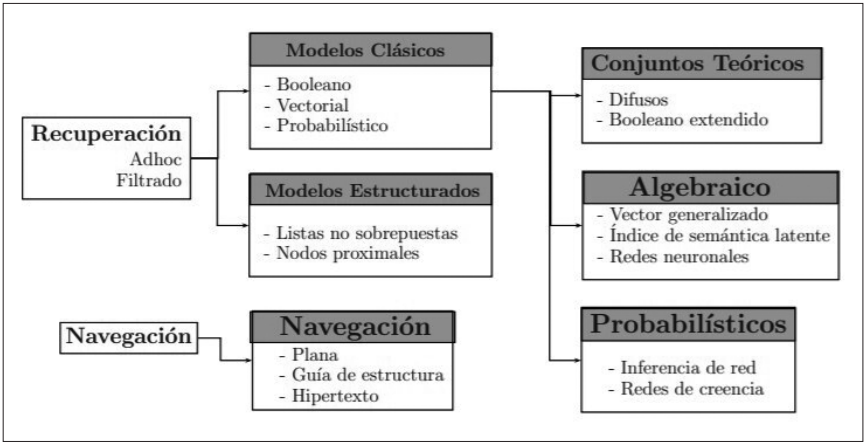
Bajo esta premisa, los modelos probabilísticos requieren de un conjunto de datos de la interacción entre el usuario y los recursos que se evalúan como relevantes para un determinado número de consultas. A su vez, los modelos algebraicos asumen que tanto las consultas como los términos están representados por vectores y que la similitud entre consultas y recursos se obtiene mediante una normalización de dichos vectores.

En los modelos lógicos, las consultas y los términos se representan mediante proposiciones, y la dependencia entre consultas y recursos viene dada por un operador de vinculación. Los modelos teóricos de la información se basan en la noción de codificar el peso de los términos en los recursos, siendo los documentos más informativos los generados por las configuraciones de términos menos probables.

Los modelos expuestos en esta representación han sido abordados con anterioridad por la corriente teórica de la RI desde la perspectiva computacional hasta llegar a considerar las posturas de las ciencias de la información. En este sentido, un modelo de RI tiene en su carácter conceptual componentes de ambos contextos disciplinarios, pues por un lado contemplan el uso de un sistema computacional para resolver la necesidad de un individuo con perfiles y características determinadas, y por otro, considera los lenguajes y vocabularios necesarios para representar y describir los recursos en dicho modelo.

De esta forma, un modelo de recuperación de información selecciona y clasifica los recursos de información que son relevantes

Figura 19. Clasificación de los modelos de recuperación de la información según Jaimes y Vega



Fuente: Jaimes y Vega 2005.

con respecto a la consulta de un usuario. Los textos de los documentos y las consultas se representan de la misma manera, por lo que la selección y clasificación de documentos se puede formalizar mediante una función de coincidencia que devuelve un valor de estado de recuperación para cada recurso que forma parte de una colección.

Así pues, la mayoría de los sistemas de recuperación de información representan el contenido del documento mediante un conjunto de descriptores, denominados *términos*, que pertenecen a un vocabulario documental que forma parte de una determinada disciplina; aunado a ello, la recuperación de información ha recibido una fuerte influencia de las tecnologías digitales.

De acuerdo con Martínez Comeche (2006, 29), la recuperación de información ha cobrado un gran auge debido al crecimiento de Internet, pues ha tratado de facilitar la tarea de discernimiento de los escasos documentos relevantes que puedan existir en la red frente a los millones de documentos irrelevantes en relación con cada consulta formulada en la red.

Dado que esta inmensa "colección" carece por completo de organización, la automatización de los procesos de análisis y recuperación de los billones de documentos que configuran la red se ha convertido en una tarea de importancia capital. Por lo tanto, la RI busca no solo obtener recursos de información, sino que dichos recursos sean efectivos para la demanda informativa del usuario; es por ello que en la actualidad los métodos semánticos para recuperar información han tenido un gran repunte sobre todo en los contextos científicos.

De acuerdo con Fox y Hendler (2014, 160), la influencia de la comunidad dedicada a la inteligencia artificial y la creciente cantidad de datos disponibles en la web (lo cual ha llevado a muchos científicos a utilizar la web como su "equipo de cómputo" principal) han conducido a los investigadores de la web semántica a enfocarse tanto en cuestiones formales de los lenguajes de representación semántica, como en el desarrollo de aplicaciones semánticas de propósito general. Estos lenguajes se encuentran en una constante estandarización y las diversas comunidades están recurriendo a ellos para adaptarlos en diversos sistemas de recuperación mediante el uso de esquemas de representación como las ontologías y los datos abiertos enlazados.

Aunado a ello, las tecnologías de la web semántica son utilizadas para propiciar el procesamiento de información de forma automática mediante la utilización de métodos y algoritmos de inteligencia. Con esto se pretende comprender la demanda informativa expresada por el usuario en una determinada consulta y dotar la recuperación de un significado, identificando y brindando información confiable.

Para realizar la búsqueda semántica se emplean buscadores semánticos que son "[...] sistemas de recuperación de información que entienden la necesidad del usuario y analizan la información disponible en la web mediante el uso de algoritmos que simulan comprensión o entendimiento" (Viltres *et al.* 2018, 105). De esta manera, los modelos para recuperar información se encuentran cada vez más ligados al uso de métodos y técnicas de inteligencia artificial, pues actualmente no basta solo con recuperar recursos

de información tradicionales y textuales, sino que el amplio espectro de estos recursos ha sido transformado por los elementos multimedia y audiovisuales que convergen en la actualidad en el desarrollo de dichos recursos.

Con la popularidad de la tecnología multimedia, los contenidos de la web han cambiado de manera versátil en los últimos años. Estos contenidos, en la mayoría de los casos, se relacionan con recursos de información que han sido utilizados para su desarrollo y colocación en diferentes fuentes del ciberespacio.

Recientemente los sitios web como los de comercio electrónico y los de centros comerciales manejan mucha información basada en imágenes. Para encontrar una imagen específica de estas fuentes de imágenes, generalmente se utilizan motores de bases de datos de imágenes o motores de búsqueda web. Pero las capacidades de recuperación basadas en las características de estos sistemas son bastante limitadas, especialmente para las imágenes web.

Al navegar por la web con una colección tan amplia de documentos multimedia vinculados, los usuarios pueden perderse fácilmente en sus profundidades. En este sentido, los modelos de recuperación multimedia también plantean a los usuarios solucionar problemas para encontrar los recursos adecuados y extraer información de los documentos multimedia.

Las bases de datos de texto y relacionales se pueden buscar por contenido y términos de indexación. Sin embargo, para encontrar información en imágenes, video y voz, el usuario depende del alcance de la descripción semántica del recurso asignado por el indexador de la base de datos. En este sentido, es necesario identificar los métodos de búsqueda de los usuarios para desarrollar la tecnología que utilizarán. Y, por supuesto, la web debe volverse mucho más inteligente si quiere optimizar su propio rendimiento, así como su núcleo de conocimiento para responder a preguntas cada vez más complejas.

Algunos usuarios saben lo que buscan y tratan de satisfacer sus necesidades siguiendo los enlaces correspondientes. Estos usuarios pueden o no encontrar algo de interés, pero pueden fácilmente pasar por alto otros documentos más relevantes lejos de sus rutas de navegación actuales.

De esta manera, los modelos de RI también contemplan el uso de agentes inteligentes. Un agente en la web puede describirse como un programa que recopila información o realiza algún otro servicio sin su presencia inmediata y con un horario regular. Por lo general, un agente utiliza parámetros proporcionados por el usuario, busca en todo o en parte de la web para recopilar la información que le interesa.

Uno de los roles más importantes de los agentes de recuperación de información es buscar y filtrar información de fuentes web distribuidas. Por lo tanto, comprender y desarrollar el comportamiento correcto de búsqueda de información para la recuperación de información es un desafío. Por ello, los modelos actuales de RI se concentran en generar un enfoque holístico para recuperar amplios fragmentos de información distribuidos en diversos puntos de la web y que estos sean compatibles con los patrones de búsqueda manifestados por los usuarios.

De acuerdo con Skilton y Hovsepian (2018, 101), un agente es un sistema informático que se encuentra en algún entorno, y que es capaz de realizar acciones autónomas en este entorno con el fin de cumplir con sus objetivos delegados. El uso de agentes dentro de los modelos de RI puede apreciarse con cierta claridad en el desarrollo de modelos de datos enlazados y ontologías contextuales.

Estos agentes son un elemento importante al implementar un determinado sistema de información que persiga el propósito de recuperar información mediante el uso de fuentes distribuidas en el ambiente web. Los datos abiertos enlazados hacen uso de agentes semánticos para propiciar la vinculación de datos disponibles en diversas fuentes con la intención de recuperar recursos, contenidos y datos de atributos similares.

Los agentes semánticos operan en la web semántica. A su vez, la web semántica opera sobre el modelo orientado a objetos de clases y propiedades, cada uno con sus propios atributos e instrucciones. Es una extensión de la web actual en la que a la información se le da un significado bien definido.

Un agente semántico introduce el concepto de ontología. La ontología es un medio de describir información. Es un conjunto de descriptores que incluye el vocabulario, las interconexiones

semánticas y algunas reglas simples de inferencia y lógica. Las ontologías permiten definir con precisión la información en la web. Esto permite que las computadoras, utilizando agentes, devuelvan un conjunto de resultados más significativo al usuario.

Para que un sistema de este tipo funcione, se requiere un medio de comprensión del lenguaje natural. Esto se realiza a grandes rasgos en tres etapas de análisis: sintaxis, semántica y pragmática. El análisis de sintaxis se ocupa de la estructura de una oración en términos de las posiciones relativas de las palabras y sus partes del discurso.

A su vez, el análisis semántico examina el significado de las palabras y comienza a construir una representación interna del significado de la oración. Esta tarea no se puede completar sin la pragmática; es decir, el conocimiento sobre el dominio de la discusión. Para ello, se necesita una comprensión de la pragmática para resolver la incertidumbre y completar el conocimiento asumido sobre el dominio.

Así pues, en la década de 1990 la recuperación de información ha experimentado un cambio de paradigma. Los sistemas de razonamiento aproximado abrieron la puerta a componentes de valor agregado más inteligentes. La gran cantidad de documentos de texto disponibles en bases de datos y en fuentes de información arbitradas de la web ha llevado a una demanda de métodos inteligentes en la recuperación de textos y a una investigación considerable en esta área.

La necesidad de un mejor preprocesamiento para extraer conocimiento de los datos se ha convertido en una forma importante de mejorar los sistemas. Los enfoques estándar prometen peores resultados que los sistemas adaptados a los usuarios, el dominio y las necesidades de información. Hoy, la mayoría de las técnicas desarrolladas en IA se han aplicado a sistemas de recuperación con más o menos éxito. Cuando los datos de los usuarios están disponibles, los sistemas suelen utilizar el aprendizaje automático para optimizar sus resultados.

Así pues, en este apartado se han abordado los modelos para recuperar información desde un punto de vista integrador; es

decir, considerando los factores tecnológicos que han incidido en su desarrollo y evolución. Notablemente estos dos elementos están relacionados con la proliferación de tecnologías digitales como la web semántica, lo cual ha influido notablemente en el surgimiento de métodos, herramientas y sistemas para organizar, buscar, recuperar y acceder a la información.

A su vez, no es posible entender la función de los modelos abordados en este apartado, sino se contempla que su adaptación y propósito quedan definidos en la presencia de los sistemas de recuperación de información. Por lo tanto, en el siguiente apartado se abordan los conceptos, atributos y tipos de SRI disponibles actualmente en el contexto de la información.

Sistemas para la recuperación de información

Los SRI forman parte integral del desarrollo y la evolución de la recuperación de información, ya que este proceso no puede concebirse sin herramientas digitales que propicien las tareas de almacenamiento, búsqueda, recuperación y acceso a los recursos de información.

Los SRI son un tipo de sistema de información fundamentados en el uso de bases de datos que les permiten realizar consultas en los registros de información que almacenan, con formatos de codificación específicos. Un SRI debe soportar una serie de operaciones básicas sobre los recursos de información que son almacenados; por ejemplo, introducción de nuevos registros de recursos, modificación de los que ya se encuentran almacenados y eliminación de éstos.

Un SRI puede definirse como una herramienta digital conformada por módulos y componentes integrales que permiten el almacenamiento, la búsqueda, la recuperación y el acceso a la información que está registrada en recursos de información análogos y digitales. Estos sistemas permiten realizar consultas y ecuaciones de búsqueda para recuperar la información contenida en los recursos. De acuerdo con Pinto (2018, s.p.), un SRI tiene dos componentes esenciales:

- Documentos estructurados (recursos de información). Es necesario establecer un proceso donde se establezcan herramientas de indización y control terminológico.
- Bases de datos donde estén almacenados los documentos. Definir lenguajes de interrogación y operadores que soportarán la base de datos y establecer qué tipo de ecuaciones serán permitidas.

De esta manera, los recursos de información antes de ser ingresados al SRI deberán ser analizados bajo los principios de la organización de la información y del análisis documental, pues estos permiten estructurar los registros de los recursos que serán representados en dicho sistema.

Por otra parte, la base de datos es el soporte que permitirá al SRI realizar consultas y ecuaciones de búsqueda mediante la interacción con el usuario. Es preciso recordar que dicha interacción estará fundamentada en la necesidad informativa del individuo, la cual será trasladada en la formulación de una consulta o ecuación específica.

En la figura 20 puede apreciarse un esquema que representa las operaciones desarrolladas por el SRI para recuperar información. El cálculo de similitud permite contrastar la demanda informativa del usuario con la obtención de recursos de información que sean compatibles con los patrones de la consulta o búsqueda. Esto permite obtener un listado con los recursos que posiblemente podrían ayudar a satisfacer la necesidad informativa del usuario.

La necesidad de información de un usuario en particular puede satisfacerse mejor si existe algún conocimiento sobre las necesidades específicas del usuario, sus habilidades y su interacción a corto y largo plazo. Ése es el campo de los sistemas de información personalizados que explotan los perfiles de usuario. Un perfil de usuario (o modelo de usuario) es un conocimiento almacenado sobre un usuario en particular. El perfil generalmente consiste en palabras clave que describen el área de interés del usuario desde su primera interacción con el sistema.

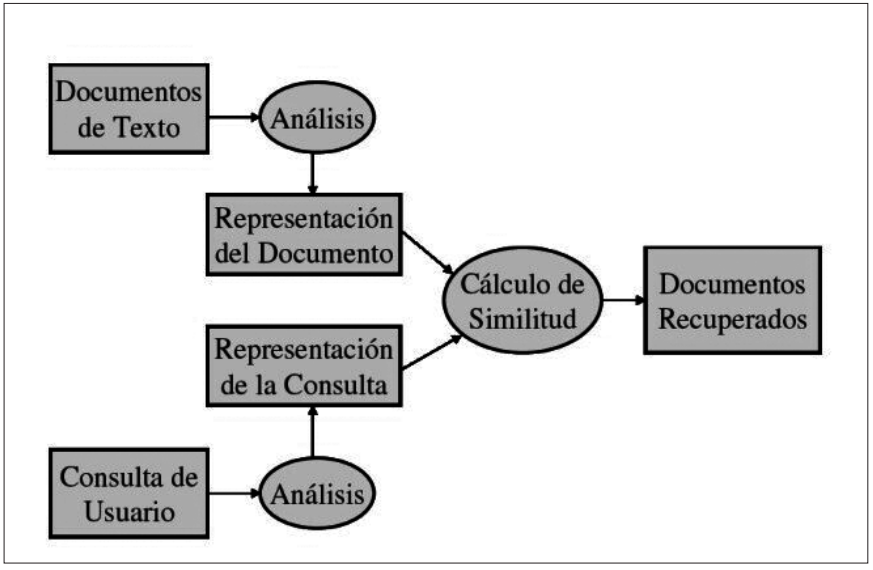
No obstante, el análisis de estos perfiles no asegura en su totalidad la satisfacción de la demanda informativa del usuario, pues

las operaciones para recuperar información presentan problemáticas que interfieren con la compatibilidad de los componentes de la consulta y el listado de resultados obtenidos como parte de un proceso de recuperación de información.

Los problemas relativos a la satisfacción de la demanda informativa en sistemas de recuperación de información son analizados tomando en cuenta las principales limitaciones de la recuperación de información como la indexación de documentos, la evaluación de las consultas y la evaluación del propio sistema.

Entonces, el sistema de recuperación de información también consta de dos componentes: el sistema de indexación y el sistema de consulta. La primera de ellas se encarga de analizar los recursos almacenados y con la creación de índices que luego permiten realizar consultas de búsqueda, mientras que el segundo es la interfaz visible del motor de búsqueda; es decir, la parte con la que interactúan los usuarios.

Figura 20. Operaciones para la recuperación de información



Fuente: López Herrera 2006.

Si un motor de búsqueda es capaz de responder preguntas en espacios de tiempo tan cortos a los que el usuario se ha acostumbrado (típicamente fracciones de segundo), es porque no exploran el entorno de datos en su totalidad en tiempo real (es decir, cómo y cuándo se realiza la consulta), sino que utilizan un índice que se actualiza periódicamente (varias veces al día).

En este sentido, la indexación de documentos tiene el objetivo principal de encontrar significados importantes y crear una representación interna de los recursos que son almacenados en un determinado sistema. Los factores por considerar en este rubro son la precisión para establecer la semántica, la exhaustividad y la facilidad para representar con exactitud el contenido intelectual del recurso que ayude a satisfacer la demanda informativa del individuo.

A su vez, la evaluación de consultas en un sistema de recuperación de información debe responder a los siguientes cuestionamientos: ¿cómo se puede representar un documento con las palabras clave seleccionadas y cómo se comparan los documentos y las representaciones de consultas para calcular una valoración del uso del recurso? Aunado a ello, la recuperación de información se ocupa de cuestiones como la incertidumbre y la vaguedad en los sistemas de información.

De esta manera, el principio de incertidumbre se refiere a la representación que no suele reflejar la verdadera semántica de recursos de información como imágenes, videos, grabaciones sonoras, etc. Por otra parte, el concepto de vaguedad manifiesta que la información que el usuario requiere carece de claridad, ya que solo se expresa vagamente a través de una consulta, retroalimentación o acción en un SRI.

El concepto de evaluación del sistema pone de manifiesto la importancia de determinar el impacto de la información proporcionada en los logros del usuario. Aquí vemos si la eficiencia del sistema en particular se relaciona con el tiempo y el esfuerzo que el usuario dedica para resolver su demanda informativa.

Por lo tanto, las operaciones esenciales del sistema de recuperación de información que deseamos considerar se limitan a aquellas

que se han cubierto en esta investigación y que se consideran parte del núcleo de todo el problema relacionado con la efectiva recuperación de los recursos. Al respecto existe una serie de acciones importantes relacionadas con la indexación de los recursos que resaltan las problemáticas de la recuperación de información que persisten actualmente en los SRI. Estas acciones son expuestas cuando un usuario determinado pretende satisfacer su demanda informativa a través de dichos sistemas:

a) indexación de conceptos de los recursos de información,
b) traducir el concepto de indexación al lenguaje descriptor,
c) ingresar información codificada en el núcleo del sistema,
d) análisis del concepto de la consulta,
e) traducir el análisis del concepto en el lenguaje de los descriptores,
f) extraer información codificada del núcleo del sistema.

Si bien la indexación de los recursos es un proceso intelectual que tiene el propósito de representar el contenido informativo de los recursos, dentro del propio SRI en ocasiones este proceso puede verse alterado por una serie de factores que van desde la inadecuada selección de los atributos descriptivos de los recursos, hasta olvidar la figura del usuario, que será quien utilizará y buscará dichos recursos en el sistema.

Según Romá Ferri (2015, 6), la necesidad de información del usuario se debe entender como la conducta activa para obtener información. Es la conducta que conlleva la transformación de la expresión, escueta y significativa en lenguaje natural, al lenguaje documental en el SRI. A esta expresión transformada se le conoce como "estrategia de búsqueda" y es con la cual se establece la directriz o lógica con la que se realiza la acción de formular una consulta en el sistema de recuperación de información. En dicha estrategia se emplean los términos o descriptores representativos de la necesidad de información; dicha expresión está condicionada por las funcionalidades del SRI.

De esta manera, la indexación es un método de adquisición de información (desarrollo de información) mediante el cual los recursos de información se recopilan y clasifican en función de palabras clave. Posteriormente se desarrolla un índice que es el puente entre la necesidad del usuario y el recurso en cuestión.

Así, los recursos indexados, en su mayoría contenido de texto, se preparan para la búsqueda de un recurso específico o palabra clave y se les proporcionan descriptores. Si se busca una palabra clave y los recursos relacionados, lo ideal es que se muestre el contenido más relevante en función del descriptor seleccionado.

Por ejemplo, en una biblioteca los descriptores pueden ser datos como los números de autor, título o ISBN. En principio, lo mismo sucede con una consulta en Internet. En otras palabras, el término indexación denota la formación de un índice en el que los recursos web se recogen y clasifican utilizando varios descriptores (tales como palabras clave) y se ponen a disposición para búsquedas posteriores, fomentando con ello la recuperación de información.

De acuerdo con Kagolovsky y Moehr (2003, 401), una de las razones de esta dificultad está relacionada con el problema básico de la semántica, pues la relación entre una entidad del mundo real, las construcciones mentales (conceptos) sobre esta entidad y la terminología asociada generalmente se representa utilizando el "triángulo de significado". Esta construcción teórica demuestra que la única conexión entre las entidades del mundo real y la terminología relacionada para describir los recursos de información se produce a través de construcciones mentales. Esta conexión explica la ambigüedad con la que los seres humanos utilizan la terminología para desarrollar búsquedas en un determinado sistema.

Uno de los ejemplos de esta ambigüedad es que las personas a menudo usan los mismos términos cuando se refieren a diferentes conceptos y entidades. A esto se le llama "polisemia". Otra variante, llamada "sinonimia", es el uso de diferentes términos para identificar entidades similares y construcciones cognitivas.

De esta manera, la polisemia y la sinonimia son problemas terminológicos de la recuperación de información que impactan en

un sistema particular y en la actualidad siguen siendo un reto para obtener una recuperación más eficiente dentro de cualquier sistema de información, pues la elección de un lenguaje documental para representar el contenido intelectual y la información de los recursos no asegura con totalidad su recuperación efectiva dentro de un SRI. "El lenguaje documental normaliza y controla, a través de los conceptos, toda aquella terminología que hay en el lenguaje natural como son sinónimos, formas variantes de escritura, términos redundantes o en desuso" (Sánchez 2012, 76).

Se estima que el uso de estos lenguajes documentales ayude sobre todo a recuperar información de índole especializada, pues no toda la información contenida en los recursos puede ser susceptible de describirse y emparejarse mediante el uso de un determinado lenguaje documental.

Bajo esta premisa, conviene reconocer los límites e inconvenientes del lenguaje documental frente al lenguaje libre, pues los lenguajes documentales permiten una indización menos específica y exhaustiva que el lenguaje libre; su actualización es difícil (necesitan un estudio y análisis previo a la incorporación de nuevos términos); encarecen los procesos de análisis documental, y abren la posibilidad a errores puesto que la elección de puntos de acceso está en manos de profesionales y no de los autores.

Mucho se ha discutido acerca de la efectividad de estos lenguajes para recuperar información, pues se afirma que son la alternativa idónea para representar fielmente a los temas que forman parte del contenido de los recursos. De hecho, un SRI recibe una fuerte influencia de los temas que son asignados para representar los recursos de información; debido a ello han surgido los problemas que hemos mencionado con anterioridad. Por lo tanto, la efectiva recuperación de información mediante el análisis temático de los recursos es un reto persistente en los SRI actuales.

El tratamiento temático de la información, que tradicionalmente ha sido controlado por bibliotecarios y profesionales de la información con la ayuda de herramientas de clasificación, catalogación y control de vocabulario, es un área importante de investigación en el entorno de los SRI.

Mientras tanto, frente a enormes cantidades de información, los usuarios no necesitan que los bibliotecarios solo brinden un servicio simple como la búsqueda de literatura. Las necesidades de los usuarios de bibliotecas se están diversificando y, a medida que los usuarios están mejor educados, su demanda también es mayor. Es por ello que en la actualidad es necesario contar con SRI adaptables y compatibles con dichas necesidades.

Los llamados sistemas para la recuperación de información de nueva generación contemplan otorgar acceso a distintos recursos como artículos, pre-prints, datasets, imágenes, software, etc. Estarán centrados en los objetos o fuentes identificados inequívocamente por su URI. Sobre estos objetos se construirá una capa enriquecida de servicios añadidos.

Las nuevas generaciones de SRI conformarán una red con conexiones y enlaces entre sus diferentes recursos, y permitirán el desarrollo de servicios compartidos. Sus contenidos no serán estáticos; permitirán comentarios, diferentes versiones, enlaces entre diferentes contenidos, y no esperarán a ser recolectados, sino que serán ellos los que se comuniquen con otros sistemas para ofrecer sus novedades. De esta manera, los componentes de un sistema de recuperación de información son:

1. La base de datos documental. Es el espacio en donde se almacenan los recursos de información y sus respectivos registros. Los datos que conforman los recursos de información son estructurados bajo un formato específico dentro de esta base de datos, con lo cual se permite la representación del recurso en el SRI.
2. El subsistema de consulta. Es el módulo que permite al usuario interactuar con los registros de los recursos almacenados dentro del SRI. A través de este módulo, el usuario puede realizar las operaciones o consultas que le permitirán recuperar un recurso de información en específico, o bien un listado de resultados que contenga los términos de búsqueda aplicados en una determinada consulta. Este subsistema está conformado por una interfaz que permite

al usuario formular dichas consultas y por un analizador sintáctico que toma la consulta escrita por el usuario y la desglosa en sus partes integrantes. En la mayoría de los casos, los usuarios de los SRI realizan sus consultas basándose en la estructura de consultas booleanas. Cada uno de los elementos básicos de consulta puede ser un término (descriptor o concepto).

3. El subsistema de evaluación. Este módulo es el encargado de calcular el grado en el que las representaciones de los recursos de información satisfacen los requisitos expresados en la consulta y recupera aquellos recursos que son relevantes a la misma. Destacan dos modalidades de evaluación: una fórmula que empareja los recursos individualmente con la consulta, uno por uno, y otra que los empareja en su conjunto. La evaluación del SRI es un aspecto de suma importancia, sobre todo si se considera que es un método por el cual es posible conocer el grado de satisfacción informativa del usuario al momento de interactuar y recuperar recursos del SRI.

Asimismo, los SRI poseen componentes esenciales que les permiten dotarlos de una estructura que integra sus alcances y limitaciones dentro de un contexto informativo en particular. Como puede apreciarse, estos sistemas guardan una relación cercana con los usuarios que interactuarán con ellos, pues todo sistema de información debe contener un nivel de consideración hacia la demanda informativa del usuario.

Como hemos abordado en este capítulo, el objetivo de un sistema de recuperación de información es obtener de forma eficaz recursos de información relevantes para la consulta de un usuario. A través del tiempo, se han desarrollado muchos modelos de recuperación de información; por ejemplo, el modelo booleano, de espacio vectorial, probabilístico, modelos booleanos difusos y extendidos. Cada uno tiene un conjunto único de ventajas y desventajas, pues la información tiene características muy diversas que sería imposible adaptar un solo modelo y sistema para propiciar su búsqueda y recuperación de una manera uniforme.

Bajo esta premisa, la investigación en sistemas de recuperación de información es un campo que ofrece las posibilidades de plantear y debatir una multitud de cuestiones relevantes y metodológicas sobre gran parte de los fundamentos y fenómenos que conciernen a la evolución de la información y su adaptación en nuevos sistemas para recuperar la información.

En este sentido, la investigación sobre los SRI permitirá abordar los principios teóricos y metodológicos en las pruebas de concepto formuladas para implementar SRI con nuevas características y enfoques. Aunado a esto, los sistemas de recuperación de información basados en datos abiertos enlazados ponen de manifiesto la incursión en nuevas metodologías y principios para organizar y representar la información que se encuentra disponible en el ambiente digital.

Datos abiertos enlazados y recuperación de información

La aplicación de los datos abiertos enlazados en los sistemas para la recuperación de información se encuentra en constante investigación y desarrollo. Son amplios los estudios de caso que exponen diferentes metodologías, técnicas y principios para llevar a cabo la implementación de ambos elementos en un entorno interoperable y susceptible de sistematizarse.

Los datos abiertos enlazados son el componente principal de la web semántica. Estos datos representan una metodología para publicar de manera abierta e interoperable los datos que forman parte de diferentes fuentes disponibles en la web; por lo tanto, un sistema basado en estos datos deberá permitir la búsqueda, recuperación y visualización de los datos de diferentes fuentes que se encuentren interconectadas con atributos comunes dentro de sus estructuras. El elemento principal para el desarrollo de estos sistemas consiste en la apertura total de los datos que se interconectarán. El postulado anterior nos lleva a contemplar que los datos plenamente abiertos deben estar libres de barreras económicas, legales y técnicas.

"Los datos abiertos [enlazados] son datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen" (Open Knowledge Foundation 2021, párr. 3). En este sentido, el concepto de apertura pone de manifiesto los siguientes aspectos:

- Disponibilidad y acceso: la información debe estar disponible como un todo y a un costo razonable de reproducción, preferiblemente descargándola de Internet. Además, la información debe estar disponible en una forma conveniente y modificable.
- Reutilización y redistribución: los datos deben ser provistos bajo términos que permitan reutilizarlos y redistribuirlos, e incluso integrarlos con otros conjuntos de datos.
- Participación universal: todos deben poder utilizar, reutilizar y redistribuir la información. No debe haber discriminación alguna en términos de esfuerzo, personas o grupos. Restricciones "no comerciales" que prevendrían el uso comercial de los datos o restricciones de uso para ciertos propósitos (por ejemplo, sólo para educación) no son permitidas.

Por lo tanto, la recuperación de la información desde la perspectiva de los datos abiertos enlazados no puede concebirse sin la plena apertura de los datos que serán susceptibles de implementarse en el contexto de los SRI. En la actualidad es sabido que los fines comerciales de la información tienen un fuerte impacto en el desarrollo de sistemas de información, pues estos fines son tomados en cuenta para generar ingresos y valor comercial como parte de la comercialización de la información.

Bajo esta premisa, la industria de la información se define, a nivel internacional, como la industria que engloba las empresas que manejan contenidos y su administración, así como aquellas que la transportan, distribuyen y le dan valor agregado (Turnbull 2005, 1).

Además de eso, la industria de la información agrupa una serie de actividades relacionadas con el propio acceso a la información.

En la actualidad, el acceso a la información es un proceso costoso para los actores que forman parte de los contextos académicos y de investigación, pues esta información es distribuida y publicada por grandes cadenas comerciales que buscan un beneficio económico mediante el procesamiento y la validación de información que puede utilizarse para la generación de nuevos conocimientos.

Bajo esta premisa, la industria de la información ha trascendido al punto de considerar los datos como una fuente de gran valor, sobre todo cuando su reutilización fomenta la generación de nuevos métodos de revisión y validación de investigación. De acuerdo con Harari (2019, 419), esta realidad responde a una cuestión mucho más compleja, pues a menudo el sistema global de procesamiento de datos se vuelve omnisciente y omnipotente; conectarse con el sistema se convierte en el origen de todo sentido.

Los humanos quieren fusionarse con el flujo de datos porque cuando formas parte del flujo de datos, formas parte de algo mucho mayor que tú. Es decir, los datos hoy en día forman parte de todos y cada uno de los procesos que se desarrollan en la actividad humana, consciente o inconscientemente se generan datos que están en constante fluctuación en nuestro entorno.

Los datos abiertos enlazados fueron conceptualizados con el afán de propiciar el desarrollo de un entorno digital interoperable, en donde la recuperación de información permita descubrir nuevos hallazgos fundamentados en el uso y la reutilización de datos disponibles en diferentes fuentes y plataformas del entorno digital. En este sentido, se considera oportuno mencionar las tres "leyes" sugeridas por Eaves (SPUK 2015) respecto a la apertura y el uso de los datos en el ambiente digital:

1. Si el dato no puede ser encontrado o indexado en la web, no existe.
2. Si el dato no está disponible en un formato abierto y legible por computador, no puede ser reutilizado.

3. Si el marco legal no permite que el dato sea compartido, no es útil.

En esencia, si los datos abiertos enlazados no cumplen con las características anteriores, muy difícilmente su aplicación y funcionalidad podrán verse reflejados en el ambiente digital. Además, la interoperabilidad global de los datos es otro principio que se ve fuertemente alterado por la presencia de estas caracterizaciones.

La interoperabilidad es la posibilidad de que distintos tipos de computadoras, redes, sistemas operativos y aplicaciones trabajen juntos de forma eficaz, sin comunicación previa, de tal forma que puedan intercambiar información de manera útil y con sentido.

Hay tres aspectos que se deben tener en cuenta en la interoperabilidad: semántica, estructura y sintáctica. De acuerdo con Gómez (2007), la interoperabilidad es la capacidad de sistemas múltiples con diversas plataformas del *hardware* y del *software*, estructuras de datos e interfaces para intercambiar datos con la pérdida mínima de contenido y funcionalidad. En el contexto de los datos abiertos enlazados, la interoperabilidad es trascendental al momento de adaptar e implementar protocolos y estándares para la estructuración de los datos y su aprovechamiento en el ambiente digital. Por lo tanto, el proceso de RI mediante estos datos también deberá ser de carácter interoperable y contextual.

La interoperabilidad en un SRI puede ser alcanzada mediante la adopción de estándares abiertos y el desarrollo de políticas. Un estándar de estas características es una especificación técnica que normalmente proviene de organizaciones encargadas de generar normas internacionales o bien, de organismos regionales y locales. Los estándares poseen la característica principal de ser desarrollados, mantenidos, aprobados y ratificados mediante consenso en un organismo especializado que cuenta con integrantes cualificados para validar la aplicación de dicho estándar.

Este tipo de estándares están disponibles públicamente sin ningún costo o mediante precios razonables que faciliten su adopción y puesta en práctica por cualquier comunidad interesada. La adopción de recomendaciones es otra manera de fomentar la

interoperabilidad entre sistemas. Este tipo de estrategia incluye especificaciones técnicas de organismos, por ejemplo, el W3C. Si bien las recomendaciones no son estándares, sí incluyen especificaciones para su adopción en el ámbito de los sistemas.

Gómez (2007) manifiesta que aunque los investigadores han estado luchando para lograr la interoperabilidad por más de 20 años, éste ha sido un problema crítico y lo será para el futuro próximo porque el número de los sistemas informáticos, depósitos de la información, usos y los usuarios se están multiplicando en una forma exponencial. Lo anterior obliga a que en el análisis y diseño de los sistemas de información del futuro esté siempre presente el componente de interoperabilidad asociada.

De esta manera, la interoperabilidad entre sistemas puede mejorar la comunicación, eficiencia y precisión de la transmisión de información, lo que eventualmente conduce a mejoras en los resultados de recuperación de información y en la respuesta a demandas informativas de la comunidad usuaria.

A su vez, las tecnologías semánticas permiten establecer procesos de interoperabilidad de datos abiertos enlazados más allá del punto de alineación pura de formato y estructura de dichos datos. Durante los últimos años, la actividad de la web semántica ha cobrado impulso con la publicación generalizada de datos estructurados como RDF.

En este sentido, el paradigma de datos enlazados ha evolucionado de una idea de investigación práctica a un candidato muy prometedor para abordar uno de los mayores desafíos en el área de la visión de la web semántica: la explotación de la web como plataforma para la integración de datos y un medio idóneo para recuperar información de variada naturaleza en diversas fuentes del entorno digital.

Se puede apreciar cómo aunque todavía no se han alcanzado algunos de los objetivos de la web semántica, varias aplicaciones conocidas y exitosas ya están utilizando tecnologías semánticas, como Knowledge Graph de Google, Satori de Microsoft o Graph Search de Facebook.

Un grafo de conocimiento (*knowledge graph*) representa una colección de descripciones interconectadas de entidades: objetos,

eventos o conceptos. Los grafos de conocimiento ponen los datos en contexto a través de enlaces y metadatos semánticos y de esta manera proporcionan un marco para la integración, unificación, análisis e intercambio de datos.

Los grafos de conocimiento son un componente evolutivo de la aplicación de los datos abiertos enlazados en el contexto de la recuperación de información. Forman parte de las tecnologías de la web semántica que han sido adaptadas por diferentes aplicaciones y plataformas disponibles en la web.

Debido al crecimiento explosivo de la información, los servicios de información en línea se han vuelto cada vez más importantes para que las personas obtengan información y comprendan de una mejor manera los fenómenos que acontecen día a día en el mundo, pues el tsunami de datos que se presenta en el entorno digital actual ha motivado la necesidad de contar con nuevas estrategias para obtener información confiable que pueda ser utilizada para validar hechos y acontecimientos de la vida diaria.

Por ejemplo, algunos portales de noticias también han comenzado a utilizar las tecnologías semánticas como los grafos de conocimiento para asegurar una mejor asimilación de los contenidos por parte de su comunidad usuaria. De acuerdo con Liu y colegas (2019, 1), un grafo de estas características incluye las relaciones de colaboración de entidades codificadas en artículos de noticias y los comportamientos de navegación de los usuarios.

Estas relaciones de colaboración revelan la similitud del contexto de las entidades en las noticias. Por ejemplo, las entidades que aparecen con frecuencia en artículos o en las que los mismos usuarios hacen clic suelen estar muy relacionadas con el dominio donde aparecen dichas noticias. Por lo tanto, un grafo de conocimiento también considera el entorno en donde los datos abiertos enlazados se ubican, pues se trata de establecer un mecanismo capaz de ofrecer relaciones semánticas que expliquen la unión entre datos que forman parte de un mismo contexto y que puedan tener una conexión con entornos de atributos similares.

En este sentido, un grafo de conocimiento permite representar una colección de descripciones interconectadas de entidades —objetos y eventos del mundo real, o conceptos abstractos (por

ejemplo, documentos)— donde las descripciones tienen una semántica formal que permite que tanto las personas como las computadoras las procesen de manera eficiente y sin ambigüedades.

Las descripciones de entidades se contribuyen unas a otras y forman una red donde cada entidad representa parte de la descripción de las entidades relacionadas con ella y proporciona un contexto para su latente interpretación. En este sentido, los grafos de conocimiento combinan características de varios paradigmas de gestión de datos:

- Utilizan bases de datos, porque los datos se pueden explorar mediante consultas estructuradas.
- Hacen uso de grafos, porque se pueden analizar como cualquier otra estructura de datos en red.
- Emplean una base de conocimiento, porque contienen semántica formal, que se puede utilizar para interpretar los datos e inferir nuevos hechos (Ontotext 2021).

Además, los grafos de conocimiento, representados en RDF, proporcionan un marco común para la integración, unificación, vinculación y reutilización de los datos porque combinan:

- Expresividad: los estándares pilares de la web semántica (RDF(S) Y OWL) permiten una representación fluida de varios tipos de datos y contenido, por ejemplo: esquema de datos, taxonomías y vocabularios.
- Rendimiento: Todas las especificaciones han sido pensadas y probadas en la práctica para permitir una gestión eficiente de gráficos de miles de millones de hechos y propiedades.
- Interoperabilidad: Existe una gama de especificaciones para seriar datos, acceso (protocolo SPARQL para puntos finales), administración (SPARQL Graph Store) y federación. El uso de identificadores únicos a nivel mundial facilita la integración y publicación de datos, lo que fomenta un ecosistema interoperable de datos.
- Estandarización: todo lo anterior está estandarizado a través de la comunidad W3C para garantizar que se satisfagan

los requisitos de los diferentes actores, desde los lógicos hasta los profesionales de la gestión de datos y los equipos de operaciones del sistema.

Actualmente se experimenta un cambio de paradigma significativo en el acceso e intercambio de información en la web. Ésta no es la primera vez que la web ha cambiado drásticamente la forma en que cooperamos y nos comunicamos. Con el correo electrónico surgió la comunicación en línea (casi) instantánea y con la web una plataforma mundial para el intercambio de información. Además del proceso para recuperar información, en la actualidad los dispositivos inteligentes y computacionales ponen de manifiesto nuevos comportamientos para acceder a la información.

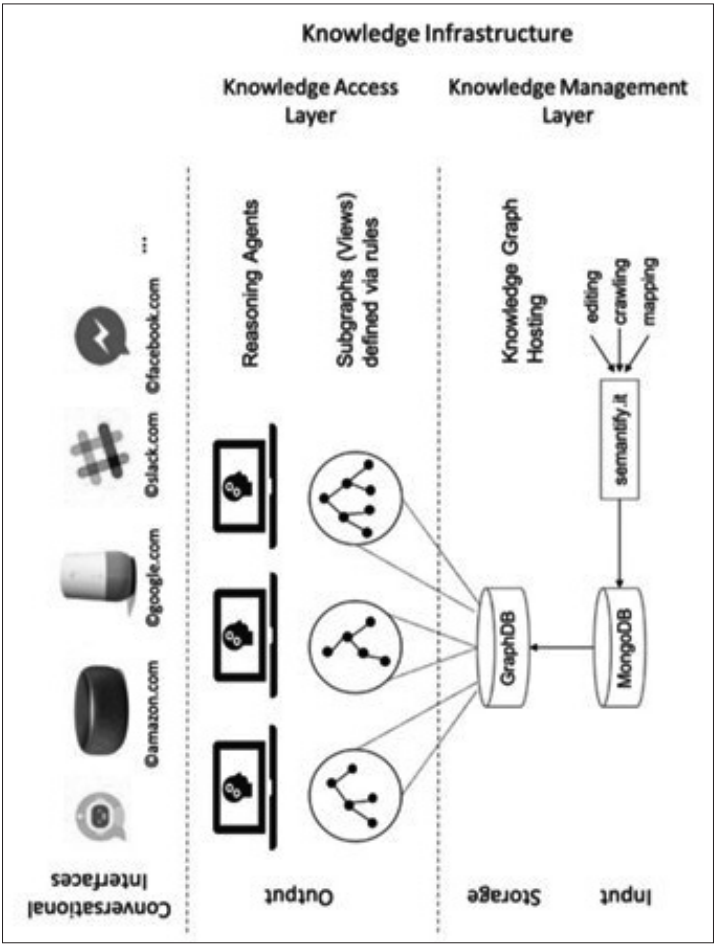
En la figura 21 pueden apreciarse los componentes de la infraestructura de conocimiento que permiten el procesamiento y razonamiento de datos mediante el uso de aplicaciones y agentes computacionales. En este escenario, la recuperación de información se desarrolla a través de diferentes dispositivos, ejemplificando con ello la interoperabilidad de los datos que prevalece en el contexto digital actual.

Esto nos lleva a considerar que la recuperación de información mediante el uso de datos abiertos enlazados es un proceso que también está relacionado con el tipo de dispositivo en donde el usuario desea acceder a la información; por esta razón es muy importante contemplar la interoperabilidad que debe estar presente dentro de un contexto de datos abiertos enlazados mediante el uso de grafos de conocimiento que permitan recuperar de una manera semántica los datos que están ubicados en diversas fuentes.

Con el creciente volumen de recursos de información en la web, la necesidad de un mecanismo automático para extraer conocimiento se ha vuelto más crítico y necesario. Se estima que los datos abietos enlazados aplicados a la RI puedan ayudar con el cumplimiento de este propósito.

La construccción automática de relaciones semánticas dirigidas entre conceptos para ayudar a los usuarios de la información a recuperar de manera eficiente el conocimiento apropiado de la web

Figura 21. Agentes de razonamiento que acceden a un grafo de conocimiento



Fuente: D. Fensel *et al.*, Knowledge Graphs 2020. Disponible en https://doi.org/10.1007/978-3-030-37439-6_3. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthKAWOW_AqimUzVhNyA?e=H2Zo1d.

es uno de los propósitos de la aplicación de los grafos de conocimiento que hemos abordado con anterioridad. No obstante, el rápido desarrollo de la tecnología de la información y la informática ha contribuido al aumento de la recopilación y el almacenamiento de datos digitales. Sin embargo, el rápido aumento asociado en los volúmenes de datos digitales no se correlaciona automáticamente con nuevos conocimientos y avances en nuestra comprensión de esos datos.

Por lo tanto, se requiere de metodologías que permitan a los datos disponibles en plataformas transformarlos en conocimiento asimilable e interpretable por agentes computarizados capaces de ofrecer a los usuarios una nueva vía para satisfacer sus necesidades informativas mediante procesos de recuperación integrales y empáticos con las características del usuario final. Pues a menudo los sistemas para recuperar información no contemplan el factor humano que se adhiere al perfil del usuario de la información.

Las características de la recuperación de información mediante el uso de datos abiertos enlazados ponen de manifiesto las siguientes consideraciones:

- Los sistemas de recuperacion de información diseñados bajo el uso de estos dos componentes deben fomentar el uso interoperable de los datos para recuperar piezas de información disponibles en diferentes fuentes de la web.
- Además de recuperar información, deben motivar la consulta a las vinculaciones semánticas que explican la relación manifestada entre los datos.
- Propiciar la visualización de grafos de conocimiento que permitan al usuario navegar, interactuar y descubrir nuevos hallazgos basados en datos e información.
- Ejercer la búsqueda textual de los datos, pero motivar la interacción intuitiva entre nodos y aristas, mediante los grafos generados a partir del procesamiento de los datos.
- Permitir al usuario acceder a los datos y la información de una manera abierta; es decir, libre de restricciones económicas, técnicas y legales para fomentar el principio de apertura a los datos, la información y el conocimiento.

- Traducir la necesidad informativa del usuario mediante estrategias que fomenten el uso y la conexión interoperable entre los datos que forman parte de diferentes fuentes en el entorno de la web.
- Propiciar la recuperación de manifestaciones, expresiones y obras que relacionan con un determinado dominio informativo mediante el análisis de las entidades que forman parte de un contexto en particular.

Aunado a ello, Dimou y colegas (2014, 247) manifiestan que las visualizaciones ofrecen una interfaz interactiva sobre los datos abiertos enlazados publicados que permite a los usuarios que no estén familiarizados con las tecnologías de la web semántica explorar los datos y sus enlaces.

De esta manera, la recuperación de información mediante el uso de datos abiertos enlazados propiciará la aparición de una nueva generación de SRI, los cuales deberán ser flexibles e interoperables con las características actuales de la web. Esto podría generar un cambio de paradigma en la manera de concebir la información y los métodos para acceder a ella, tomando en cuenta las demandas informativas que los usuarios puedan manifestar en el presente, donde la información y los datos son cada vez más abundantes y de características complejas.

Cientos de millones de personas se involucran en la recuperación de información todos los días cuando utilizan un motor de búsqueda en la web o buscan en su propio correo electrónico. La recuperación de información se está convirtiendo rápidamente en la forma dominante de acceso a la información, superando la búsqueda tradicional de base de datos, al grado de utilizar sofisticados sistemas computacionales para obtener la información que ayude a resolver una demanda informativa.

Bajo esta premisa, los sistemas de recuperación de información también se pueden distinguir por la escala a la que operan. En la búsqueda web, el sistema tiene que proporcionar búsquedas en millones de documentos almacenados en enormes cantidades de computadoras. Los problemas distintivos en este contexto radican en la necesidad de recopilar documentos para indexarlos, y de

esta manera poder construir sistemas que funcionen de manera eficiente a esta enorme escala y manejar aspectos particulares de la web. Por ejemplo, la explotación del hipertexto y no ser engañados por los proveedores del sitio que manipulan el contenido de la página en un intento por mejorar su clasificación en los motores de búsqueda. Por este motivo es necesario implementar las tecnologías de la web semántica para obtener resultados significativos en cuanto al contenido de la información y su relevancia para resolver la necesidad informativa del individuo.

Entonces, será relevante que el usuario defina con claridad su necesidad informativa al momento de interactuar con el sistema. Se han realizado numerosos estudios de usuarios involucrados en el proceso de búsqueda y los resultados pueden ayudar a guiar el diseño de interfaces de búsqueda de carácter intuitivo e interactivo, esto acorde a los requerimientos de la recuperación de información mediante el uso de datos abiertos enlazados.

Una observación común es que los usuarios a menudo reformulan sus consultas con ligeras modificaciones, ya que esto puede ser más fácil que intentar especificar la consulta precisamente en el primer intento. Cuando se muestran los resultados de la búsqueda, los recursos de información deben mostrarse completos o se debe presentar al buscador algún tipo de representación del contenido de esos recursos.

El sustituto del recurso se refiere a la información que resume el propio contenido del recurso y es una parte clave del éxito de la interfaz de búsqueda. El diseño de recursos sustitutos y la visualización de resultados de recuperación es un área activa de investigación y experimentación.

Aunado a ello, las interfaces de búsqueda deben permitir la visualización de los datos abiertos enlazados, pues se trata de un elemento crucial en la conformación y estructura de sistemas de recuperación capaces de retroalimentar la búsqueda del usuario, en concordancia con el dominio temático de la información que desea recuperar.

Así pues, una interfaz de usuario se diseña a través de un proceso interactivo en el que los objetivos y las tareas se aclaran a través de la investigación del usuario, y luego se crean los diseños

iniciales, a menudo basados en diseños existentes, pero que potencialmente incluyen nuevas ideas. Estos diseños iniciales se prueban con los posibles usuarios, y luego son evaluados y rediseñados, y evaluados nuevamente, en un ciclo que puede repetirse numerosas veces.

En el diseño de sistemas para recuperar información mediante los principios de los datos abiertos enlazados, será necesario sujetarse a un modelo que fundamente las etapas y los procesos a seguir mediante la experimentación y prueba conceptual. Estos elementos serán abordados en el siguiente capítulo de esta obra.

Modelo para la recuperación de información con datos abiertos enlazados

Fundamentación

El modelo de datos abiertos enlazados para recuperar información pone de manifiesto la implementación de los principios universales de la web semántica, aquellos que se refieren al uso de datos para crear un entorno interoperable de información con significado para el usuario final. La web semántica es una web de datos, fechas, títulos, números, propiedades químicas y cualquier otro dato que se pueda imaginar. El conjunto de tecnologías de la web semántica (RDF, OWL, SKOS, SPARQL, etc.) proporciona un entorno donde su aplicación permite consultar esos datos, hacer inferencias utilizando vocabularios, esquemas de metadatos y ontologías.

Sin embargo, para hacer de la web de datos una realidad es importante tener la enorme cantidad de datos en la web disponible en un formato estándar, accesible y manejable por las herramientas de la web semántica. Además, la web semántica no solo necesita acceso a los datos, sino que las relaciones entre los datos también deben estar disponibles para crear una web de datos (en contraposición a una mera colección de conjuntos de datos). Esta colección de conjuntos de datos interrelacionados en la web de una manera abierta es el reflejo de los datos abiertos enlazados.

Para lograr consolidar y crear datos abiertos enlazados, las tecnologías deben estar disponibles en un formato común (RDF). Esto les permite realizar conversiones o acceso sobre la marcha a las bases de datos existentes (relacionales, XML, HTML, etc.). También es importante configurar los puntos de consulta para acceder a esos datos de manera más conveniente; por ejemplo, mediante la creación de motores de búsqueda semánticos. En este sentido, el W3C proporciona una gama de tecnologías (por ejemplo, RDFa y RIF) para obtener acceso a los datos que han sido publicados.

De esta manera, los datos abiertos enlazados son el corazón de la web semántica, su integración a gran escala y el razonamiento sobre los datos dependerá de su sistematización en el ambiente web. Casi todas las aplicaciones enumeradas en la mayoría de los estudios de caso de la web semántica se basan esencialmente en la accesibilidad e integración de los datos abiertos enlazados en varios niveles de complejidad y especificidad.

Un ejemplo típico de esta situación es el gran conjunto de datos abiertos enlazados que representa DBPedia. Esta plataforma esencialmente hace que el contenido de Wikipedia esté disponible en RDF. La importancia de DBPedia no solo radica en la disponibilidad de los datos de Wikipedia, sino también en que incorpora enlaces a otros conjuntos de datos en la web, por ejemplo, a Geonames (nombres geográficos).

Al proporcionar estos enlaces adicionales (en términos de triples RDF), las aplicaciones pueden aprovechar el conocimiento adicional (y posiblemente más preciso) de otros conjuntos de datos al desarrollar una aplicación con capacidad de buscar y recuperar datos desde un punto de vista semántico; en virtud de la integración y el acceso a las vinculaciones entre varios conjuntos de datos, la aplicación puede proporcionar una mejor experiencia de recuperación al usuario.

Los datos abiertos enlazados son parte de una visión integradora y universal para el progreso actual de la web. Este entorno digital despegó como una red de documentos con hipervínculos que fueron una innovación con antecedentes nunca antes vistos, pero dichos documentos no se pueden utilizar de forma eficaz y flexible como los datos.

De hecho, gran parte de la web se basa en datos, y por largos periodos de tiempo los datos se han ocultado en archivos dentro de servidores, plataformas y concentradores. Si bien los documentos hablan de personas y cosas, se necesitan datos con un significado establecido para poder identificar información de una manera más precisa y relacionar datos con diversas manifestaciones, expresiones y entidades disponibles en el ambiente digital.

A medida que la web ha evolucionado, también se ha proliferado de datos, los cuales pueden ser de diversa naturaleza y tipología. Los estándares de los datos abiertos enlazados han permitido publicar datos de una manera que pueden ser leídos por personas y procesados por computadoras para que los flujos de datos previamente ocultos puedan ser más evidentes y accesibles para los usuarios.

Quizás los datos abiertos enlazados pueden no ser tan interesantes de leer como una web de hipertexto, pero son más interesantes en términos de hacer que todo funcione de manera más eficaz, desde los negocios hasta el trabajo científico de investigación. Las computadoras pueden leer, procesar y combinar datos enlazados de forma mucho más eficaz y propiciar la generación de descubrimientos intensivos basados en datos.

En este sentido, los datos abiertos enlazados han alcanzado la mayoría de edad en los últimos años. En la actualidad, hemos visto a Google anunciar su grafo de conocimiento y adoptar el formato de seriación JSON-LD para Gmail y producir un gran conjunto de términos para uso general en schema.org.

Por otra parte, la empresa IBM ha anunciado desde el 2014 que la base de datos DB2 se convertirá en un servidor de datos enlazados, y Facebook ha expuesto datos vinculados a través de su API Graph. Otras grandes empresas y organizaciones gubernamentales han seguido esta línea, por ejemplo: Data.gov del gobierno de los Estados Unidos (https://www.data.gov/), data.gov.uk del Reino Unido (https://data.gov.uk/) y el proyecto Land (https://landportal.org/developers/what-is-linked-open-data).

De esta manera, los componentes básicos de los datos abiertos enlazados no son particularmente nuevos. La propuesta original

para la World Wide Web ya incluía una primera noción para construir hipervínculos con semántica. Estos hipervínculos tenían que formar parte de un sistema complejo capaz de relacionar los datos con obras, manifestaciones y expresiones disponibles en la web, lo que daba lugar a sistemas de recuperación de información de nueva generación capaces de combinar las metodologías clásicas de la RI con las estrategias de visualización que permiten consultar grandes cantidades de datos en un solo escenario integrador.

Por lo tanto, en este modelo se plantean los componentes que pueden implementarse en un sistema de recuperación de información que combine las metodologías clásicas y las tendencias actuales para consultar y acceder a los datos abiertos enlazados que han sido desarrollados en un dominio específico.

En este apartado se explicará cómo los datos abiertos enlazados pueden implementarse a través de un modelo metodológico que permite abordar las etapas del procesamiento de los datos y su puesta a disposición mediante una interfaz que permite interactuar con los datos mediante un proceso de recuperación de información.

Bajo esta premisa, el entorno de la información ha cambiado vertiginosamente. Es necesario establecer metodologías para comprender los comportamientos de los nuevos recursos de información emergentes. Al respecto, los conjuntos de datos están siendo utilizados como una nueva fuente capaz de acelerar el descubrimiento de nuevos conocimientos; claro está, con la ayuda de sistemas sofisticados que ayuden a comprender de mejor manera las necesidades del usuario final.

En este modelo, se destaca la función del usuario, pues es un modelo desarrollado por y para atender las demandas informativas de los individuos. De hecho, los datos abiertos enlazados han surgido como una necesidad de resolver demandas complejas relacionadas con el uso y la recuperación de información que forman parte de diversos dominios informativos.

Por lo tanto, se trata de un modelo holístico que contempla la totalidad de la sistematización de los datos abiertos enlazados, partiendo del análisis de sus partes y contemplando la función y

presencia del factor humano que hace posible la interacción del usuario con el sistema en cuestión.

En suma, este modelo representa una estructura que ayudará a comprender la naturaleza de los datos abiertos enlazados y su aplicación en el contexto de la recuperación de información. El modelo describe los componentes y la metodología que se contemplan al momento de concebir un entorno sistematizado de datos abiertos capaces de relacionarse semánticamente.

Además, siempre que se pretende visualizar el comportamiento de los datos, se toman en cuenta sus valores y se convierten de forma sistemática y lógica en elementos visuales que permitirán el desarrollo de un grafo final. En este sentido, la visualización de datos tiene que transmitir con precisión el detalle de los datos. No debe engañar ni distorsionar los atributos que representan. Esto es, si un número es dos veces más grande que otro, pero en la visualización parecen ser aproximadamente iguales, entonces la visualización es incorrecta.

Al mismo tiempo, una visualización de datos debe ser estéticamente agradable. Las buenas presentaciones visuales tienden a realzar el mensaje de la visualización. Si una figura contiene colores discordantes, elementos visuales desequilibrados u otras características que distraen, al usuario le resultará más difícil inspeccionar los datos e interpretarlos correctamente.

Por lo tanto, este modelo contemplará de manera enfática la visualización de los datos que conlleva implementar un método para recuperar información desde la perspectiva de los datos abiertos enlazados. A su vez, los principales atributos de la visualización de datos abiertos enlazados contemplan los siguientes aspectos:

- El usuario podrá navegar dentro de un conjunto específico y a través de múltiples conjuntos de datos enlazados para respaldar el descubrimiento de nuevos hallazgos.
- El usuario podrá interactuar con el sistema de datos abiertos enlazados mediante el análisis de la estructura y alineación de los datos, realizar consultas de información y plantear consultas de nivel básico, intermedio y avanzado.

- Mediante el modelo, se plantea la posibilidad de enriquecimiento del contenido a través de la anotación de datos y la identificación o derivación de relaciones semánticas entre los datos.
- Además, presentar e intercambiar datos y resultados derivados de su análisis a diferentes tipos de usuarios.

En el siguiente apartado se muestra la estructura del modelo de datos abiertos enlazados planteado en el desarrollo de esta investigación; cabe señalar que esta estructura ha sido construida a través del análisis de propuestas, estudios de caso y proyectos relacionados con la adopción de datos abiertos enlazados en diversos contextos informativos.

Estructura

El modelo de datos abiertos enlazados para la recuperación de información se encuentra estructurado en seis etapas principales. Las etapas enunciadas en este modelo tienen la particularidad de fundamentarse en los principios y estándares de los datos abiertos enlazados. El modelo tiene una fuerte influencia de las herramientas computacionales que pueden utilizarse para llevar a cabo las tareas de construcción de un entorno interoperable de datos abiertos enlazados, susceptible de aplicar metodologías para recuperar información mediante la interacción del usuario con un sistema determinado.

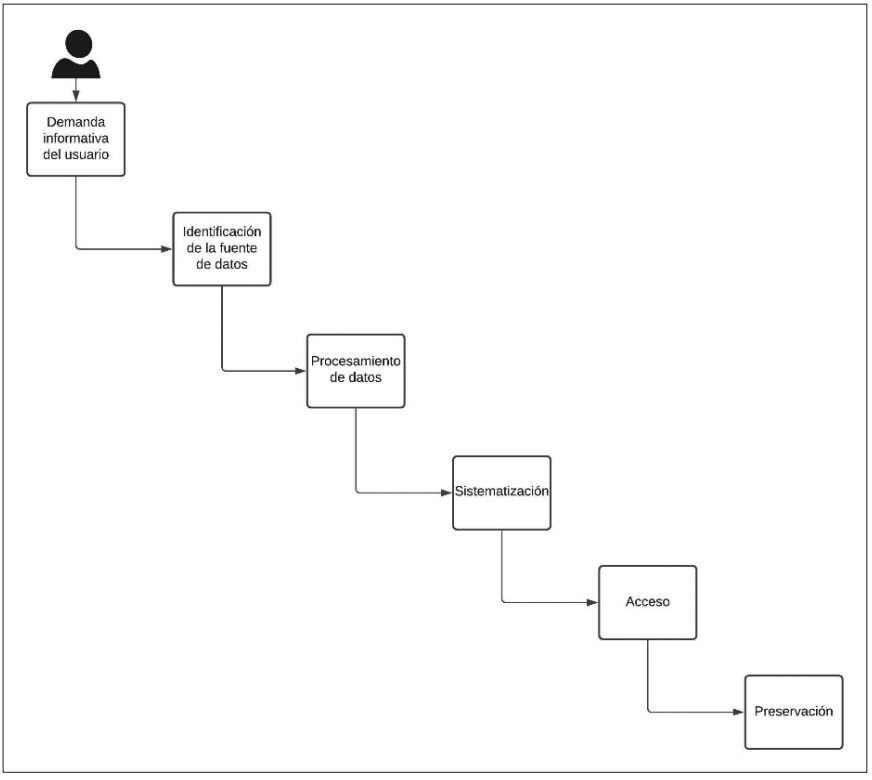
Un sistema de información basado en el uso de datos abiertos enlazados deberá propiciar la búsqueda y recuperación de diferentes recursos de información tomando en cuenta su correcta apertura, y permitiendo al usuario acceder al recurso sin tener que enfrentar restricciones legales, técnicas y económicas de acceso.

De acuerdo con Speicher, Arwe y Malhortra (2015), los recursos expuestos en la web se encuentran dispersos en una gran variedad de fuentes con diferente información, estructura y semántica,

lo que ocasiona problemas de heterogeneidad, lo que muestra la necesidad de contar con una solución que permita la integración de los datos y el conocimiento embebido en los mismos de manera eficiente. Es decir, se debe permitir a los usuarios acceder a los datos almacenados en fuentes de datos heterogéneas, presentando una única vista unificada de esos datos, de forma que el usuario no llegue a percibir esta heterogeneidad.

En la figura 22 se presentan las etapas que conforman la estructura del modelo planteado en esta investigación. El origen de este

Figura 22. Estructura del modelo de datos abiertos enlazados para la recuperación de información



Fuente: elaboración propia, 2021.

modelo es la necesidad informativa del usuario, pues un modelo de estas características fomenta la satisfacción de la necesidad informativa a través de la recuperación de información. A continuación se detalla cada uno de los elementos que forman parte de dichas etapas:

- Demanda informativa del usuario. Las necesidades de información son la carencia de conocimientos e información sobre un fenómeno, objeto, acontecimiento, acción o hecho que tiene una persona, razón por la cual ésta se coloca en un estado de insatisfacción que la motiva a presentar un comportamiento para buscar la satisfacción (Calva González 2004, 76). Al momento de buscar la satisfacción informativa, el usuario requiere de datos e información que lo ayuden a resolver su duda o responder a sus interrogantes. Este proceso marca el inicio de la interacción del usuario con los datos abiertos enlazados.
- Identificación de la fuente de datos. Como se ha manifestado con anterioridad, las fuentes de datos pueden ser ampliamente diversas. En este sentido, deben contemplarse las fuentes de datos que ayuden a resolver una demanda informativa o a satisfacer una necesidad de información. Así, los estudios de usuarios pueden ser una herramienta muy útil para conocer las demandas de la comunidad y ajustarlas a las variables que permitan identificar una determinada fuente de datos susceptible de incorporarse en una latente sistematización.
- Procesamiento de datos. En esta etapa intervienen procesos como la limpieza de los datos, que tiene el objetivo de eliminar inexactitudes y errores que pudieran existir en los datos al momento de ser registrados; involucra tareas relacionadas con el mapeo semántico de los datos mediante el uso de vocabularios y ontologías que permitan la construcción de vínculos semánticos entre ellos; además se llevan a cabo tareas para la asignación de metadatos que permitan representar los datos que fueron colocados en un

determinado dominio. El procesamiento de datos es crucial para establecer la conexión semántica de los datos; por lo tanto, es una etapa que permitirá realizar consultas complejas entre datos que previamente han sido manejados para ser consumidos por el usuario final.

- Sistematización. Es el proceso que engloba las normas y los procedimientos que harán posible el funcionamiento de los datos abiertos enlazados en el contexto de la recuperación de información. En esta etapa se contempla la adaptación de los datos que fueron previamente procesados, a través de una arquitectura que permite la búsqueda y el acceso a los datos y a la información que está contenida en el sistema. En esta etapa se valora la usabilidad y accesibilidad a los datos, mediante el funcionamiento del sistema, el cual deberá adaptar estrategias y mecanismos idóneos para recuperar y visualizar los datos que han sido procesados y colocados en un contexto relacionado.
- Acceso. Como se ha mencionado con anterioridad, la accesibilidad a los datos mediante el sistema en cuestión deberá estar libre de restricciones económicas, legales y técnicas. Esta condición otorgará la posibilidad de establecer un medio para satisfacer demandas informativas, apegado a los principios del acceso universal a la información. Para este cometido, existen en la actualidad estándares y lineamientos basados en licenciamientos que permiten a los datos contar con la libertad para ser reutilizados por los usuarios; sin embargo, las cuestiones de acceso abierto a la información y los datos se encuentran en constante debate académico y sociocultural.
- Preservación. Es el proceso que garantiza el acceso en el futuro a los datos, independientemente de los cambios que puedan existir en el entorno donde se ubican y la información que los rodea. En este proceso se contempla no solo la preservación de los datos, sino de las vinculaciones que se establecen entre ellos. Es la etapa más compleja del modelo, pues representa el cúmulo de esfuerzos para asegurar

> que los datos vinculados puedan seguir estableciendo relaciones de significado entre las piezas de información que forman parte del universo de información. La preservación es un proceso planificado, fundamentado en el uso de políticas que tienen el objetivo de propiciar el acceso efectivo a los datos y los recursos de información que se representan en un determinado contexto, esto acorde a la demanda y la satisfacción de las necesidades informativas del usuario.

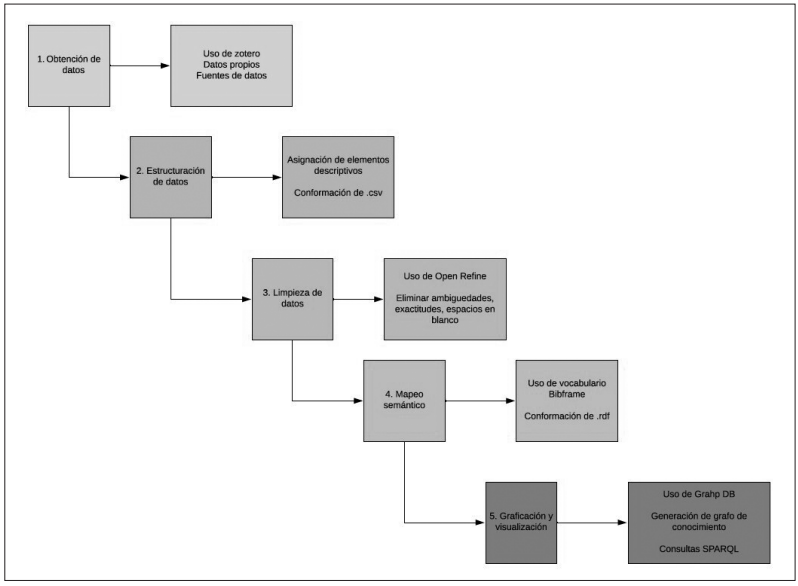
La estructura de este modelo está fundamentada en un proceso pragmático que involucra el desarrollo de datos abiertos enlazados para ponerlos a disposición del usuario final. Se trata de una estructura que está influenciada por el uso de normas y estándares que remiten a un corpus teórico de la organización de la información y el conocimiento, pero también a la naturaleza de los estudios de la información. Por lo tanto, la utilidad de este modelo no puede concebirse sin la función y utilidad de una metodología explícita que permita poner en marcha la generación de datos abiertos enlazados en un sistema que fomente la recuperación de información.

Metodología para el procesamiento de los datos

La metodología expone los pasos a seguir para el desarrollo de los datos abiertos enlazados. El primer paso para establecer esta metodología consiste en seleccionar la fuente de los datos para llevar a cabo la representación de la recuperación de información mediante herramientas de índole semántica.

En la figura 23, se encuentra el esquema que expone la metodología en cuestión. Cada una de las etapas de esta metodología está dirigida a la conformación de un grafo RDF mediante el cual sea posible visualizar y recuperar los datos que han sido procesados mediante el análisis de *datasets*. Cabe señalar que el procesamiento de datos se desarrolla a partir del uso de *software* libre de índole semántica.

Figura 23. Esquema de la metodología para el desarrollo de datos abiertos enlazados



Fuente: elaboración propia, 2021. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthKBspgNxZJzIrRrTOA?e=T5o02j.

La etapa uno, referente a la obtención de datos, pone de manifiesto el uso de datos abiertos que permitan aplicar los principios de Linked Open Data en sus estructuras. Fuentes como GitHub (https://github.com/), Zenodo (https://zenodo.org/), re3data (https://www.re3data.org/) y Dryad (https://datadryad.org/stash) permiten descargar datos de manera abierta y libre de restricciones.

Muchos de los datos colocados en estas fuentes carecen de normalización e integridad. La integridad de los datos, junto con la confidencialidad y su disponibilidad, es uno de los tres aspectos fundamentales de la seguridad de los datos. La integridad consiste en garantizar que los datos sean y sigan siendo fiables y que no hayan sido manipulados o alterados por error.

Las fallas de *hardware* o *software*, los errores humanos y los actores malintencionados pueden ser amenazas para dicha integridad.

Este puede ser un problema particularmente crítico cuando se trata de enormes cantidades de datos, pues debido al volumen y la variedad de datos almacenados y procesados, los problemas de integridad se manifiestan en mayor medida.

La integridad de los datos se ocupa de cuestiones como la "confianza" y la "idoneidad para su uso" (Lagoze 2014, 104). Incluso cuando los datos se han recopilado, almacenado y procesado correctamente, los problemas de representatividad y calidad de los datos pueden hacer que las conclusiones y su uso efectivo no sean tan confiables (Lazer *et al.* 2014).

En esta etapa de la metodología, se contempla el uso de gestores como Zotero (https://www.zotero.org/), que permitan recolectar los datos de manera orgánica e interoperable mediante un proceso de cosecha. La cosecha de datos significa obtener los datos y la información del recurso en línea como bases de datos, repositorios y plataformas. Este concepto es intercambiable con el *web scraping*, *web crawling* y la extracción de datos. La cosecha de datos es el proceso para extraer datos valiosos de las fuentes y ponerlos en su base de datos en un formato específico y estructurado.

Por lo tanto, en la etapa dos, denominada *estructuración de los datos*, se toma en consideración la organización y representación que los datos abiertos enlazados deben tener para llevar a cabo su procesamiento. En este sentido, una estructura de datos es una forma particular de organizar y almacenar datos en una computadora de manera que se pueda acceder a ellos y modificarlos de manera eficiente. Más precisamente, una estructura de datos es una colección de valores de datos, las relaciones entre ellos y las funciones u operaciones que se pueden aplicar en ellos para cumplir un propósito específico. Para el análisis de datos, es importante comprender que existen tres tipos comunes de ordenaciones:

- Datos estructurados. Son datos que se adhieren a un modelo de datos predefinido y, por lo tanto, son fáciles de analizar. Los datos estructurados se ajustan a un formato tabular con relación entre las diferentes filas y columnas. Ejemplos

comunes de datos estructurados son archivos de Excel o bases de datos SQL. Cada uno de éstos tiene filas y columnas estructuradas que se pueden ordenar. Los datos estructurados dependen de la existencia de un modelo de datos, un modelo de cómo se pueden almacenar, procesar y acceder a ellos. En un modelo de datos, cada campo es discreto y se puede acceder a ellos por separado o conjuntamente con datos de otros campos. Esto hace que los datos estructurados sean extremadamente poderosos: es posible agregar rápidamente datos de varias ubicaciones en la base de datos. Los datos estructurados se consideran la forma más "tradicional" de almacenamiento de datos, ya que las primeras versiones de los sistemas de gestión de bases de datos (DBMS) podían almacenar, procesar y acceder a datos estructurados.

- Datos no estructurados. Los datos no estructurados son información que no tiene un modelo de datos predefinido o no está organizada de una manera predefinida. La información no estructurada suele contener mucho texto, pero también puede contener datos como fechas, números y hechos. Esto da como resultado irregularidades y ambigüedades que dificultan la comprensión del uso de programas tradicionales en comparación con los datos almacenados en bases de datos estructuradas. Los ejemplos comunes de datos no estructurados incluyen archivos de audio, video o bases de datos sin SQL. La capacidad de almacenar y procesar datos no estructurados ha crecido enormemente en los últimos años con la llegada al mercado de muchas tecnologías y herramientas nuevas que pueden almacenar tipos especializados de datos no estructurados. MongoDB, por ejemplo, está optimizado para almacenar documentos. Apache Graph, como ejemplo opuesto, está optimizado para almacenar relaciones entre nodos. La capacidad de analizar datos no estructurados es especialmente relevante en el contexto de Big Data, ya que una gran parte de los datos en las organizaciones no están estructurados. Piense en

imágenes, videos o documentos PDF. La capacidad de extraer valor de los datos no estructurados es uno de los principales impulsores del rápido crecimiento del Big Data.

- Datos semiestructurados. Los datos semiestructurados son una forma de datos estructurados que no se ajusta a la estructura formal de los modelos de datos asociados con bases de datos relacionales u otras formas de tablas de datos, pero que, no obstante, contienen etiquetas u otros marcadores para separar elementos semánticos y hacer cumplir las jerarquías de registros y campos dentro de los datos. Por lo tanto, también se conoce como estructura autodescriptiva. Los ejemplos de datos semiestructurados incluyen JSON y XML. La razón por la que existe esta tercera categoría (entre datos estructurados y no estructurados) es porque los datos semiestructurados son considerablemente más fáciles de analizar que los datos no estructurados. Muchas soluciones y herramientas de Big Data tienen la capacidad de "leer" y procesar JSON o XML. Esto reduce la complejidad para analizar datos estructurados, en comparación con datos no estructurados.
- Metadatos: datos sobre datos. Una última categoría de tipo de datos son los metadatos. Desde un punto de vista técnico, ésta no es una estructura de datos separada, pero es uno de los elementos más importantes para el análisis de los datos abiertos enlazados y las soluciones de Big Data. Los metadatos son datos sobre datos. Proporcionan información adicional sobre un conjunto específico de datos. En un conjunto de fotografías, por ejemplo, los metadatos podrían describir cuándo y dónde se tomaron las fotografías. Luego, los metadatos proporcionan campos para fechas y ubicaciones que, por sí mismos, pueden considerarse datos estructurados. Por este motivo, las soluciones de Big Data utilizan con frecuencia metadatos para el análisis inicial y son un punto de partida para describir los datos abiertos enlazados que se vincularán en diversos contextos de la web.

Tomando en cuenta las categorías de datos mencionadas, la metodología de esta investigación trabajará con datos estructurados que han sido agrupados y organizados mediante filas y columnas a través de un proceso de tabulación. Para ello, se opta por explotar el formato .csv, que permite asignar elementos descriptivos a los datos.

En la tercera etapa de la metodología, denominada *limpieza de los datos*, se aborda el proceso para detectar y corregir (o eliminar) registros de datos corruptos o inexactos. La limpieza de datos consiste en aplicar el proceso de preparar datos para su análisis mediante la eliminación o modificación de datos incorrectos, incompletos, irrelevantes, duplicados o con formato incorrecto. Por lo general, estos datos no son necesarios ni útiles cuando se trata de analizar datos, porque pueden dificultar el proceso o proporcionar resultados inexactos. Existen varios métodos para limpiar los datos, dependiendo de cómo se almacenen junto con las respuestas que se buscan.

La limpieza de datos no se trata simplemente de borrar información para dejar espacio a nuevos datos, sino de encontrar una manera de maximizar la precisión de un conjunto de datos sin necesariamente eliminar información. Por un lado, la limpieza de datos incluye más acciones que eliminar datos, como corregir errores de ortografía y sintaxis; estandarizar conjuntos de datos, y corregir errores como campos vacíos, códigos faltantes e identificar puntos de datos duplicados.

La limpieza de datos se considera un elemento fundamental de los conceptos básicos de la ciencia de datos, ya que juega un papel importante en el proceso analítico y el descubrimiento de respuestas confiables. El objetivo de la limpieza de datos es crear conjuntos de datos estandarizados y uniformes para permitir que las herramientas de análisis de datos e inteligencia artificial accedan fácilmente y encuentren los datos correctos para cada consulta.

Independientemente del tipo de análisis o visualizaciones de datos que se necesite, la limpieza de datos es un paso vital para garantizar que las respuestas que se generen sean precisas. Cuando se recopilan datos de varios flujos mediante entrada manual, la

información puede contener errores, introducirse incorrectamente o tener lagunas. La limpieza de datos ayuda a garantizar que los datos abiertos enlazados siempre coincidan con los campos correctos, al tiempo que facilita que las herramientas de inteligencia artificial interactúen con los conjuntos de datos para encontrar y recuperar información de manera más eficiente. Uno de los ejemplos de limpieza de datos más comunes es su aplicación en repositorios de datos.

Por otra parte, el mapeo semántico de datos (véase etapa 4 del diagrama de la metodología) es un paso crucial para el diseño de proyectos de migración, integración y transformación de datos. Las soluciones modernas utilizan inteligencia artificial para mapear campos de datos desde un formato de origen a un formato de destino. Como consecuencia de eso, los usuarios de los datos pueden establecer relaciones entre modelos de datos separados de fuentes o sistemas dispares. Esto tiene un impacto en el análisis de la información, la previsión y la toma de decisiones. Por lo tanto, el mapeo de datos no solo es importante para los procesos de integración de datos, sino también para el incremento de su reutilización.

En este contexto, "[...] el mapeo semántico es una estrategia visual para la expansión del vocabulario y la extensión del conocimiento al mostrar en categorías datos relacionados entre sí" (Khoii y Sharifafar 2013, 202). Estos vocabularios son muy importantes para la conformación de la web semántica y del establecimiento de los datos abiertos enlazados, sobre todo cuando se intenta alcanzar un alto nivel de interoperabilidad entre ellos.

Además, las especificaciones de mapeo de datos son particularmente valiosas en los siguientes tipos de proyectos que engloba la presencia de los datos abiertos enlazados:

- Migración de datos: cuando los datos de origen se migran a un nuevo repositorio de datos de destino.
- Integración de datos: cuando los datos de origen se envían a un repositorio de datos de destino de forma regular y las dos fuentes de datos no comparten un modelo de datos

común. La integración puede ocurrir cada hora, día, semana, mes o incluso en tiempo real, como normalmente se requiere para la integración de un sistema basado en datos abiertos enlazados.

La creación de una especificación de mapeo de datos requiere descubrir y resolver problemas potenciales antes de que se implemente el mapeo de datos. En las migraciones e integraciones de datos, cualquier cantidad de diferencias entre la forma en que se almacenan los datos puede provocar la pérdida o propiciar la representación incorrecta de los datos. Estos asuntos son de suma relevancia cuando se intenta establecer proyectos de datos abiertos enlazados de grandes magnitudes, pues la integración de múltiples plataformas requiere realizar un mapeo semántico de sus datos para alcanzar interoperabilidad en la consulta y recuperación de información.

Por ejemplo, puede ser que los datos de origen tengan un campo de texto y el repositorio de destino utilice una lista de fechas enumerada. Sin analizar los datos y proporcionar la lógica para asignar los valores de texto a los valores de lista permitidos (o iniciar los esfuerzos de limpieza de datos adecuados), es probable que se experimenten errores inesperados durante la migración de datos entre sistemas. En este sentido, el uso de vocabularios facilita el desarrollo de mapeos que permitan registrar datos de diferentes modelos.

Bibframe 2.0 (https://www.loc.gov/bibframe/docs/index.html) es un vocabulario semántico que permite establecer mapeo de datos de tipo documental, sobre todo aquellos datos que están disponibles en unidades de información y bibliotecas. El vocabulario BIBFRAME consta de clases y propiedades RDF. Las clases incluyen las tres clases principales (obra, instancia y ejemplar), así como varias clases adicionales, muchas de las cuales son subclases de las clases principales (Library of Congress 2021).

Las propiedades describen las características del recurso que se describe, así como las relaciones entre los recursos. Por ejemplo: una Obra puede ser una “traducción de” otra Obra; una Instancia

puede ser una "instancia de" una Obra BIBFRAME en particular. Otras propiedades describen atributos de Obras e Instancias. Por ejemplo: la propiedad BIBFRAME "sujeto" expresa un atributo importante de una Obra (de qué trata la Obra), y la propiedad "extensión" (por ejemplo, número) expresa un atributo de una Instancia. Este vocabulario será utilizado para representar el procedimiento enmarcado en la metodología de estudio.

Una vez que los datos han sido mapeados, se da paso a su graficación y visualización (véase etapa 5 de la metodología). Esta etapa es significativa para acceder y recuperar los datos abiertos enlazados y sus respectivas relaciones de significado. Los grafos RDF son representaciones que permiten obtener una visión global de los datos con los respectivos enlaces que se gestan en el entorno digital.

La visualización de datos es la práctica de traducir la información a un contexto visual, como un mapa o un gráfico, para facilitar que el cerebro humano comprenda y extraiga información de los datos (Brush 2021). El objetivo principal de la visualización de datos es facilitar la identificación de patrones, tendencias y valores atípicos en grandes conjuntos de datos. El término se usa a menudo indistintamente con otros, incluidos gráficos de información, visualización de información y grafos estadísticos.

La visualización de datos es uno de los pasos del procesamiento de datos que establece que después de que los datos han sido recopilados, normalizados y modelados, deben visualizarse para poder sacar conclusiones. La visualización de datos también es un elemento de la disciplina más amplia de la arquitectura de presentación de datos (DPA) que tiene como objetivo identificar, ubicar, manipular, formatear y entregar datos de la manera más eficiente posible.

La visualización de datos es importante para casi todas las áreas del conocimiento. Por ejemplo, los profesores pueden usarlo para mostrar los resultados de las pruebas de los estudiantes; los científicos informáticos que exploran los avances en inteligencia artificial, o los ejecutivos que buscan compartir información con las partes interesadas.

Lo anterior también juega un papel importante en proyectos de Big Data y Data Analytics. Aunado a ello, a medida que las personas

acumulaban colecciones masivas de datos durante los primeros años de la tendencia de Big Data, necesitaban una forma de obtener rápida y fácilmente una descripción general de sus datos. Las herramientas de visualización encajaron de forma natural.

La visualización es fundamental para la analítica avanzada por razones similares. Cuando un científico de datos está escribiendo análisis predictivos avanzados o algoritmos de aprendizaje automático, es importante que pueda visualizar los resultados para monitorear los resultados y asegurarse de que los modelos estén funcionando según lo previsto. Esto se debe a que las visualizaciones de algoritmos complejos generalmente son más fáciles de interpretar que las salidas numéricas.

Bajo esta premisa, la visualización de datos se ocupa del desarrollo, el diseño y la aplicación de la representación gráfica de datos y facilita la comprensión del sentido de los datos. También se conoce como visualización científica o visualización de información.

El uso de imágenes, gráficos, tablas y mapas para comprender los datos y la información se ha utilizado durante siglos. Debido al avance de las computadoras, ahora es posible manejar y procesar una gran cantidad de datos a una velocidad muy alta. Hoy en día la visualización de datos se está convirtiendo en una combinación de arte y ciencia que traerá un cambio visible en los próximos años (Gandhi y Pruthi 2020, 18).

Así pues, un grafo RDF puede utilizarse como método general para la descripción conceptual y la visualización de los datos que están disponibles en los recursos web. Como hemos visto con anterioridad, el modelo de datos RDF se basa en la idea de hacer declaraciones sobre recursos web en forma de expresiones sujeto-predicado-objeto. Estas expresiones se conocen como triples en la terminología RDF. Un triple RDF consta de un sujeto, un predicado y un objeto. El triple RDF permite visualizar los datos abiertos enlazados que se establecen en un determinado contexto o en diversos contextos.

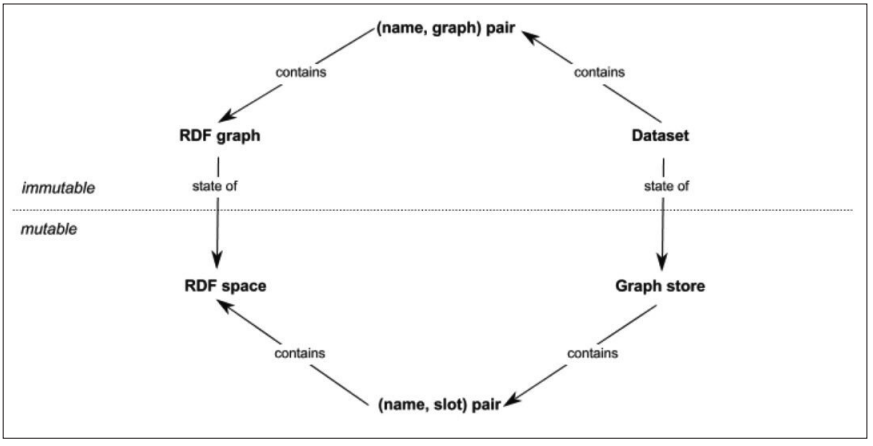
En la figura 24 puede apreciarse la interacción entre los grafos RDF y los conjuntos de datos que se utilizan para representarlos en un marco general común en donde los datos y los grafos usados

para posicionarlos visualmente son interoperables entre sí y con la normativa que se involucra dentro de su procesamiento.

Por lo tanto, para que la metodología presentada en este apartado sea susceptible de ser funcional y significativa para los intereses de la recuperación de la información, es necesario comprender la manera de hacer consultas bajo la lógica de los datos abiertos enlazados y el uso de grafos RDF. Pues los principales objetivos de la visualización son presentar, transformar y convertir los datos en una representación visual, de modo que los usuarios puedan analizar y consultar los datos de manera eficiente.

Para aumentar la efectividad de una herramienta de visualización, dicha herramienta debe permitir a los usuarios explorar dinámicamente la representación visual de los datos para que puedan comprender los datos de forma más rápida y sencilla. Los usuarios pueden procesar y analizar representaciones visuales de datos notablemente más rápido y más efectivamente que hacerlo leyendo la representación numérica o textual de los mismos datos.

Figura 24. Relaciones entre espacios RDF, grafos, conjuntos de datos y almacenes de grafos



Fuente: https://www.w3.org/2012/08/RDFNG.html, 2012. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthKEb0vETtdXhYWsUyQ?e=Bwaotk.

Desarrollo de consultas SPARQL

SPARQL es el lenguaje estándar y protocolo de consulta para datos abiertos enlazados y bases de datos RDF. Este lenguaje fue diseñado para consultar una gran variedad de datos con el propósito de extraer de manera eficiente, información oculta en datos no uniformes y almacenada en varios formatos y fuentes del entorno web.

A diferencia de SQL, las consultas SPARQL no están limitadas a trabajar dentro de una base de datos, sino que las consultas federadas pueden acceder a múltiples almacenes de datos (puntos finales). Esto es técnicamente posible porque SPARQL es más que un lenguaje de consulta. También es un protocolo de transferencia de datos basado en HTTP, donde se puede acceder a cualquier punto final SPARQL a través de una capa de transferencia estandarizada.

Los resultados de búsqueda en RDF se pueden devolver en varios formatos de intercambio de datos y las entidades RDF se identifican mediante identificadores de recursos uniformes (URI). La codificación de datos con URI permite hacer referencia a los datos de forma inequívoca en todas las aplicaciones y supera las limitaciones que plantea una búsqueda local. En consecuencia, se pueden desarrollar aplicaciones específicas de un determinado servicio de datos abiertos enlazados.

SPARQL está diseñado para habilitar datos abiertos enlazados que contribuyan al desarrollo de la web semántica. Su objetivo es enriquecer los datos al vincularlos a otros recursos semánticos globales, compartiendo, fusionando y reutilizando datos de una manera más significativa. Como resultado, los alcances de SPARQL junto con la flexibilidad de RDF pueden reducir los tiempos y esfuerzos en el desarrollo de consultas que permitan recuperar la combinación de resultados en múltiples fuentes de datos.

De esta manera, una consulta SPARQL consta de un conjunto de patrones de triples en los que cada elemento (sujeto, predicado y objeto) puede ser una variable (comodín). SPARQL puede desarrollar los siguientes cuatro tipos de consultas:

- *ASK*. Para preguntar si hay al menos una coincidencia del patrón de consulta en los datos del grafo RDF.
- *SELECT*. Para seleccionar todas o algunas de esas coincidencias en forma de tabla (incluida la agregación, el muestreo y la paginación a través de OFFSET y LIMIT).
- *CONSTRUCT*. Para desarrollar un grafo RDF sustituyendo las variables de estas coincidencias en un conjunto de plantillas triples.
- *DESCRIBE*. Para describir las coincidencias encontradas en los datos y construir un grafo RDF con los datos seleccionados.

Las principales bases de datos orientadas a grafos que admiten SPARQL tienen editores SPARQL intuitivos con funciones de autocompletado, explorador y muchas otras características que facilitan la creación de consultas para acceder a datos que no son tan evidentes con una búsqueda común.

Aunado a ello, los enfoques modernos para la gestión de bases de datos se están volviendo extremadamente populares, lo que permite discutir los alcances y limitaciones de las últimas alternativas de motores de bases de datos relacionales. Un tipo de base de datos orientadas a grafos que ha crecido rápidamente en los últimos años son los *triplestores* o almacenes RDF.

Los *triplestores* tienen la capacidad de ingerir diversos datos, proporcionando flexibilidad para su consulta con respecto a cambios de esquema y asignaciones en los cuales han sido representados. También permiten una mayor libertad para el manejo eficiente de consultas sofisticadas y respuesta a necesidades de información de mayor complejidad.

Los *triplestores* emplean soluciones inteligentes de gestión de datos que combinan la búsqueda de texto completo con el análisis de grafos y razonamiento lógico para producir resultados de recuperación de información integrales y profundos.

Las bases de datos relacionales son los caballos de batalla de muchas aplicaciones de informes y análisis de información, pero el auge del Big Data ha llevado al surgimiento de modelos

alternativos de bases de datos NoSQL. Este tipo de bases de datos se adaptan mejor a la naturaleza de los datos y al tipo de preguntas que se formularán en relación con los datos.

Las *triplestores* son una especie de base de datos NoSQL que almacenan datos en "triples" en lugar de la estructura relacional tradicional. A diferencia de las bases de datos relacionales que almacenan datos en tablas, las *triplestores* almacenan datos como declaraciones en la forma sujeto-predicado-objeto, como "Jessica enseña Ciencias de la Computación"; cada declaración se llama triple y cada triple representa el vínculo entre datos con atributos similares.

Estas bases de datos otorgan mayor flexibilidad para el manejo de datos, pues no es necesario definir un esquema por adelantado y no es necesario que entidades artificiales, como tablas, representen una relación de varios a varios. La falta de un esquema de datos predefinido significa que modificar el modelo de datos es fácil. A diferencia de las consultas que utilizan SQL, que se vuelven complicadas e ineficaces si la base de datos no se diseñó con columnas para que la búsqueda sea eficiente, las *triplestores* pueden manejar fácilmente consultas complejas, alineando los datos con un patrón en común que permita recuperarlos de una manera integral.

Además, estas bases de datos fomentan el uso compartido de los datos ya que utilizan URIs, lo que se convierte en una ventaja para los programas de análisis que necesitan reunir datos de múltiples fuentes. Aunado a ello, el descubrimiento de relaciones es una característica muy importante de los *triplestores*, ya que cuando se combinan con ontologías que definen formalmente los objetos y sus tipos de relaciones, apoyan la inferencia que permite el descubrimiento de hechos y relaciones implícitas en diversos contextos de información. Para fomentar el descubrimiento de nuevos hallazgos basados en el uso de datos abiertos enlazados, es importante el aspecto de la visualización y recuperación que se ejerce en un determinado dominio.

Visualización y recuperación con datos abiertos enlazados

La visualización es el proceso de transformar datos, información y conocimiento en presentaciones gráficas para apoyar tareas como el análisis de datos, la exploración de información, la definición de información, la predicción de tendencias, la detección de patrones, el descubrimiento de información, etcétera (Zhang 2008, 3).

En términos generales, la visualización se puede clasificar en dos categorías: visualización científica y visualización de información. La visualización científica se usa a menudo como un apoyo al sistema sensorial humano para mostrar cosas que están en escalas de tiempo demasiado rápidas o lentas para que el ojo las perciba, o estructuras mucho más pequeñas o más grandes que la escala humana, o fenómenos como rayos X o radiación infrarroja que la gente no puede percibir directamente.

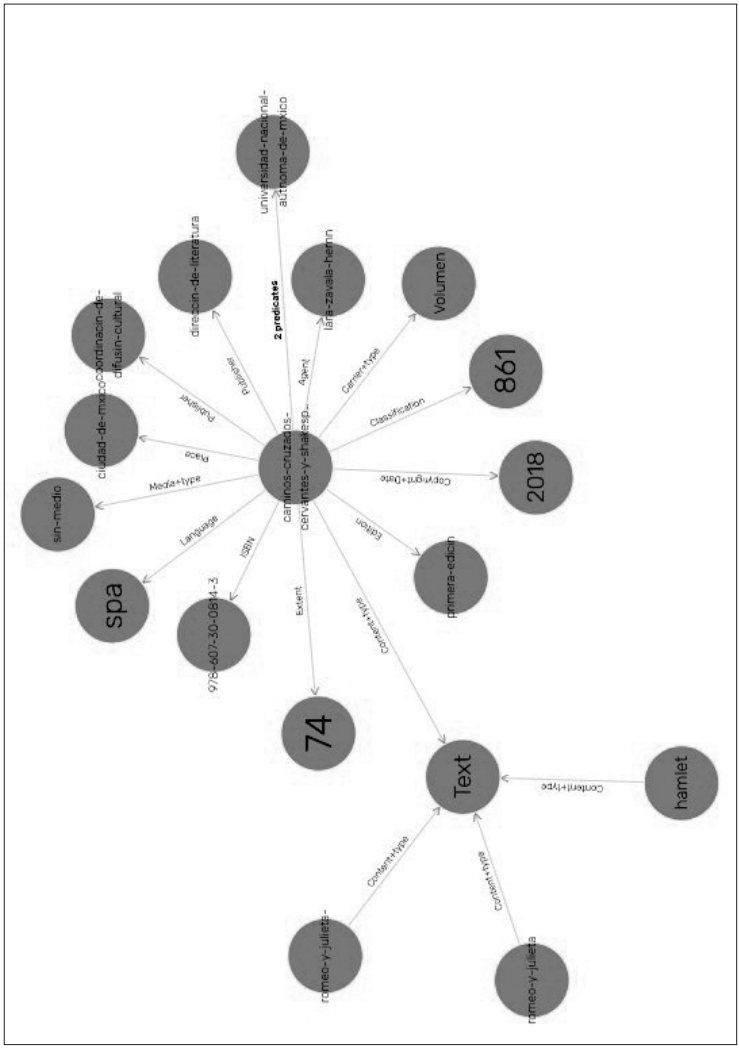
La visualización de información no tiene una estructura espacial inherente o una geometría de datos para mostrar, mientras que la visualización científica posee una estructura espacial inherente de datos para ilustrar. En otras palabras, a diferencia de la visualización científica, se debe crear una estructura espacial o marco para las relaciones semánticas entre los datos en la visualización de la información.

La visualización de información se utiliza generalmente para consultar información abstracta. Una lista incompleta de ejemplos para la aplicación de visualización de información incluye razonamiento visual, modelado de datos visuales, programación visual, visualización de recuperación de información, visualización de la ejecución del programa, lenguajes visuales, razonamiento y visualización sistemática.

En el caso de los datos abiertos enlazados, la visualización es un aspecto esencial de la recuperación de información bajo esta lógica, pues se requiere de la implementación de técnicas visuales para consultar la información que ha sido representada mediante el uso de relaciones semánticas y de grafos y tripletes.

En la figura 25 puede apreciarse un grafo RDF correspondiente a la obra *Caminos cruzados: Cervantes y Shakespeare a 400 años*

Figura 25. Ejemplo de visualización de datos abiertos enlazados mediante el uso de grafos RDF



Fuente: elaboración propia, 2021. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthKFKOTVW-k5lJIJG3w?e=Vz2pHO.

de Hernán Lara Zavala. En dicha representación puede apreciarse la funcionalidad de la visualización en el contexto de la recuperación de información, pues la visualización de datos es la presentación de datos en formato pictórico o gráfico. Permite a los usuarios de la información acceder a los análisis presentados visualmente, de modo que puedan captar conceptos difíciles o identificar nuevos patrones.

Debido a la forma en que el cerebro humano procesa la información, usar tablas o gráficos para visualizar grandes cantidades de datos complejos es más fácil que estudiar minuciosamente hojas de cálculo o informes. Entonces, la visualización es una forma rápida y fácil de transmitir y asimilar conceptos e información de manera concisa y clara para el usuario final.

En el contexto de los datos abiertos enlazados, los grafos proporcionan un medio para obtener información muy valiosa y única permitiendo construir conocimiento a partir del análisis de los datos. El análisis de grafos saca a la luz relaciones complejas, fomentado la toma de decisiones. Por lo tanto, la visualización es fundamental para ese proceso. Ser capaz de consultar las relaciones visualmente es fundamental para la comprensión, ya sean características de los datos sin procesar o características específicas destacadas por análisis de grafos.

Debido a que los grafos tienen una larga historia en el contexto de las matemáticas, las discusiones sobre el análisis de grafos y su visualización tiende a incluir muchos términos esotéricos y confusos como borde y grado. El área de estudio responsable de esto se conoce generalmente como teoría de grafos (Brath y Jonker 2015, 3).

En el análisis de grafos derivados de la aplicación de los datos abiertos enlazados, estos términos son muy importantes para describir de manera correcta la representación de los grafos en un determinado dominio. Por ejemplo, en la figura 26, se presenta un grafo construido a partir del análisis de datos abiertos de índole documental, los cuales están disponibles en el catálogo en línea de una biblioteca.

Los nodos de este grafo (representados mediante círculos) muestran los datos bibliográficos que forman parte de la descrip-

ción de un libro. A su vez, los vértices del grafo (líneas que conectan a los nodos) son las relaciones semánticas que se establecen en los datos mediante el uso de RDF. Para ello, fue necesario utilizar un vocabulario semántico, como BIBFRAME, que permitiera vincular los datos mediante relaciones compatibles con el dominio documental que caracteriza a todos los elementos.

Los grafos basados en datos abiertos enlazados han surgido como una abstracción central para incorporar el conocimiento humano en sistemas inteligentes. Este conocimiento está codificado en una estructura basada en gráficos cuyos nodos representan entidades del mundo real, mientras que los bordes definen múltiples relaciones entre estas entidades.

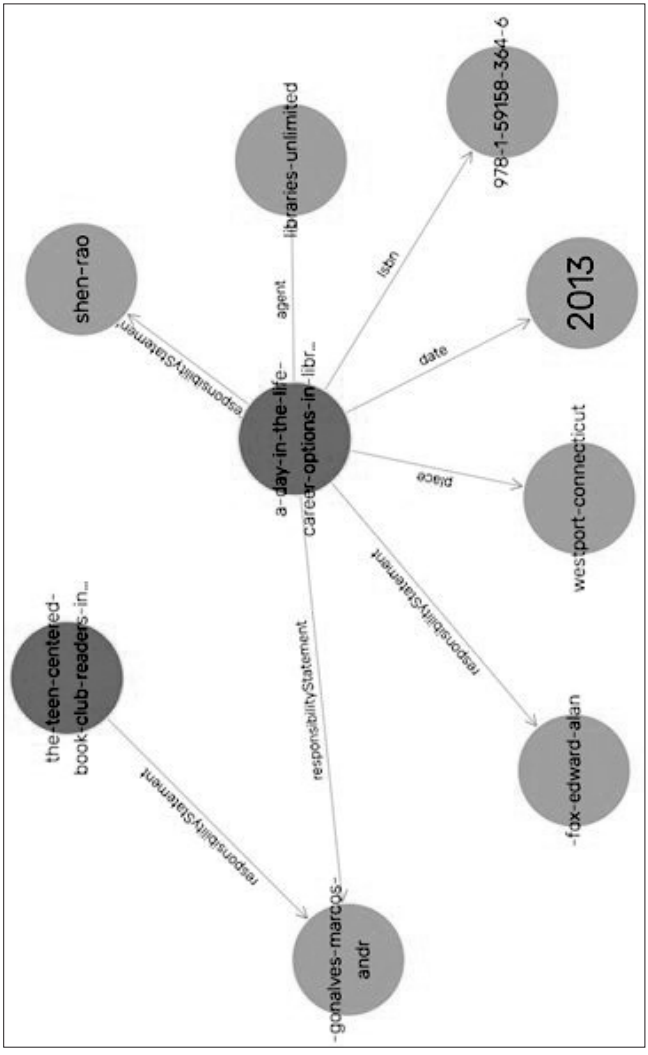
Los grafos están atrayendo la atención tanto de la industria como del mundo académico porque brindan una forma flexible de capturar, organizar y consultar una gran cantidad de datos multirrelacionales. Formalmente, un grafo se puede definir como $G = \{E, R, T\}$, donde G es un grafo múltiple etiquetado y dirigido, y E, R, T son el conjunto de nodos, bordes y triples, respectivamente.

Cada triple se formaliza como $(u, e, v) \in T$, donde $u \in E$ es el nodo principal, $v \in E$ es el nodo de cola, y $e \in R$ es el borde que conecta u y v . En el contexto semántico, un triple se formaliza como un hecho en el que (u, e, v) corresponde a (s, r, o), donde s y o son dos entidades, el sujeto y el objeto del hecho respectivamente, mientras que r es la relación que conecta s y o. La relación también puede interpretarse como el predicado que conecta al sujeto con el objeto en un grafo.

Los vocabularios semánticos pueden utilizarse para representar a las relaciones que permitan unificar los datos que forman parte de contextos con múltiples dominios. Esto ayudaría a fomentar la interoperabilidad global entre los datos, un elemento que resulta trascendental para la implementación de los datos abiertos enlazados.

Cuando se habla de la interoperabilidad de los datos abiertos enlazados, se pueden distinguir varios niveles para abordar este tema de una manera holística y orientada a la tecnología. Según Janssen, Estevez y Janowski (2014), se pueden definir los siguientes cuatro niveles principales de interoperabilidad:

Figura 26. Ejemplo de visualización de datos abiertos enlazados mediante el uso de grafos RDF



Fuente: elaboración propia, 2021. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthKFRlucRnN7TmFWT7g?e=dGddeJ.

1. Técnico: este nivel se refiere a una interconectividad basada en la red entre sistemas para poder intercambiar datos, por ejemplo, por transacción o vía transmisión en tiempo real. Al emplear enfoques de X-as-a-Service (XaaS), se pueden eliminar incompatibilidades tales como diferentes sistemas operativos o lenguajes de programación aplicados.
2. Sintáctico: este nivel se refiere al uso de estándares en términos de formatos de intercambio, por ejemplo, XML o JSON en un nivel de interfaz web, es decir, para que los servicios web intercambien datos.
3. Semántico: este nivel se refiere a reducir la ambigüedad en términos de interpretabilidad de los datos. Esto, a su vez, requiere tecnologías semánticas y metadatos bien definidos, por ejemplo, a través del uso de ontologías.
4. Pragmático: este nivel se refiere a la calidad y la confianza de una organización en general. En esta perspectiva se incluyen acuerdos de nivel de servicio (SLA) o compatibilidad al contexto en términos de significado y partes interesadas involucradas en el tratamiento y procesamiento de los datos.

Aunado a ello, la complejidad de la integración e interoperabilidad de los datos enfatiza los niveles de almacenamiento de datos, la estructura y los niveles en los que los datos se pueden integrar y operar como una sola entidad. La recopilación y el mantenimiento de grandes conjuntos de datos es costoso, por lo que las organizaciones tienden a adaptarse a las metodologías de la nube para almacenar los datos y reutilizarlos (Klischewski y Scholl 2006).

Además, los datos de fuentes heterogéneas podrían dar lugar a incoherencias en los niveles de datos, por lo que se requieren más recursos para optimizar los datos no estructurados. Los datos estructurados permiten realizar las operaciones de consulta para analizar, filtrar y utilizar estos datos para las decisiones basadas en información.

En este escenario, donde están involucrados grandes conjuntos de datos, los datos no estructurados residen en volúmenes más

altos. Esto se puede localizar utilizando los métodos de etiqueta y clasificación, lo que permite buscar los datos mediante palabras clave. Ahora bien, para la conformación de datos abiertos enlazados, también es necesario llevar a cabo tareas de transformación, pues para implementar este tipo de datos en un contexto sistematizado, el procesamiento es una tarea vital que también contempla aspectos de su visualización.

El proceso para transformar datos de diversas fuentes en datos abiertos enlazados implica la implementación de un conjunto de reglas para transformar los datos del origen al destino. Esto también implica la unión de datos de todas las fuentes, la clasificación, la derivación de valores y la aplicación de las reglas. De hecho, el uso de un vocabulario semántico responde a este tipo de reglas, las cuales pueden ser lógicas, de descripción y de representación.

Los retos en el manejo de los datos mencionados anteriormente impactan las prácticas organizacionales tradicionales para la integración e interoperabilidad de datos. Si se abordan todos estos desafíos, esto podría proporcionar un buen alcance para las organizaciones que conforman los dominios dentro de sí mismos y también se integran a los niveles de múltiples organizaciones que desean incursionar en el ambiente digital de los datos abiertos enlazados.

Internet se ha convertido en un espacio digital que fomenta la generación de abismales cantidades de datos, un entorno que hace más hincapié en la contribución, la participación, el intercambio y la comunicación entre las personas. El diseño y la organización de los sistemas de información de nueva generación presta más atención a la experiencia del usuario, que a cambio atraerá y generará demandas de información mucho más complejas.

De acuerdo con Kubernátová, Friedjungová y van Duijn (2019, 4), los sistemas de nueva generación tienen la capacidad de recomendar visualizaciones de datos acordes al perfil del usuario y la demanda informativa que manifiestan estos sistemas realizan el siguiente tipo de recomendaciones acorde a los atributos que se ejemplifican:

- Orientado a características de datos. Estos sistemas recomiendan visualizaciones basadas en características de datos.
- Orientado a tareas: estos sistemas recomiendan visualizaciones basadas en objetivos de representación, así como en características de datos.
- Orientado al conocimiento del dominio. Estos sistemas mejoran el proceso de recomendación de visualización con conocimiento de dominio.
- Orientado a las preferencias del usuario. Estos sistemas recopilan información sobre los objetivos y las preferencias de presentación del usuario a través de la interacción del usuario con el sistema de visualización.

La línea entre las diferentes categorías de sistemas de recomendación es bastante delgada y algunos sistemas pueden tener clasificaciones ambiguas, como se abordará más adelante en el apartado de integración y sistematización de los datos abiertos enlazados, pues no es posible concebir la idea de este tipo de datos sin considerar el impacto que tendrían en el contexto tecnológico y digital, que es proliferado por la presencia de sistemas con nuevas características y alcances.

Integración y sistematización

Como se ha mencionado anteriormente, uno de los objetivos de este trabajo es abordar el problema de facilitar el consumo de datos abiertos enlazados al usuario final. En este sentido, es necesario analizar cómo el enfoque metodológico planteado anteriormente puede ser factible y eficiente a través de su sistematización y acorde al punto de vista del usuario.

El objetivo del tratamiento de datos es dejar los datos listos para ser consumidos por el sistema. El dominio y los conjuntos de datos son elementos que se deben elegir antes de que comience esta fase. Para explorar la posibilidad de utilizar varios conjuntos

de datos de tripletes, es necesario construir una capa de acceso a los datos. El objetivo del acceso es hacer que los datos disponibles puedan recuperarse como un triple mediante una consulta.

Por ejemplo, en la figura 27 puede apreciarse un ejemplo de la integración de consultas de SPARQL en un entorno sistematizado, el cual está caracterizado por la división en triples de cada uno de los datos que han sido almacenados en el sistema, mediante un conjunto de datos determinado.

Después de que los datos se extraen de una fuente determinada, se procesan y almacenan localmente. Pero no es suficiente almacenar los datos, estos datos también deben integrarse. Esta es una fase importante, ya que cada conjunto de datos tiene su propio contexto, el cual permite describir sus términos en función de los atributos que presenten.

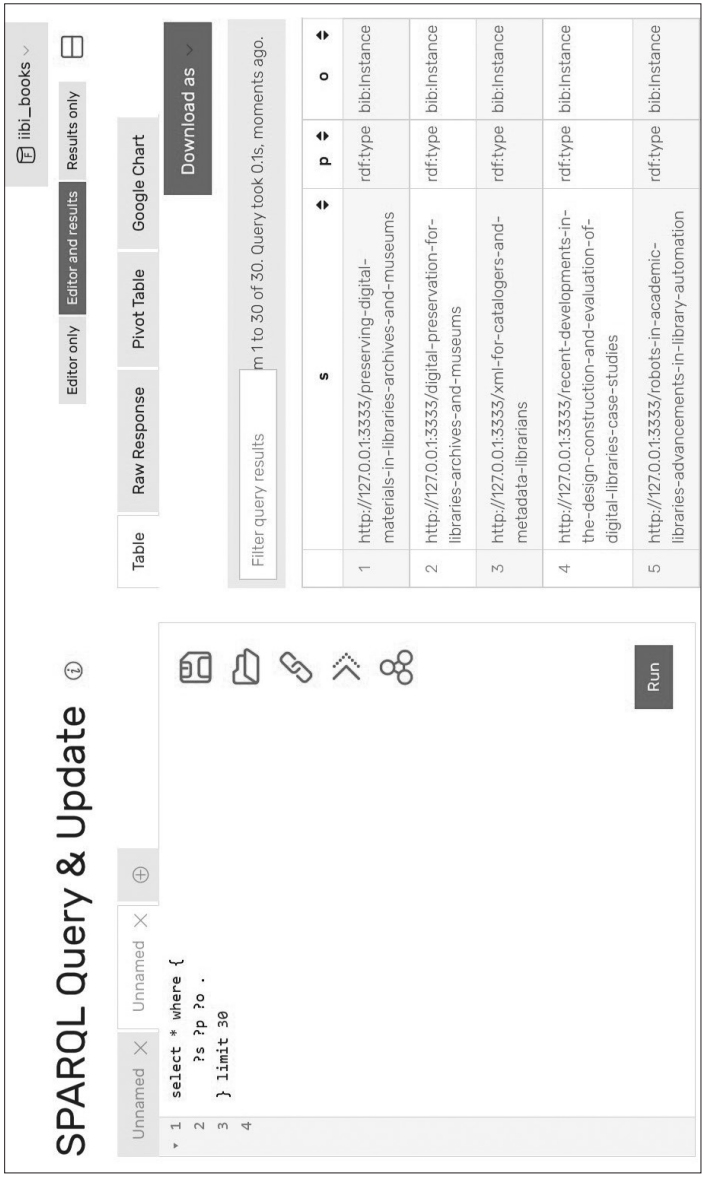
Además, en un entorno sistematizado de datos abiertos enlazados, las consultas SPARQL son capaces de expandirse y de llevar a cabo una labor de descubrimiento a través del propio método de recuperación. En la figura 28, se observa este comportamiento de las consultas en donde se ha expandido una consulta sobre un título de libro determinado. Es decir, las consultas SPARQL, además de permitir la búsqueda de datos, también propician el descubrimiento de nuevos datos incluidos en las variables de la consulta.

De esta manera, en la figura 29, se puede apreciar un grafo RDF derivado de la misma consulta SPARQL que se ha desarrollado con anterioridad. Puede observarse que la integración de los datos es una constante, pues las propias consultas permiten desarrollar grafos de una manera automática e intuitiva.

Como hemos abordado con anterioridad, SPARQL es el lenguaje de consulta estándar para recuperar y manipular la información contenida en los grafos RDF. A diferencia de los motores de búsqueda habituales, donde las consultas en lenguaje natural (LN) pueden plantearse, el uso de SPARQL requiere conocimiento sobre las entidades en el dominio que se va a consultar, así como una comprensión de la sintaxis y semántica del lenguaje.

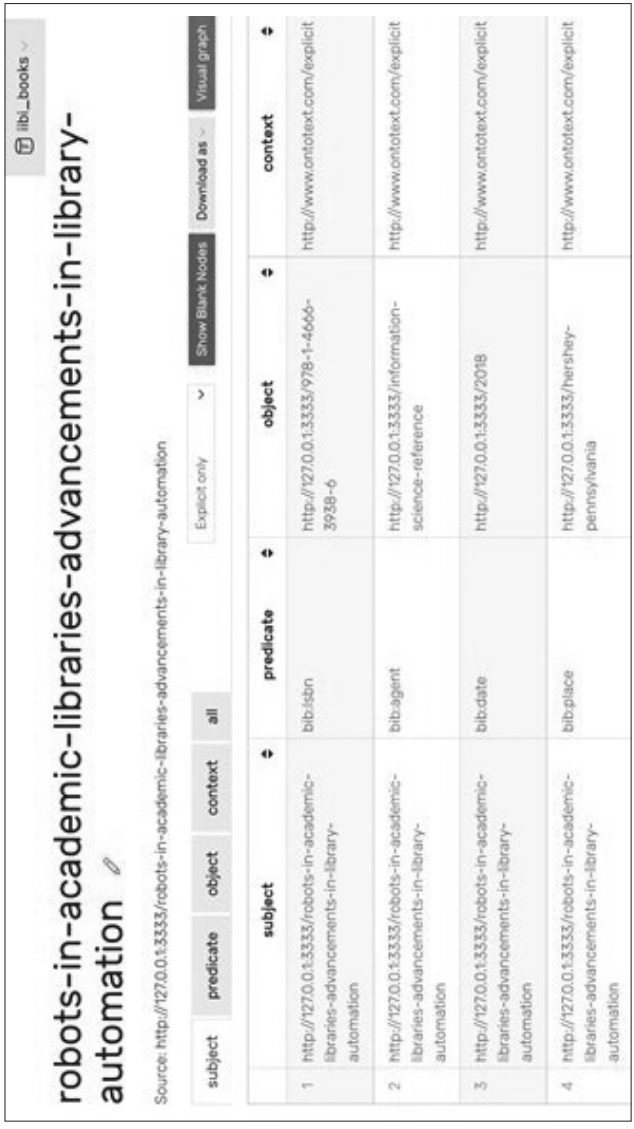
Por esta razón, el uso de las consultas SPARQL está generalmente limitado a un grupo de expertos en web semántica competente

Figura 27. Ejemplo de consulta y representación de datos abiertos enlazados en un entorno sistematizado



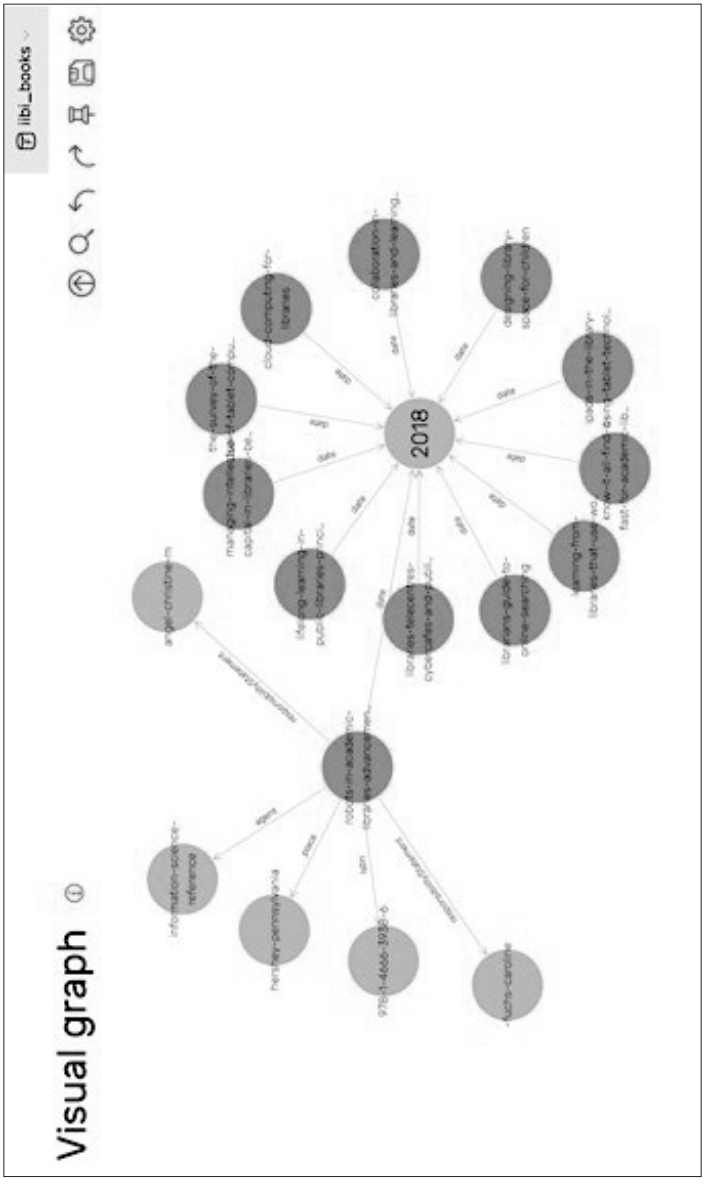
Fuente: elaboración propia, 2021. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/sAkMKIvpX0suthKFa6yBBHV34NqBmEA?e=y9bSSd.

Figura 28. Expansión de consulta de datos abiertos enlazados en un entorno sistematizado



Fuente: elaboración propia, 2021. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthKFbYtEJEUi2GobUhQ?e=TnHgOC.

Figura 29. Grafo RDF derivado de la consulta SPARQL en un dominio de libros digitales



Fuente: elaboración propia, 2021. Esta figura es para efectos ilustrativos; para verla de forma detallada, consultarla en: https://1drv.ms/u/s!AkMKlvpX0suthMF3bG6Ri6erkQWZcg?e=hyvA4N.

en el lenguaje de consulta. Difundir su accesibilidad a una audiencia más amplia y, por lo tanto, transmitir el conocimiento para el desarrollo de consultas con SPARQL permitirá traducir automáticamente las preguntas plasmadas en LN a consultas de SPARQL.

Aunado a ello, la integración de la gran cantidad de datos dispersos disponibles en la web a través de diferentes formatos (tablas HTML, CSV, JSON y, en general, información recuperada a través de servicios API) es un ingrediente clave para construir datos abiertos enlazados a gran escala. "En los estudios de la web semántica, un enfoque común para combinar información de múltiples fuentes heterogéneas explota ontologías de dominio para producir una descripción semántica de las fuentes de datos a través de un proceso llamado mapeo semántico" (Futia *et al.* 2018, 188).

El mapeo de consultas se lleva a cabo identificando y analizando las posibles asociaciones entre las palabras clave de la consulta y los términos de la base de datos. De acuerdo con Ramada, da Silva y Leitão-Júnior (2020, 5), este proceso requiere comprender el significado de las palabras clave, tanto por separado como en conjunto, y la construcción de una base de datos consulta estructurada (por ejemplo, una expresión SQL) que proporciona una coherencia interpretación de la consulta original. Por lo tanto, es necesario mapear las palabras clave para las estructuras de la base de datos, por ejemplo, relaciones, atributos y valores de atributo.

Una vez que se ha realizado un mapeo, los vínculos entre la consulta y un sistema de datos abiertos enlazados se pueden utilizar para crear perfiles de usuarios en función de las consultas que realizan. Este sistema también puede ser utilizado para enriquecer datos disponibles en la nube de datos abiertos enlazados, por ejemplo, al ver una consulta como una anotación (generada por el usuario) de los datos que han sido vinculados. Esta consulta se puede utilizar para etiquetar imágenes en las que un usuario hace clic como resultado de una búsqueda determinada. Este procedimiento refleja el método intuitivo e interactivo que se ejerce cuando se consumen datos en un sistema basado en la lógica de los datos enlazados.

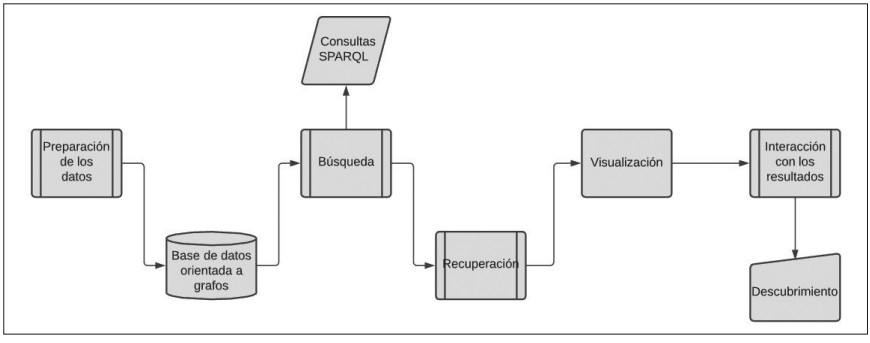
El paradigma de los datos abiertos enlazados tiene como objetivo transformar a la web y encaminarla hacia el desarrollo de nue-

vas prácticas de estructuración de la información, con un impacto comparable al auge de las bases de datos relacionales. Por lo tanto, pensar el desarrollo de sistemas de datos abiertos enlazados, también supone el desarrollo de una web mejor contextualizada y organizada.

De esta manera, en la figura 30 se aprecia el esquema del modelo de recuperación de información mediante la implementación de datos abiertos enlazados. La preparación de los datos es la fase inicial de este modelo, la cual está relacionada con el uso de la metodología para el procesamiento de datos que se ha planteado con anterioridad en este capítulo. Una vez que los datos han sido procesados, son susceptibles de almacenarse en una base de datos orientada a grafos, la cual conserva los datos en forma de triples para su posterior búsqueda a través de una interfaz de usuario.

La búsqueda de datos abiertos enlazados es un proceso que se desarrolla a través de consultas SPARQL. Los usuarios del sistema interactúan con un motor de búsqueda que conserva la lógica de este tipo de consultas. Una vez que las consultas son efectuadas, se manifiesta el proceso de recuperación de información, en donde se arrojan resultados relacionados con los datos y su enriquecimiento semántico. Las relaciones semánticas entre los datos pueden apreciarse a través de un proceso de visualización; tal y

Figura 30. Modelo esquemático de la recuperación de información con datos abiertos enlazados



Fuente: elaboración propia, 2021.

como abordamos con anterioridad, este proceso motiva la interacción del usuario con los resultados obtenidos a través de proceso de recuperación, lo que a su vez propicia una labor de descubrimiento de información basado en el uso de los datos.

El descubrimiento basado en datos es un proceso que puede resultar fortuito, pues permite conocer información que en primera instancia el usuario desconocía. Este proceso comienza a despertar interés en el área de la generación de sistemas de información, lo que propicia la evolución de dichos sistemas en una nueva generación de ellos. Los sistemas de información de próxima generación ofrecen una toma de decisiones autónoma a través del ofrecimiento de datos complejos en tiempo real e involucran tareas de aprendizaje automatizado, por ejemplo, mediante el uso de agentes de *software*.

Los objetivos de los sistemas de próxima generación son lograr un nivel amplio de interoperabilidad entre sistemas al incorporar atributos uniformes entre dichos sistemas, lo cual les permita aprovechar las tecnologías y componentes compatibles en la web. Esto fomentaría la integración de contenidos, recursos y datos globales existentes en diversas infraestructuras, específicamente aquellas destinadas a la identificación de, por ejemplo, contribuciones, datos de investigación, colaboradores, instituciones y financiadores de proyectos. Los sistemas de próxima generación buscan fomentar la aparición de servicios de valor agregado que utilizan estos atributos uniformes para apoyar el descubrimiento, acceso, anotación, curación en tiempo real, uso compartido, evaluación de calidad, transferencia de contenido, análisis, rastreo de procedencia, etcétera.

La explosión global de datos es una gran oportunidad para hacer las cosas de nuevas formas y hacer que el mundo digital funcione mejor. Sin embargo, al mismo tiempo, este tsunami de datos amenaza con abrumar y alterar las fuentes de datos que son el corazón de la nueva economía de las aplicaciones y los sistemas de información.

En este sentido, es preciso entender que un sistema de información de próxima generación deberá integrar varios datos de

manera aislada; se puede acceder directamente a la base de datos conformada e integrada para llevar a cabo una consulta de información completa en lugar de consultar varios sistemas de información de una forma manual. Eso significa que los datos que se almacenan en la base son una integración orgánica y un almacenamiento asociado de varios datos heterogéneos, pero no solo se acumulan en la base de datos de manera simple y aislada, sino que interactúan interoperablemente para ofrecerle al usuario una visión global de los recursos y contenidos que se relacionan con un dato determinado.

En este sentido, la sistematización e integridad de los datos abiertos enlazados son conceptos muy importantes para el fomento de la accesibilidad de dichos datos en el ambiente digital. La integridad de los datos significa que los datos se extrajeron por completo de la fuente que se conecta con el sistema; de esta manera es posible establecer relaciones entre los datos y facilitar su procesamiento para mejorar su eficiencia en respuesta a una demanda informativa determinada.

Debido a que los diferentes sistemas de información se desarrollan para necesidades específicas, se construyen con poca consideración en las características que presentan las fuentes de los datos, por lo que existen diferencias semánticas entre los diferentes datos que se almacenan en ellas. Estas diferencias semánticas provocarán una variedad de generación de información incompleta o incluso incorrecta, una serie de conflictos en la semántica de los nombres (diferentes nombres representan el mismo concepto) y un modelo diferente para expresar la misma información en una estructura.

En este sentido, el conflicto semántico traerá redundancia en los resultados de la integración de datos. Por lo tanto, la integración de los datos debe basarse en el modo de conversión de datos y las reglas para una estructura de datos unificada y un campo de transcodificación semántica; por esta razón los datos abiertos enlazados adoptaron como principio la utilización de la norma RDF.

Dado que los contenidos de la base de datos pertenecen a diferentes unidades, cada sistema de información tiene sus propios

derechos de usuario y modos de administración, lo que impide administrar los datos de una manera integrada; por este motivo, la interoperabilidad es trascendental para modificar parámetros entre los datos que han sido enlazados con anterioridad.

La naturaleza y el volumen de los datos han afectado a los cambios tecnológicos de los últimos años, que a su vez constituyen un problema importante en las técnicas de gestión y recuperación de datos. La comunicación de información ha cambiado por completo casi todos los aspectos de nuestras vidas. Una vez que se pensó como un sueño poco realista, Data finalmente se ha hecho realidad, lo que ha permitido que las computadoras nos comprendan e interactúen con nosotros mientras procesan su pensamiento.

La recuperación de información en los próximos años tendrá una fuerte influencia de la analítica de datos, la minería de datos, la inteligencia artificial y el aprendizaje automático. Los datos abiertos enlazados son una faceta de la aplicación de elementos que forman parte de estos campos de conocimiento y su integración en los sistemas de información. En este sentido, el proceso de aprendizaje automático es similar al de la minería de datos. Ambos sistemas buscan entre los datos para encontrar patrones.

Sin embargo, en lugar de extraer los datos para la comprensión humana, como hacen las aplicaciones de minería de datos, el aprendizaje automático utiliza esos datos para detectar patrones en los datos y ajustar las acciones del programa en consecuencia. Los algoritmos del aprendizaje automático se clasifican a menudo como supervisados o no supervisados. Los algoritmos supervisados pueden aplicar lo que se ha aprendido en el pasado a nuevos datos. Los algoritmos no supervisados pueden extraer inferencias de conjuntos de datos.

En este trabajo, hemos manifestado la gran relevancia que los sistemas de próxima generación tendrán en la interacción con el usuario, pues estos sistemas no pueden concebirse sin la presencia humana que permite generar las demandas de información y trasladarlas en un contexto más significativo, pues la propia semántica de los datos obligará a contar con sistemas más sofisticados, pero también más humanos. El reto consiste en obtener

sistemas de información más inteligentes, pero también más cercanos a las complejas demandas informativas que se manifiestan día con día.

Por lo tanto, la recuperación de información deberá considerarse un proceso en constante actualización y aplicable en los diferentes contextos de información y datos de diferentes características y naturaleza. Sin embargo,

> [...] para evitar una cantidad excesiva de adaptaciones aisladas, las interacciones entre documentos y diferentes fuentes de datos se requiere de una capa semántica genérica e interoperable sobre las bases de datos. Tales estructuras harán más accesibles los datos a los motores de búsqueda genéricos, mediante búsquedas de palabras clave y consultas en lenguaje natural (Ginsparg 2009, 203).

La posibilidad de tener los datos accesibles de esa manera alentará a que más administradores de bases de datos proporcionen interfaces semánticas locales, con lo que incrementarán su integración en la red global de datos y amplificarán los beneficios comunitarios de disponer de un acceso abierto a los recursos, contenidos y a los propios datos.

Por lo tanto, la recuperación de la información en este sentido no podrá concebirse sin la concepción de un entorno digital cada vez más abierto, libre de restricciones y susceptible de integrarse en nuevas modalidades para satisfacer las demandas informativas de las comunidades de usuarios.

Bajo esta premisa, el movimiento de acceso abierto se enfrenta a un aspecto ampliamente criticado de las prácticas de publicación tradicionales: que la investigación no puede ser de libre acceso para todos. Por ejemplo, los científicos de los países más pobres tienen ahora un mejor acceso a los últimos avances en su campo y están en menor desventaja.

El público en general (que, según muchos, tiene todo el derecho a leer investigaciones financiadas con fondos públicos) también puede aprender más sobre ciencia. Hacer que la ciencia sea

abierta puede ayudar a abordar la desconfianza del público en los valiosos avances científicos, como la manipulación genética y el desarrollo de vacunas.

Los desarrollos tecnológicos como blogs, motores de búsqueda y redes sociales han revolucionado las publicaciones en línea. Estas herramientas están aumentando rápidamente la exposición de la información publicada en línea. Los científicos se benefician de esta exposición más amplia de su trabajo. Por lo tanto, el futuro de los datos abiertos enlazados estará ampliamente relacionado con los principios y la prospectiva del movimiento del acceso abierto.

La publicación en acceso abierto promueve el intercambio de información. Los recursos de información publicados en línea no limitan lo que se puede imprimir. Los artículos en línea pueden presentar una mayor riqueza de información, como videos de técnicas de imágenes de última generación.

Además, se pueden cargar y compartir grandes conjuntos de datos sin procesar. Este intercambio de datos promueve la ciencia abierta y fomenta la confiabilidad y reproducibilidad. Además, publicar en revistas de acceso abierto será un aliciente para continuar con la conformación de un entorno digital cada vez más accesible.

A su vez, las direcciones futuras en la recuperación de información se pueden considerar revisando las teorías que están actualmente bajo consideración activa y estudiando tecnologías que probablemente prevalecerán en el futuro previsible. Las teorías de mayor interés se ocupan de los sistemas de procesamiento del lenguaje natural que utilizan representaciones ampliadas del contenido de la información, es decir, desde la óptica de la semántica de la información.

Bajo este escenario, las nuevas tecnologías influyen de manera sobresaliente en la conceptualización de una nueva forma de concebir la recuperación de la información. Uno puede esperar que los nuevos avances teóricos podrían, con el tiempo, acoplarse a las nuevas tecnologías, lo que da oportunidad a la implementación de nuevos sistemas flexibles que sean de apoyo al usuario final.

Como hemos manifestado con anterioridad, los sistemas de próxima generación deberán implementar interfaces útiles para el usuario, es decir, que sean capaces de interactuar con los diferentes comportamientos que manifiestan los distintos tipos de usuarios. En el contexto de los datos abiertos enlazados, estas interfaces serán de gran trascendencia para facilitar el vínculo entre la demanda informativa del usuario y los datos, contenidos y recursos que se encuentran en distintas fuentes del ambiente digital.

Aunado a ello, de mayor interés podrían ser los sistemas que pudieran decidir por sí mismos, dada una consulta en particular, cuál de los muchos recursos debería ser utilizado para proporcionar la respuesta específica a la demanda de usuario. En este modo, el sistema sería capaz de conocer las capacidades de diferentes recursos y ayudar al usuario a tomar la decisión correcta sobre qué recurso de información utilizar.

El diseño de tales sistemas de información inteligentes está todavía algo más allá del estado de la técnica. Sin embargo, se están logrando avances considerables en el desarrollo de sistemas de propósito especial que faciliten el acceso a muchos recursos diferentes y alivien al usuario de los detalles operativos que normalmente deben ser dominados antes de que se pueda obtener acceso a la información.

La prospectiva de la implementación de los datos abiertos enlazados en el entorno de la sistematización aún está en una fase de desarrollo y adaptación. Se estima que las restricciones legales, económicas y técnicas que imperan en el movimiento del acceso abierto se flexibilicen y de esta manera marquen la pauta para poder incursionar en el progreso de los sistemas de próxima generación. Este suceso permitiría dar una mayor y mejor continuidad al estudio teórico de los datos abiertos enlazados en el contexto de la sistematización.

Consideraciones finales

El estudio de los datos abiertos enlazados y su aplicación en el proceso de recuperación de información ha permitido descubrir los principios teóricos y metodológicos que bien podrían considerarse para llevar a cabo su sistematización. El entorno digital actual exige contar con procesos más eficientes y eficaces para obtener información y se estima que los datos abiertos enlazados pueden ayudar a conseguir este propósito.

Sin embargo, para ello será necesario tener en cuenta las características del usuario final que ayuden a perfilar estrategias específicas para el desarrollo de consultas de información más sofisticadas. En este sentido, la consulta de datos abiertos enlazados requerirá de un usuario mejor formado con habilidades informativas que le permitan interactuar de mejor manera con la lógica planteada en el proceso de recuperación de información mediante datos abiertos enlazados.

Los usuarios de la información presentan cada día nuevas y complejas demandas de información. Se estima que el modelo planteado en esta investigación sirva como una metodología para establecer nuevas técnicas para la obtención de información, sobre todo de una manera significativa y acorde a las necesidades de dicho usuario.

El modelo expuesto en esta investigación es desarrollado por y para el usuario final; por ende, una variable muy importante para su implementación será tomar en cuenta el comportamiento

informativo del usuario, pues estos elementos permitirán definir con mayor claridad las características del tipo de datos y de información que el usuario puede consumir para satisfacer su necesidad informativa.

Los datos abiertos enlazados pueden ser utilizados para cubrir una amplia gama de necesidades informativas, desde aquellas elementales relacionadas con las actividades de la vida cotidiana, hasta la incursión en actividades académicas, o bien transitar en necesidades más complejas relacionadas con el uso de información especializada para el desarrollo de trabajos altamente demandantes o especializados. En suma, la recuperación de información con datos abiertos enlazados es multivariable y altamente contextual; su uso depende del tipo de datos, el contenido y los recursos que serán colocados en un dominio determinado.

Ahora bien, las relaciones que se establecen entre datos, contenidos y recursos de información dan la posibilidad de descubrir fuentes de conocimiento con atributos similares entre sí. Esto es posible gracias al análisis semántico que puede realizarse a la información contenida en todos estos elementos.

La idea de construir una nube de datos abiertos enlazados recae en la oportunidad de desarrollar una gran red de información interconectada semánticamente. En el modelo que hemos planteado en esta investigación se aborda la interacción entre recursos y contenidos disponibles en distintas fuentes del entorno digital, con la particularidad de utilizar a los datos que se utilizan para representar sus atributos y de esta manera establecer relaciones significativas entre ellos.

Las relaciones semánticas entre datos son interconexiones que se establecen mediante el análisis del significado de los elementos alfanuméricos que son utilizados para representar a los datos. En el modelo que se ha planteado en esta investigación, dichas relaciones son un puente entre los datos, los contenidos y los recursos que pueden estar disponibles en una fuente en particular.

Una relación semántica puede ser multivariable; es decir, conectar datos de diferentes fuentes, pero con atributos similares. Por ejemplo, conectar obras con sus respectivas manifestaciones

y expresiones. En el dominio documental este tipo de relaciones fomentan en mayor medida la organización y sistematización de la información, ya que permiten tener un acercamiento preciso con las derivaciones de las diversas obras intelectuales que se han generado en el universo de información.

En suma, los datos abiertos enlazados aplicados al proceso de recuperación de información permiten obtener una panorámica de las conexiones existentes entre diversas piezas de información. Este tipo de recuperación es trascendental para las áreas en donde el descubrimiento intensivo basado en investigación requiere localizar, identificar y acceder a información precisa en cortos periodos de tiempo. Además, puede concebirse también como una metodología para el análisis del comportamiento de la información en diferentes contextos.

La presente investigación se enmarca en un entorno caracterizado por el incremento excesivo de datos, una visión que los expertos en datos han denominado la era del "Dataismo". En un contexto datificado como el que acontece en la actualidad, es necesario contar con mecanismos y estrategias que ayuden a los usuarios a enfrentar el tsunami de datos. De hecho, los datos abiertos enlazados han surgido como propuesta que busca controlar y organizar de mejor manera a estos datos.

Se estima que una mayor organización de los datos propicia su mejor utilización por parte de los usuarios, lo que fomenta mejores prácticas al momento de compartir y utilizar los datos en el ambiente digital. El papel de los datos abiertos enlazados no se reduce únicamente a su aplicación en el proceso de recuperación de información, sino que su visión puede ampliarse para ayudar a resolver problemas relacionados con el uso ético de la información y la desinformación que se presenta con gran ahínco en la actualidad.

Por lo tanto, la recuperación de información mediante el uso de datos abiertos enlazados deberá ser un proceso altamente certero que propicie confiabilidad para el uso de la información que se está conectando y la cual será utilizada por el usuario en sus labores de descubrimiento, pues la certeza en el uso de la información es

una constante en un contexto que se encuentra inmerso de abismales cantidades de datos.

Aunado a ello, la revolución de los datos pone de manifiesto la incursión del profesional de la información en temas que requieren una mayor comprensión de la generación, problemas y utilidad que los datos pueden tener para ayudar a resolver demandas informativas. Sobre todo, se debe tomar en cuenta que los usuarios actuales están experimentando una transformación social que impacta en su comportamiento informativo, sobre todo relacionado con la interacción que tienen con datos de diversas tipologías y naturaleza, mediante dispositivos que se encuentran proliferados de datos.

En tiempos adversos de pandemia, el uso efectivo y correcto de los datos es un problema que vislumbra la necesidad de alfabetizar a los individuos en el uso de los datos. En este contexto los datos abiertos enlazados y su aplicación en la recuperación de información requieren de usuarios formados en el uso de datos y de profesionales que sepan cómo obtener el mayor provecho a los mismos.

Aunado a ello, la alfabetización en datos es un tema demasiado recurrente en la actualidad, pues se requiere que los usuarios desarrollen su forma de pensar en términos de datos; es decir, que adquieran la capacidad para interpretar, analizar y argumentar datos. En este sentido, la recuperación de información basada en datos abiertos enlazados es un proceso que requiere de un usuario capacitado en el manejo y uso de los datos para satisfacer su demanda informativa, pues, por sí solos, los datos no tienen la capacidad de resolver una necesidad informativa específica.

El desarrollo actual hacia una "sociedad basada en datos" y la proliferación de la "datificación" mediante la cual los fenómenos en la naturaleza y la sociedad se cuantifican y cualifican cada vez más con el objetivo de obtener nuevos conocimientos a través del análisis de datos plantean un desafío social.

La datificación y el creciente volumen de datos disponibles públicamente en la web son en parte el resultado de nuestro amplio uso de tecnologías digitales, que se acompaña de consecuencias

tanto positivas como negativas. Por un lado, la datificación tiene el potencial de ayudar en nuestra transición hacia un futuro más sostenible, pero también existe el riesgo de que los datos se puedan utilizar contra ciudadanos débiles o como medio de evaluaciones y vigilancia electrónicas no deseadas. Estos desarrollos abren la cuestión de qué competencias son necesarias para la ciudadanía activa en una sociedad donde los datos son centrales en la toma de decisiones tanto individuales como colectivas.

Bajo esta premisa, resulta trascendental que el profesional de la información incursione en el estudio de los datos para ofrecer a los usuarios de la información las herramientas necesarias para interactuar en el contexto datificado actual. Las prácticas y herramientas de la alfabetización de datos están cambiando a un ritmo rápido, por lo que la alfabetización de datos se convierte en una de las "alfabetizaciones de lo digital" que requieren el desarrollo de competencias de por vida.

Las definiciones más específicas de alfabetización de datos que se han utilizado como base para la intervención educativa van desde centrarse en las habilidades en estadística y visualización de datos, hasta capacidades y competencias más generales para analizar y resolver problemas con la ayuda de datos.

Por lo tanto, el estudio teórico y epistemológico de los datos es una constante que debe imperar en el desarrollo de teorías que ayuden a fundamentar el uso de dichos datos en los diversos contextos de la actividad humana, pues la manera de abordar críticamente los datos será a partir de su estudio como entidad epistemológica y con un trasfondo formalmente fundamentado.

El uso de datos abiertos enlazados en un contexto sistematizado promete revolucionar la producción de conocimiento dentro y más allá de la ciencia, al permitir formas novedosas y altamente eficientes de buscar, recuperar, difundir y evaluar los datos y la información que se requieren para el desarrollo de la investigación.

Las últimas décadas han sido testigos de la creación de formas novedosas de producir, almacenar y analizar datos, que culminó con la aparición del campo de la ciencia de datos, que reúne técnicas computacionales, algorítmicas, estadísticas y matemáticas para

extrapolar conocimientos a partir de del uso de grandes cantidades de datos.

En el contexto de los estudios de la información, será pertinente analizar los fenómenos derivados de la interacción de los datos con el usuario y la sociedad en general. Se trata de abordar los problemas relativos al uso y manejo de los datos, pero también a las maneras en que los individuos conviven con ellos, generando fenómenos que requieren el uso de metodologías para comprender en mayor medida su utilización en la respuesta a demandas informativas y en la toma de decisiones.

Finalmente, transformar el papel de los datos en un proceso para recuperar información, aumenta su estatus como elementos clave para el desarrollo de la investigación científica y académica. Los desarrollos tecnológicos y metodológicos que se han abordado en esta investigación tienen implicaciones que bien pueden ser encaminadas para el desarrollo de conceptualizaciones filosóficas acerca de los datos, los procesos inferenciales y el conocimiento científico, así como para la forma en que se conduce, organiza, gobierna y evalúa la investigación fundamentada en el descubrimiento intensivo en datos.

Recomendaciones

Los datos abiertos enlazados se utilizan como una propuesta aplicable al proceso de la recuperación de información. Por ende, se trata de un conjunto de principios que también reflejan una metodología para el tratamiento, manejo y procesamiento de los datos.

El modelo que se ha planteado en esta investigación se basa en el manejo de datos abiertos, pues actualmente existen muchos datos en el ambiente digital que quizás no puedan utilizarse debido a las restricciones técnicas, económicas y legales que los constituyen. Se recomienda trabajar con datos abiertos para fomentar la capacidad de los datos para enlazarse de manera interoperable en el ambiente digital.

El diseño de un sistema de información basado en los principios de los datos abiertos enlazados plantea la necesidad de integrar diversos componentes tecnológicos, sobre todo aquellos que se refieren a bases de datos enfocadas en grafos y mecanismos de visualización de datos. Estas áreas pueden ofrecer un nuevo umbral de oportunidades para los sistemas de información que se desarrollan en el área documental; sin embargo, es necesario formar recursos humanos con las habilidades y conocimientos necesarios que les permitan gestionar dichos sistemas.

El paradigma de la recuperación de información es ampliamente dinámico, lo cual ha provocado la aparición de

una amplia gama de metodologías, procedimientos y técnicas para obtener la información deseada por el usuario final. Sin embargo, bajo el contexto del uso de los datos abiertos enlazados, la recuperación de la información se verá ampliamente favorecida mediante el desarrollo de estudios de usuarios que permitan identificar los diversos comportamientos del usuario e identificar sus demandas informativas con la intención de trasladarlas al dominio del modelo que se ha planteado en este trabajo.

No debe pasar desapercibido que los datos abiertos enlazados se encuentran en una constante evolución; por ende, su aplicación en el contexto de la recuperación de la información traerá consigo una constante actualización en los estándares y las normas que permitan configurar la generación de los próximos sistemas de recuperación de información. Por ende, el profesional de la información también deberá mantenerse en constante actualización sobre dicha evolución.

Referencias bibliográficas

Amati, Giambattista. 2018a. "Information Retrieval". *Encyclopedia of Database Systems*, editado por Ling Liu y M. Tamer Özsu, 1970-75. Nueva York: Springer. https://doi.org/10.1007/978-1-4614-8265-9_915.

Baeza-Yates, Ricardo y Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison Wesley.

Berners-Lee, Tim; James Hendler y Ora Lassila. 2001. "The semantic web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities". *Scientific American,* vol. 284, núm. 5: 34-43.

Bizer, Christian; Maria-Esther Vidal y Hala Skaf-Molli. 2018. "Linked Open Data". *Encyclopedia of Database Systems*, editado por Ling Liu y M. Tamer Özsu, 2096-2101. Nueva York: Springer. https://doi.org/10.1007/978-1-4614-8265-9_80603 (Consultado el 04 de marzo de 2021).

Bizer, Christian; Tom Heath y Tim Berners-Lee. 2009. "Linked Data: the story so far." *International Journal on Semantic Web and Information Systems*, 5 núm. 3. http://doi:10.4018/jswis.2009081901 (Consultado el 27 de marzo de 2021).

Brath, Richard y David Jonker. *Graph analysis and visualization: discovering business opportunity in linked data*. Indianapolis, Indiana: Wiley, 2015.

Brush, Kate. "What is data visualization and why is it important?". *Search Business Analytics.* https://searchbusinessanalytics.techtarget.com/definition/data-visualization (Consultado el 02 de julio de 2021).

Cabrera, María, René Elizalde y Nelson Piedra. 2018. "Semantic representation of bibliographic resources cataloged in the UTPL Library". Trabajo presentado en la IEEE 37th Central America and Panama Convention. Concapan: IEEE, 1-6.

Calva González, Juan José. 2011. "Surgimiento de las necesidades de información". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 25, núm. 53. https://doi.org/10.22201/iibi.0187358xp.2011.53.27466 (Consultado el 22 de mayo de 2021).

___. 2004. *Las necesidades de información: fundamentos teóricos y métodos.* 1. ed. México: Universidad Nacional Autónoma de México, Centro Universitario de Investigaciones Bibliotecológicas.

Casalini, Michele. 2017. "BIBFRAME and Linked Data practices for the stewardship of research Knowledge". *DH. Opportunities and Risks. Connecting Libraries and Research.* Berlín. https://hal.archives-ouvertes.fr/hal-01636351.

Christophides, Vassilis. 2009. "Resource Description Framework (RDF) Schema (RDFS)". *Encyclopedia of Database Systems*, editado por Ling Liu y M. Tamer Özsu, 2425-28. Boston: Springer US. https://doi.org/10.1007/978-0-387-39940-9_1319.

Daquino, Marilena; Silvio Peroni, David Shotton, Giovanni Colavizza, Behnam Ghavimi, Anne Lauscher, Philipp Mayr, Matteo Romanello y Philipp Zumstein. 2020. "The OpenCitations Data Model". *The Semantic Web – ISWC 2020*, editado por Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne y Lalana Kagal, 12507:447-63. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-62466-8_28.

Davies, Tim y Ania Calderon. 2020. "Open Data". *The Palgrave Encyclopedia of Interest Groups, Lobbying and Public Affairs*, editado por Phil Harris, Alberto Bitonti, Craig S. Fleisher y Anne Skorkjær Binderkrantz, 1-8. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-13895-0_102-1.

Deliot, Corine. 2014. "Publishing the British National Bibliography as Linked Open Data". http://vll-minos.bl.uk/bibliographic/pdfs/publishing_bnb_as_lod.pdf. (Consultado el 21 de abril de 2021).

Dimou, Anastasia; Laurens De Vocht, Geert Van Grootel, Leen Van Campe, Jeroen Latour, Erik Mannens y Rik Van de Walle. 2014. "Visualizing the Information of a Linked Open Data Enabled Research Information System". *Procedia Computer Science* 33: 245-52. https://doi.org/10.1016/j.procs.2014.06.039.

Dominich, S. 2000. "A unified mathematical definition of classical information retrieval". *Journal of the American Society for Information Science*, núm 51, vol. 7: 614-624.

Fox, Peter y Hendler, James. 2014. "La e-ciencia semántica: la codificación del significado en la ciencia meorada digitalmente de la siguiente generación". *El cuarto paradigma: descubrimiento científico intensivo en datos*. Hey, Anthony J. G., Stewart Tansley, Kristin Michele Tolle, José Luis Acosta Rodríguez, Rodrigo Cambray-Núñez y Abdiel Macías Arvizu, eds. México: Universidad Autónoma Metropolitana.

Futia, Giuseppe; Antonio Vetro, Alessio Melandri y Juan Carlos de Martin. 2018. "Training Neural Language Models with SPARQL Queries for Semi-Automatic Semantic Mapping". *Procedia Computer Science* 137: 187-98. https://doi.org/10.1016/j.procs.2018.09.018.

Gandhi, Parul y Jyoti Pruthi. 2020. "Data Visualization Techniques: Traditional Data to Big Data". *Data Visualization: Trends and Challenges Toward Multidisciplinary Perception*, S. Margret Anouncia, Hardik A. Gohel, y Subbiah Vairamuthu, eds.: 53-74. Singapur: Springer Singapore. https://doi.org/10.1007/978-981-15-2282-6_4.

Gartner, Richard. 2016. *Metadata*. Cham: Springer International Publishing, https://doi.org/10.1007/978-3-319-40893-4.

Ginsparg, Paul. 2014. "El texto en un mundo centrado en datos". *El cuarto paradigma: descubrimiento científico intensivo en datos*. Hey, Anthony J. G., Stewart Tansley, Kristin Michele Tolle, José Luis Acosta Rodríguez, Rodrigo Cambray-Núñez, y Abdiel Macías Arvizu, eds. México: Universidad Autónoma Metropolitana.

Gómez, Laureano Felipe. 2007. "Interoperabilidad en los sistemas de información documental". *Revista Códice* vol. 3, núm. 1: 23-39.

Gottron, Thomas y Steffen Staab. 2018. "Linked Open Data". *Encyclopedia of Social Network Analysis and Mining*, eds.: Reda Alhajj y Jon Rokne, 1211-13. Nueva York: Springer, 2018. https://doi.org/10.1007/978-1-4939-7131-2_111.

Gunjal, Amit. 2016. "The Process of Information Retrieval". https://amitgunjal.wordpress.com/2016/11/21/the-process-of-information-retrieval-from-scratch/. (Consultado el 2 de septiembre de 2021).

Hansen, Charles, Chris Johnson, Valerio Pascucci y Claudio Silva. 2014. "Visualización para la ciencia intensiva en datos". *El cuarto paradigma: descubrimiento científico intensivo en datos.* Hey, Anthony J. G., Stewart Tansley, Kristin Michele Tolle, José Luis Acosta Rodríguez, Rodrigo Cambray-Núñez, y Abdiel Macías Arvizu, eds. México: Universidad Autónoma Metropolitana.

Haslhofer, Bernhard y Erich J. Neuhold. 2011. "A retrospective on semantics and interoperability Research". *Foundations for the Web of Information and Services: A Review of 20 Years of Semantic Web Research*. Dieter Fensel, ed.: Berlín, Heidelberg: Springer: 3-27. https://doi.org/10.1007/978-3-642-19797-0_1.

Hernández Salazar, Patricia. 2017. "El sentido de la información: un enfoque centrado en el usuario". *Significados e interpretaciones de la información desde el usuario. Hernández Salazar, Patricia*. México: Instituto de Investigaciones Bibliotecológicas y de la Información. https://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/L127/2/significados_informacion_usuario_k.pdf.

Hidalgo Delgado, Yusniel y Rafael Rodríguez Puente. 2013. "The semantic web: a brief overview". *Revista Cubana de Ciencias Informáticas* 7: 76-85.

Hua, Jiang. 2009. "Study on information retrieval model based on rough set theory". *International Symposium on Intelligent Ubiquitous Computing and Education*: 440-44. https://doi.org/10.1109/IUCE.2009.106.

Janssen, Marijn, Elsa Estevez y Tomasz Janowski. 2014. "Interoperability in big, open, and linked data organizational maturity, capabilities, and data portfolios". *Computer* 47, núm. 10: 44-49. https://doi.org/10.1109/MC.2014.290.

Kagolovsky, Yuri y Jochen Moehr. 2003. "Terminological Problems in Information Retrieval". *Journal of Medical Systems*, 10. https://doi.org/10.1023/a:1025687220609.

Khoii, Roya y Samira Sharififar. 2013. "Memorization versus semantic mapping in L2 vocabulary acquisition". *ELT Journal* 67: 199-209. https://doi.org/10.1093/elt/ccs101.

Klischewski, R. y H.J. Scholl. 2006. "Information quality as a common ground for key players in e-Government integration and interoperability". *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*: 72-72. https://doi.org/10.1109/HICSS.2006.221.

Kubernátová, Petra; Magda Friedjungová y Max van Duijn. 2019. "Constructing a Data Visualization Recommender System". *Data Management Technologies and Applications*, editado por Christoph Quix y Jorge Bernardino, 1-25. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-26636-3_1.

Lagoze, Carl. 2014. "Big data, data integrity, and the fracturing of the control zone". *Big Data & Society* 1. https://doi.org/10.1177/2053951714558281.

Lancaster, Frederick. 2001. "Sistemas avanzados de recuperación de información". *Procesamiento de la información científica:* 213-224, Madrid: Arco Libros.

Lancaster, F. W. y A. J. 1993. Warner. *Information Retrieval Today.* Arlington, Virginia: Information Resources Press.

Lapolla, Fred. 2013. "Perceptions of Librarians Regarding Semantic Web and Linked Data Technologies". *Journal of library metadata,* 13: 114-140.

Lau, Jesus. "Directrices sobre desarrollo de habilidades informativas para el aprendizaje permanente". https://www.ifla.org/wp-content/uploads/2019/05/assets/information-literacy/publications/ifla-guidelines-es.pdf (Consultado el 17 de abril de 2021).

Lazer, David; Ryan Kennedy, Gary King y Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis". *Science* 343: 1203-5. https://doi.org/10.1126/science.1248506.

Library Linked Data Incubator Group Final Report. https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/ (Consultado el 30 de agosto de 2021).

Liu, Danyang; Ting Bai, Jianxun Lian, Guangzhong Sun, Wayne Xin Zhao, Ji-Rong Wen y Xing Xie. 2019. "News Graph: An Enhanced Knowledge Graph for News Recommendation". *Proceedings of KaRS 2019 Second Workshop on Knowledge-Aware and Conversational Recommender Systems*, Beijing, China. Del 3 al 7 de noviembre. https://www.microsoft.com/en-us/research/uploads/prod/2019/09/kars2019.pdf.

López Herrera, Antonio Gabriel. Modelos de Sistemas de recuperación de información documental basados en información lingüística difusa. Tesis para optar por el grado de Doctor en Informática. Escuela Técnica Superior de Ingeniería Informática Departamento de Ciencias de la Computación e Inteligencia Artificia. Universidad de Granada. 2006. https://www.kimerius.es/app/download/5788695465/Modelos+de+sistemas+de+recuperaci%C3%B3n+de+informaci%C3%B3n+documental+basados+en+informaci%C3%B3n+ling%C3%BC%C3%ADstica+difusa.pdf.

Martínez Comeche, Juan Antonio. 2006. "Los modelos clásicos de recuperación de información y su vigencia". Martínez Arellano, Filiberto Felipe y Calva González, Juan José. *Memoria del Tercer Seminario Hispano-Mexicano de investigación en bibliotecología y documentación. Tendencias de la investigación en bibliotecología y documentación en México y España*. 29 al 31 de marzo de 2006: 187-206. http://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/CL952/1/3er_seminario_hispanomexicano_juan_antonio_martinez_comeche.pdf.

Martínez Méndez, Francisco Javier. 2004. *Recuperación de información: modelos, sistemas y evaluación*. Murcia: Kiosko.

Meadow, Charles, Donald Kraft y Bert Boyce. 1999. *Text Information Retrieval Systems*. San Diego: Academic Press.

Méndez, Eva. 1999. "RDF: un modelo de metadatos flexible para las bibliotecas digitales del próximo milenio". *7es Jornades Catalanes de Documentació*, Barcelona (Spain), 4 al 6 de noviembre, 1999.

Méndez, Eva y Jane Greenberg. 2012. "Datos enlazados para vocabularios abiertos y marco general de HIVE". *El profesional de la información* 21.

Mirel, Barbara. 1999. "Complex Queries in Information Visualizations: Distributing Instruction Across Documentation and Interfaces". *Proceedings of the 17th annual international conference on Computer documentation*: 1-8. https://doi.org/10.1145/318372.318378

Morato, Jorge; Sonia Sanchez Cuadrado, Alejandro Ruiz Robles y José Antonio Moreiro González. 2014. "Information visualization and retrieval in the semantic web". *Profesional de la información* 23: 319-29. https://doi.org/10.3145/epi.2014.may.12.

Morsey, Mohamed; Jens Lehmann, Sören Auer, Claus Stadler y Sebastian Hellmann. 2012. "DBpedia and the live extraction of structured data from Wikipedia". *Program* 46: 157-81. https://doi.org/10.1108/00330331211221828.

Ontotext. "What is a Knowledge Graph?: Ontotext Fundamentals". (Consultado el 02 de junio de 2021). https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/.

Open Citations. "Welcome to the OpenCitations homepage". (Consultado el 21 de abril de 2021). https://opencitations.net/.

Open Data Handbook. "¿Qué son los datos abiertos?". (Consultado el 30 de mayo de 2021). https://opendatahandbook.org/guide/es/what-is-open-data/.

Peroni, Silvio; Francesca Tomasi y Fabio Vitali. 2013. "Reflecting on the Europeana Data Model". *Digital Libraries and Archives*, Maristella Agosti, Floriana Esposito, Stefano Ferilli, y Nicola Ferro, eds.: 228-40. *Communications in Computer and Information Science*. Berlín, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-35834-0_23.

Peters, Carol, Martin Braschler y Paul Clough. 2013. "Interaction and User Interfaces". *Multilingual Information Retrieval: From Research To Practice*, Carol Peters, Martin Braschler y Paul Clough, eds.: 85-128. Berlín: Springer. https://doi.org/10.1007/978-3-642-23008-0_4.

Pinto, María. 2018. "Búsqueda y Recuperación de Información". (Consultado el 24 de mayo de 2021). http://www.mariapinto.es/e-coms/busqueda-y-recuperacion-de-informacion/.

Polleres, Axel. 2014. "SPARQL". *Encyclopedia of Social Network Analysis and Mining*, Reda Alhajj y Jon Rokne eds.: 1960-66. Nueva York: Springer. https://doi.org/10.1007/978-1-4614-6170-8_124.

Quimbert, Erwann; Keith Jeffery, Claudia Martens, Paul Martin y Zhiming Zhao. 2020. "Data Cataloguing". *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges*, Zhiming Zhao y Margareta Hellström eds.: 140-61. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-52829-4_8.

Ramada, Mariana Soller, João Carlos da Silva y Plínio de Sá Leitão-Júnior. 2020. "From Keywords to Relational Database Content: A Semantic Mapping Method". *Information Systems* 88: 101460. https://doi.org/10.1016/j.is.2019.101460.

Rodríguez-Cruz, Yunier y María Pinto. 2018. "Modelo de uso de información para la toma de decisiones estratégicas en organizaciones de información". *Transinformação* 30: 51-64. https://doi.org/10.1590/2318-08892018000100005.

Romá Ferri, María Teresa. 2015. "Los sistemas de recuperación de la información de las bases de datos documentales y la calidad de los resultados obtenidos". *Documentación e información científica en salud*, núm. 43401. https://rua.ua.es/dspace/bitstream/10045/42141/1/DICS_T4_SRI_EvaluaResultados_C14_15.pdf.

Russo, Vanessa. 2015. "Semantic Web: Metadata, Linked Data, Open Data". *Science & Philosophy* 3, núm. 2: 37-46. http://eiris.it/ojs/index.php/scienceandphilosophy/article/view/266.

Sammut, Claude y Geoffrey Webb. 2017. "Information Retrieval". *Encyclopedia of Machine Learning and Data Mining*, 671-72. Boston: Springer. https://doi.org/10.1007/978-1-4899-7687-1_403.

Sánchez Luna, Blanca Estela. 2012. "Lenguajes documentales". *Anuario de Bibliotecología*. 1.1: 61-77. http://ru.ffyl.unam.mx//handle/10391/4267.

Sánchez, Pastor. "Simple Knowledge Organization System". Wikipedia, la enciclopedia libre, (consultado el 11 de julio de 2020). https://es.wikipedia.org/w/index.php?title=Simple_Knowledge_Organization_System&oldid=127641215.

Senso Ruiz, José. "*Resource Description Framework* «Resource Description Framework". (Consultado el 30 de agosto de 2021). https://www.upf.edu/hipertextnet/numero-1/rdf.html.

SHARE Virtual Discovery Environment. (Consultado el 22 de abril de 2021). https://share-vde.org/sharevde/clusters?l=en.

Skilton, Mark y Felix Hovsepian. 2018. *The 4th Industrial Revolution: Responding to the Impact of Artificial Intelligence on Business*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-62479-2.

Sparck Jones, Karen. 1999. "Information Retrieval and Artificial Intelligence". *Artificial Intelligence* 114: 257-81. https://doi.org/10.1016/S0004-3702(99)00075-2.

Speicher, Steve; Arwe, John y Malhortra Ashok. 2015. "Linked Data Platform 1.0". (Consultado el 14 de junio de 2021). https://www.w3.org/TR/ldp/.

SPUK. 2015. "¿Qué son los "Datos Abiertos?". (Consultado el 1 de junio de 2021). https://ceweb.br/guias/dados-abertos/es/capitulo-2/.

Taylor, Arlene. 1999. *The organization of information*. Estados Unidos de América: Libraries Unlimited.

Tello, Adolfo. 2001. "Ontologías en la Web Semántica". (Consultado el 15 de mayo de 2021). http://eolo.cps.unizar.es/docencia/MasterUPV/Articulos/Ontologias%20en%20la%20Web%20Semantica.pdf.

Tillet, Barbara. 1987. Bibliographic relationships: toward a conceptual structure of bibliographic information used in cataloging. Tesis de doctorado. Universidad de California, Los Ángeles.

Turnbull Muñoz, Federico. 2005. *Industria de la información en México: estado actual y prospectiva. Foro Transfronterizo de Bibliotecas*. Chihuahua, México. http://eprints.rclis.org/9487/1/2006.F.TurnbullMunoz.ForoTransfronterizo.pdf.

Ungvarsky, Janine. 2020. "Information retrieval". *Salem Press Encyclopedia*. Salem Press.http://pbidi.unam.mx:8080/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=125600246&lang=es&site=eds-live.

Vallez, Mari y Rafael Pedraza. 2007. "El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines" (Consultado el 30 de abril de 2021). http://www.upf.edu/hipertextnet/numero-5/pln.html.

Vandenbussche, Pierre-Yves; Ghislain A. Atemezing, María Poveda-Villalón y Bernard Vatant. 2016. "Linked Open Vocabularies (lov): A Gateway to Reusable Semantic Vocabularies on the Web". *Semantic Web*. Michel Dumontier, ed. 8: 437-52. https://doi.org/10.3233/SW-160213.

Viltres Sala, Hubert; Paúl Rodríguez Leyva, Juan Pedro Febles y Vivian Estrada Sentí. 2018. "Procesamiento Semántico de información en Sistemas de Recuperación de Información". *Revista Cubana de Ciencias Informáticas* 12: 102-106.

W3C. 2004. "Vista General del Lenguaje de Ontologías Web". (Consultado el 30 de agosto de 2021). https://www.w3.org/2007/09/OWL-Overview-es.html.

___. 2013. "SPARQL 1.1 Overview" (Consultado el 5 de abril de 2021). https://www.w3.org/TR/sparql11-overview/.

Wang, Yong. 2013. "Ontology". *Encyclopedia of Systems Biology*, Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho y Hiroki Yokota, eds. 1560-61. Nueva York: Springer. https://doi.org/10.1007/978-1-4419-9863-7_488.

Wikidata. 2021. "Introducción a Wikidata". (Consultado el 30 de agosto de 2021). https://www.wikidata.org/wiki/Wikidata:Introduction/es.

Wood, David; Marsha Zaidman, Luke Ruth y Michael Hausenblas. 2014. *Linked Data: Structured Data on the Web*. Shelter Island, Nueva York: Manning.

Zhang, Jin. 2008. "Information Retrieval and Visualization". *Visualization for Information Retrieval*, Jin Zhang, ed. 1-20. The Information Retrieval Series. Berlín, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-75148-9_1.

Recuperación de información con datos abiertos enlazados. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial, Anabel Olivares Chávez; revisión especializada, Valeria Guzmán González; corrección de pruebas, Carlos Ceballos Sosa; revisión de pruebas, Valeria Guzmán González; formación editorial, Sonia Wendy Chávez Nolasco. Fue impreso en papel cultural de 90 g en los talleres de MIGAL Impresiones Digitales, 3er. Anillo de Circunvalación no. 73, Col. Barrio Santa Bárbara, Alcaldía Iztapalapa, C.P. 09000, Ciudad de México. Se terminó de imprimir en agosto de 2022.