Artículo original

ANCORP: Procedimiento para la recuperación de información en sistemas de bibliotecas mediante grafos de conocimiento

ANCORP: Procedure for information retrieval in library systems using knowledge graphs

José A. Senso Ruiz[1*] https://orcid.org/0000-0002-6553-6522
Amed Abel Leiva Mederos[2] https://orcid.org/0000-0002-9144-5018
Yorbelis Rosell León[3] https://orcid.org/0000-0003-2230-1253
Ania Rosa Hernández Quintana[3] https://orcid.org/0000-0002-1484-8892

[1]Universidad de Granada. España.
[2]Universidad Central de Las Villas "Marta Abreu". Cuba.
[3]Universidad de La Habana. Cuba.

*Autor para la correspondencia: jsenso@ugr.es

RESUMEN

Las bibliotecas y los centros de documentación carecen de una guía metodológica para transformar sus datos RDF en grafos de conocimiento, lo que impide que puedan aprovechar las facilidades de esta herramienta en la búsqueda y recuperación de información. El artículo propone una metodología para la transformación de datos bibliográficos en grafos de conocimiento. Se presenta ANCORP, a partir del análisis de las técnicas de incrustación, limpieza y chequeo de grafos de conocimiento. Esta metodología se divide en dos partes: la parte 1, dedicada a la construcción del grafo de conocimiento, y la parte 2, dedicada a resolver los procesos de recuperación de información. Con la implementación de la metodología se corroboran saltos cualitativos en la recuperación de información y en la calidad de los datos.

Palabras clave: Recuperación de información; biblioteca; Linked Data.

ABSTRACT

Libraries and documentation centers haven't a methodology guide to transform their RDF data into knowledge graphs, which prevents them from taking advantage of the facilities of this tool in the search and retrieval of information. This methodology was proposed for the transformation of bibliographic data in knowledge graphs. ANCORP was presented from the analysis of the techniques of incrustation, cleaning and checking of knowledge graphs. This methodology was divided into two parts: part I dedicated to the construction of the knowledge graph, and part II dedicated to solving the processes of information retrieval. With the implementation of the methodology, qualitative leaps in the information retrieval and in the quality of the data are corroborated.

Key words: Information retrieval; library; Linked Data.

Recibido: 16/12/2020

Aceptado: 10/03/2021

Introducción

Los datos enlazados han facilitado que la recuperación de información ofrezca mayor valor añadido a los procesos de búsqueda. Los grafos distribuidos constituyen una de las propiedades desarrolladas desde la web semántica que han propiciado la aparición de los sistemas de inferencia semántica.[1] Estos sistemas son capaces de facilitar:

- adquisición del conocimiento en la recuperación de la información;
- escalabilidad de los datos semánticos de los sistemas y su interacción con otros;
- construcción de modelos y algoritmos ad-hoc para la modelación de sistemas de recuperación de información basados en grafos de conocimiento.

Las técnicas de recuperación de información en el ámbito de la web semántica han ido evolucionando a zonas cada vez más complejas donde se mezclan las necesidades de los usuarios con las facilidades de procesamiento de los ordenadores. El uso de grafos de conocimiento ha enriquecido la forma en que se recupera información, lo que ha propiciado que las aplicaciones acogidas al paradigma de la web semántica dispongan de mecanismos muy sofisticados para localizar datos asociados a diferentes instancias.[2,3] Estos mecanismos de recuperación de información emanados del perfeccionamiento de la filosofía de *Linked Open Data* mejoran la precisión y la relevancia en la recuperación de la información adelantándose a las necesidades de los usuarios.

Los grafos de conocimiento modelan información objetiva en forma de entidades a partir del RDF. En los últimos años, la web ha evolucionado de una red de documentos vinculados a una estructura donde tanto documentos como datos están aunados y generan lo que se conoce como Web de los Datos.

El crecimiento del *Linked Data* ha convertido a la web en un espacio de intercambio dinámico de datos. *Linked Data* ha aparecido en dominios diversos como: negocios y finanzas, geografía, gobierno, medios de comunicación, bibliotecas digitales, ciencias de la vida, además de conjuntos de datos generados por usuarios. En el terreno de la documentación los usos de esta tecnología han servido para la gestión de tesauros mediante técnicas de *big data*[4] y en la generación de consultas de sistemas de repositorios.[5] Los grafos de conocimiento abren una puerta nueva a nuevas formas de recuperación de información en las entidades de información. Tradicionalmente, estas herramientas han servido para la búsqueda de información sobre datos almacenados en bases de datos locales o enlaces a diversos recursos de información.

Los cambios en los formatos bibliográficos, en los procesos de catalogación y la adopción de *bibframe* en los formatos bibliográficos abren paso a la concepción de grafos de conocimiento orientados a la gestión de información en las Bibliotecas. Sin embargo, los sistemas de gestión de la bibliotecaria aún no

disponen de mecanismos para evolucionar hacia este tipo de tecnologías de recuperación de la información, lo que acarrea problemas en la actividad de las bibliotecas.

En los últimos años se han obtenido notables resultados en la adopción de los principios de los datos enlazados, para beneficiar con la interoperabilidad semántica las funciones de las bibliotecas. Sin embargo, aún persisten problemas que deben ser resueltos:

- Insuficiente calidad de los datos enlazados publicados.
- Carencia de metodologías y guías metodológicas establecidas para la transformación de metadatos bibliográficos a *bibframe*.
- Insuficiente utilización de los metadatos bibliográficos publicados como datos enlazados en las prestaciones a los usuarios.
- Sistemas de recuperación ineficientes para el entorno de los gráficos de conocimiento.

Teniendo en cuenta las problemáticas mencionadas, el objetivo de esta investigación fue elaborar una metodología para la transformación de datos bibliográficos en grafos de conocimiento.

Métodos

Para el desarrollo de la investigación se utilizaron varios métodos que provienen de la ingeniería informática, particularmente de la web semántica. Hemos utilizado técnicas computacionales diversas; entre ellas:

- *Transformación de datos:* supone el uso de algoritmos que permiten convertir los datos de un formato a otro para que los datos sean escalables.

- *Limpieza de datos:* algoritmos que permiten localizar duplicados mediante la comparación de registros bibliográficos. Esta técnica se complementa con el uso de la similitud coseno.
- *Construcción de ontologías:* para construir y enriquecer la ontología de manejo se utilizó la metodología NEON[6].
- *Estudios de usuarios:* observar los problemas típicos en la recuperación de la información.
- *Incrustación de grafos:* para buscar relaciones dinámicas entre grafos.

Se han utilizado diversas herramientas, como *Protégé* (para la construcción de la ontología), *Virtuoso* (almacén de tripletas), *R2rml* (facilita el mapeo de los datos), *graphSPARQL* (permite las consultas semánticas) y *Lodlive* (usado para visualizar el grafo).

Procedimientos para la construcción de grafos de conocimiento y su relación con los sistemas de gestión bibliotecaria

No abundan metodologías para el desarrollo de grafos de conocimiento; por eso en este trabajo solo colocaremos algunas técnicas que sirven a la solución de los problemas que se presentan en las bibliotecas.

Las investigaciones relacionadas con los grafos de conocimiento en el terreno de la documentación no han sido utilizadas en su totalidad en la actividad de las bibliotecas. En este artículo revisaremos las técnicas de grafos de conocimiento que pueden servir para resolver problemas de eficiencia en los sistemas de gestión de las bibliotecas.

La construcción de tesauros de manera automática ha sido uno de los problemas en los sistemas de gestión de biblioteca debido a la desactualización de estos y a la necesidad de construir de forma automática sistemas lingüísticos soportados en RDF. Los procederes más eficientes para dar solución a este reto son la integración de grafos, la gestión de bases de datos distribuidas y las técnicas de lingüística computacional.

Dentro de la integración de grafos de conocimiento se perfila el trabajo de *Lesnikova*,[7] que se centra en la solución de problemas de grandes sistemas de información con diversos tesauros. La solución computacional que proponen integra varios recursos lingüísticos con estructura RDF mediante grafos de conocimiento distribuidos. En este mismo tema, en el terreno de la lingüística computacional, *Shekarpour*[8] propone un nuevo método para la reescritura automática de las consultas que sirven de entrada en el modelo de *Markov*,[a] lo que permite determinar las palabras derivadas más adecuadas de los recursos lingüísticos.

En este trabajo se utiliza el concepto de co-ocurrencia para reconocer palabras que co-ocurren en las estructuras RDF. Dentro de la lingüística computacional, también se ha revisado el trabajo de *Qian*,[9] en el que se propone una base de datos semántica de palabras en chino en la cual los grafos de conocimiento son utilizados para establecer relaciones entre los términos hipernonimia, homonimia y sinonimia.

En *Jana* y *Goyan*[10] es presentada la integración de grafos mediante diversas técnicas (*DeepWalk, LINE, node2vec*, etc.) para mejorar el completamiento de los tesauros, lo que ha convertido en una red de tesauros distribuidos con vectores de palabras.

Nie y *Sun*[11] proponen un modelo que mezcla las propiedades de la integración de grafos para realizar inferencias sobre entidades, relaciones y texto. El modelo no solo es adecuado para modelar las interacciones textuales, sino que también se utiliza en la modelación de relaciones entre las entidades en el grafo. *Silva*[12] propone un algoritmo para la vinculación de las partes del texto basado en modelos semánticos distributivos que sirven para encontrar un camino en el grafo que vincule el texto y principal y sus posibles relaciones.

Un enfoque muy bueno para generar grafos de conocimiento ha sido creado por *Martínez-Rodríguez*.[13] A partir de las relaciones binarias producidas por un

enfoque *OpenIE*, los autores presentan estrategias para favorecer la extracción y la vinculación de entidades nombradas. Para esto proponen el uso de la asociación de unidades gramaticales que facilitan la coherencia semántica.

Con el fin de mejorar el rendimiento de los diferentes métodos de desambiguación que usan como base la similitud de contexto, se propone un procedimiento denominado *SCSNED,*[14] que suministra la desambiguación entre las palabras contextuales y las informativas que aparecen en las entidades. Además, proponen una forma de incrustación con *Category2Vec*, un modelo basado en aprendizaje que opera a partir del conjunto de incrustaciones de las palabras y las categorías gramaticales.

Otra de las problemáticas asociadas hoy a los sistemas de gestión de biblioteca son las bajas capacidades de sus mecanismos de recuperación de información cuando estos tienen que garantizar la búsqueda textual y la respuesta a consultas. Los algoritmos y modelos de incrustación se han convertido en técnicas potencialmente eficientes para la solución a estos problemas.

Con esta técnica se han desarrollado modelos y métodos que permiten razonar nuevos hechos y relaciones a gran escala. *Wang*[15,16]expone un método que permite la inserción conjunta de entidades y palabras en el mismo espacio vectorial. Esta forma de incrustación permite preservar las relaciones entre las entidades del grafo de conocimiento y la frecuencia de palabras que más se repiten en el texto. Los nombres de las entidades se utilizan para alinear las incrustaciones.

Guo[17] considera un problema incrustar *Knowledge Graphs* (KG) en datos con múltiples instancias. La mayoría de los métodos existentes realizan esta tarea basándose únicamente en hechos observados. Con *Semantically Smooth Embedding* (SSE) se aprovecha al máximo la información semántica adicional generada en el grafo y se logra que el espacio de incrustación sea semánticamente fluido. Para lograr este tipo de incrustación se utilizan dos algoritmos de aprendizaje múltiples: *Laplacianos Eigenmaps* y *Locally Linear*

Embedding. Ambos se formulan como términos de regularización geométricos para restringir la tarea de incrustación.

Un gráfico de conocimiento en un sistema de gestión de biblioteca tiene que resolver los problemas de enlazar entidades aun cuando el sistema use FRBR, formato que no contiene toda la variedad de relaciones que pueden asociarse a los textos, ya que este es incapaz de predecir los diversos enlaces que pueden generarse en la búsqueda textual.

El enfoque de incrustaciones de grafos de conocimiento utilizado en los modelos *TransE* y *TransH* es una solución poco eficiente a este problema, pues construye las relaciones y las considera como un puente para la traducción entre la entidad principal y la entidad de cola. En estos modelos coexisten las entidades y relaciones dentro del mismo espacio semántico; por eso se propone *TransR*[18] para construir incrustaciones de entidades y relaciones en espacios de entidades separadas y espacios de relaciones.

Fan[1] presenta un modelo para predecir las relaciones que se establecen entre las entidades que componen un grafo de conocimiento de alta dimensión. El modelo toma como base el aprendizaje probabilístico para lograr que el grafo se expanda sobre las posibles entidades asociadas al contexto, y maximiza así la probabilidad de registrar el conocimiento observado.

Para predecir las relaciones en los grafos existen técnicas de incrustación que facilitan que las entidades y sus relaciones sean representadas con vectores de baja dimensión. Una de las funciones para lograr la incrustación de grafos se denomina *PaSKoGE*,[19] la cual determina adaptativamente para cada camino una función basada en la codificación de la correlación entre las relaciones y las rutas de relación. Elimina las actividades necesarias para la gestión de la información que obliga al grafo el recorrido de vínculos, relaciones y rutas de relación.

Dentro de los mecanismos de inferencia de grafo se expone el enfoque de *Chen,*[20] que formula un enfoque de incrustación para razonar nuevos hechos y relacionales a partir de un grafo de conocimiento a gran escala y un *corpus* de texto. El método integra entidades y palabras en forma conjunta en el mismo espacio vectorial continuo. El proceso de integración intenta preservar las relaciones entre las entidades en el grafo de conocimiento y las concurrencias de las palabras en el cuerpo del texto.

Aunque los sistemas de gestión de biblioteca se han encaminado hacia *bibframe*, se ha observado que existen datos y entidades nombradas que necesitan limpieza para evitar ruido en el proceso de búsqueda. Uno de los algoritmos más utilizados en la limpieza de los grafos es *Crumb Trail*.[2] Este algoritmo elimina los ciclos, los nodos fuera del dominio y los nodos no esenciales de forma segura sin romper la conectividad del grafo. En *Crumb Trail* se utiliza la poda topológica *top-botton* sobre la base de un conjunto de conceptos de entrada. Esta técnica se aplica a grafos de hipernonimia ruidosos, generados típicamente por algoritmos de aprendizaje de ontologías. *CrumbTrail Over* tiene las limitaciones de complejidad de tiempo y espacio de los algoritmos de última generación.

Otra dimensión de las problemáticas los sistemas de gestión de bibliotecas es que la recuperación de información se realiza sobre bases de datos estructuradas con baja eficiencia en la recuperación de la información, con bajo nivel de inferencia e interoperabilidad.

Las entidades en estas bases de datos actuales están asociadas a múltiples elementos que sitúan a los buscadores en contexto y facilitan las búsqueda y recuperación de información; sin embargo, esto no ocurre totalmente porque las consultas deben vincular las menciones de las entidades en los textos, lo que obliga a desambiguarlas, ya que la información se dispersa y se torna ruidosa. En *Tonon*[21] se proponen y evalúan nuevos métodos para encontrar el tipo de entidad más relevante según las estadísticas de recopilación y en la

estructura del grafo de conocimiento, que interconectan entidades y tipos, lo que facilita la recuperación ante el usuario final.

En los trabajos de *Dou*[22] se ha solucionado el problema de dispersión de información en bases de datos sobre la temática patrimonio cultural chino, mediante un grafo del conocimiento. La construcción de este grafo demandó la gestión de una ontología de dominio en cuya estructuración participaron expertos en patrimonio cultural inmaterial chino e ingenieros de conocimiento. En la misma temática aparece el proyecto de cultura francés llamado *DOREMUS*.[23] Tres importantes instituciones culturales francesas: la Biblioteca Nacional de Francia (BnF), Radio France y la Philharmonie de Paris se han unido para desarrollar métodos compartidos para describir semánticamente sus catálogos de obras y eventos musicales. Este proceso comprende la construcción de grafos de conocimiento que representan los datos contenidos en estos catálogos mediante una ontología que amplía *CIDOC-CRM* y *FRBRoo*, la cual facilita la vinculación de estos grafos y su publicación abierta en la web.

En la misma temática aparece la investigación de *Boer*,[24] donde se presenta una metodología para publicar, representar y enriquecer colecciones patrimoniales. Los grafos de conocimiento sirven para relacionar objetos, personas y lugares con recursos audiovisuales y eventos históricos.

Shan[25] desarrolla un modelo probabilístico que utiliza los grafos de conocimiento para inferir los objetivos de búsqueda. Su investigación se centra en el análisis de los especificadores de retorno, los modificadores, las relaciones y la ganancia de información en las palabras clave utilizadas en consulta. Esto resuelve los problemas de consulta de triple-patrón en grafos de conocimiento. Investigaciones sobre la eficiencia de la recuperación del grafo han sido desarrolladas por *Arnaout* y *Elbassuoni*,[26] quienes proponen un marco general para la búsqueda efectiva de los gráficos de conocimientos y reconocen patrones con una amplia gama de consultas con palabras clave, que proporcionan, además, una clasificación de resultados basada en la estadística para mejorar la recuperación de la información. También, con este proceder se

obtiene una diversidad de resultados en la configuración de datos RDF y se proponen mecanismos para diversificar los resultados de búsqueda utilizando la Relevancia Marginal Máxima.

El gráfico de conocimiento de gran escala contiene una serie de características semánticas basadas en el número de rutas que genera, lo que proporciona un mecanismo flexible para asignar y expandir la semántica, los atributos a las entidades y la búsqueda entre los catálogos de las entidades de información. *Chen*[20] usa la expansión de conjuntos de entidades como un ejemplo para mostrar que las características semánticas basadas en rutas se pueden utilizar de manera efectiva en una aplicación de búsqueda semántica. Proponen modelos probabilísticos para clasificar las entidades y con esto facilitan la recuperación de información y las consultas.

Se han revisado otras técnicas relativas al *bigdata* denominadas QSTR (Razonamiento Temporal Espacial) que abordan el razonamiento sobre los conjuntos de datos espaciales y temporales cualitativos a gran escala. La propuesta de *Mantle*[27] es ParQR, una aplicación que utiliza el marco de *Apache Spark* para manejar de forma distribuida las redes de restricción con millones de relaciones.

Propuesta para la aplicación de la metodología ANCORP

La metodología ANCORP (Anotación Coordinada de Registros Públicos) tiene por objetivo la transformación de datos bibliográficos en grafos, a partir del análisis de las técnicas de incrustación, limpieza y chequeo de grafos de conocimiento. Esta se divide en dos fases: fase 1, dedicada a la construcción del grafo de conocimiento y la fase 2, dedicada a resolver los procesos de recuperación de información (Fig. 1).

Fig. 1 - Componentes de ANCORP.

Fase 1: Construcción del grafo de conocimiento.

La etapa de construcción de datos es un paso muy complejo que debe permitir a la entidad de información o biblioteca la transformación óptima de los datos que han de servir de punto de partida a la construcción del grafo de conocimiento.

1.1. Extracción de datos.

La extracción de datos consiste en la selección de los formatos de datos para que el conocimiento almacenado en la biblioteca pueda ser procesado. Este paso obliga a seleccionar los formatos de las bases de datos bibliográficas; en este caso, pueden estar en formato MARC con registros asociados en PDF. La idea de este paso es generar un fichero en formato JSON.

1.2. Transformación de formatos de datos.

La limpieza de los datos obliga a aplicar técnicas para convertir los formatos de documentos a tipologías documentales necesarias para que el sistema de gestión de biblioteca pueda alojar sus datos en formatos intercambiables. El fichero JSON permite obtener ficheros con extensión ".xml", de manera que puedan ser leídos en su totalidad por la herramienta *open-refine*.[28,29]

Esta fase de la metodología implica la realización de varias acciones:

1.3. Reutilización de los esquemas ontológicos (RDF)

- Localización de vocabularios ontológicos.
- Transformación a los formatos requeridos.
- Enriquecimiento del formato con la colocación de marcadores de posición para las URI (Identificador Uniforme de Recursos, de sus siglas en inglés: Uniform Resource Identifier).
- Control de autoridades.

Si bien existen más de 650 vocabularios ontológicos en el caso de los sistemas de gestión de biblioteca, será pertinente transformar los formatos de Marc 21 a *Bibframe*.[30,31] (Fig. 2).

Fig. 2 - Estructura de Bibframe.

Para transformar un formato de *Marc 21* a *Bibframe* es necesario valerse de las herramientas *MARCXML* que transforman el formato *MARC 21* a *BIBFRAME*,

disponibles en la página del repositorio de *software GitHub* de la biblioteca del congreso. El formato MARC 21 por sí solo no genera un RDF rico en relaciones, por lo que es vital su enriquecimiento y la colocación de marcadores de posición para las URI. Por eso se revisaron experiencias en instanciación y anotación en otras bibliotecas y se diseña el modelo de trabajo que se observa en la figura 2. Este modelo refleja las relaciones entre trabajos, instancias y anotaciones.

La instancia de BIBFRAME se define como un recurso que refleja una relación con un material identificado unívocamente. Es importante escoger una URI para identificar los catálogos de las entidades. Nosotros seleccionamos las propiedades bf: hasInstance y bf: instanceOf. Una obra puede tener muchas instancias y muchas instancias pueden apuntar a una obra. También es vital realizar el Modelado de anotación usando la propiedad "Annotation: about", la cual incluye un enlace a un sitio donde podemos acceder al recurso electrónico descrito en los datos de BIBFRAME.[31](Fig. 3).

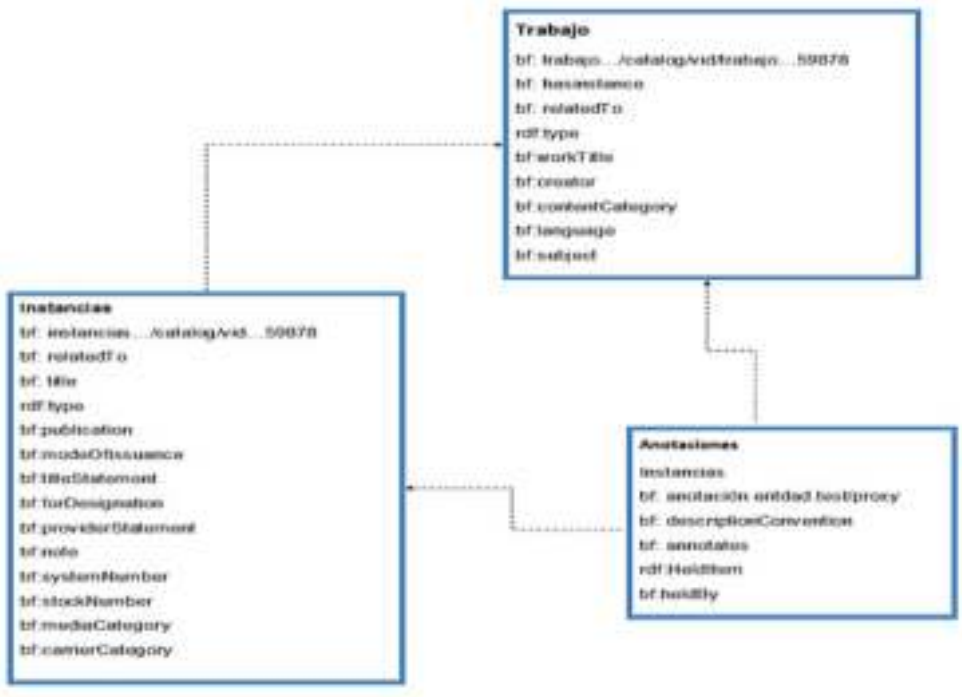


Fig. 3 - Modelo de Transformación de Bibframe.

El control de autoridades en Bibframe y la forma de instanciación de los datos es vital para el desarrollo de la migración de datos. Las reglas de trabajo con autoridad se establecen de la siguiente forma:

- Si el investigador o autor está en VIAF se conecta a la URI que los representa.
- Cuando los nombres de los autores no pueden encontrarse en VIAF se vinculan a *WorldCat Identities*. *WorldCat* tiene más de treinta millones de nombres, incluidos las personas, organizaciones y personajes ficticios.
- Para el manejo de materias y ante la ausencia de herramientas de corte lingüístico podemos vincularnos a id.loc.gov. agrovoc, al tesauro unesco, y al mesh. Esta base de datos proporciona URI para nuestros registros bibliográficos, tanto en formato duro como electrónicos, además de servir de vínculo entre los registros de autoridad.

1.4. Limpieza de datos.

Las técnicas de limpieza de datos permitirán chequear duplicados. Según *Christen*,[32] cuando el proceso de búsqueda se realiza en una sola base de conocimiento se conoce como detección de instancias duplicadas. En este caso cada instancia necesita compararse con las demás para determinar si representan la misma entidad o no. El total de comparaciones potenciales entre pares de registros es $|A|x(|A|-1)/2$, donde $|A|$ es el número de registros de la base de datos.[33] Por el contrario, cuando el proceso de búsqueda se ejecuta en dos o más ontologías o ficheros RDF se conoce como vinculación de registros u ontologías.[34] En este proceso todas las instancias de una ontología se deben comparar con los registros de la otra.[33] Para el caso de dos ontologías *A* y *B*, cada registro de *A* debe ser comparado con todos los registros de *B*. El número de comparaciones de pares de registros es $|A|x|B|$, donde $|A|$ y $|B|$ representan la cantidad de registros de *A* y *B* respectivamente.[33] En este trabajo se utiliza el término "detección de duplicados" para referirse a los dos casos.

El principal problema de rendimiento en la detección de duplicados es la costosa comparación detallada entre los valores de cada campo (o atributo) de los registros.[33,34,35] La problemática anterior adquiere mayor relevancia en la medida que el volumen de los datos aumenta, lo que hace casi imposible la comparación de todos los pares de registros.[32,33] Además, si se asume que no hay registros duplicados en una misma base de datos (un registro de *A* solo puede enlazarse con uno de *B* y viceversa), el número máximo de duplicados es el mínimo (*|A|, |B|)*. De esta manera, al trabajar con grandes bases de datos bibliográficas, la complejidad computacional aumenta cuadráticamente, mientras que el número de duplicados crece de manera lineal.[32,33,36] Teniendo *d* bases de datos de *n* registros, la complejidad computacional utilizando fuerza bruta es $O(n^d)$. Lo anterior también se aplica a una sola base de datos; en este caso, el número máximo de duplicados es siempre menor que el total de registros de esta.

En general, el proceso de detección de duplicados consta de las cinco etapas siguientes: pre-procesamiento de datos, indexado o bloqueo, comparación de pares de registros, clasificación y evaluación.[35,37](Fig. 4).

Fig. 4 - Procedimiento general de la detección de duplicados.

La limpieza de nombres es el paso más complejo de esta subetapa y consiste en eliminar los nombres repetidos, de manera que el grafo funcione correctamente y sin ruidos. Para determinar los nombres repetidos utilizamos una métrica de similitud denominada coseno. Este coeficiente es ampliamente utilizado al determinar la similitud entre documentos y se basa en el coseno del

ángulo que existe entre ellos. La ecuación que se muestra a continuación presenta la distancia de similitud coseno:

$$S(N_{i,}\ N_j) = \frac{\sum_{h=1}^{k}(peso_{ih}\ .\ peso_{jh})}{\sqrt{\sum_{h=1}^{k} peso\ \frac{2}{ih}\ .\sum_{h=1}^{k} peso\ \frac{2}{jh}}}$$

Si el coseno del ángulo es cercano a uno, los datos se consideran similares, y si es cercano a cero son considerados diferentes. También en la limpieza es vital eliminar los nodos con deficiencias y errores de sintaxis estructural. Para finalizar la transformación de los datos, finalmente debemos introducir cambios en la herramienta "protégé", en la sección de propiedades, para obtener un gafo RDF totalmente coherente.

Con todos los nombres y las instancias limpias es necesario acceder a fuentes de datos abiertas y a los recursos universales para realizar la gestión, la búsqueda y el consumo de datos abiertos vinculados en el entorno de Google.
La primera sería seleccionar un dataset en un servidor para alojar los metadatos y nuestro grafo RDF. De lo contrario, sería preciso crear nuestra dataset y colocarlo en el servicio de Google (Fig. 5).

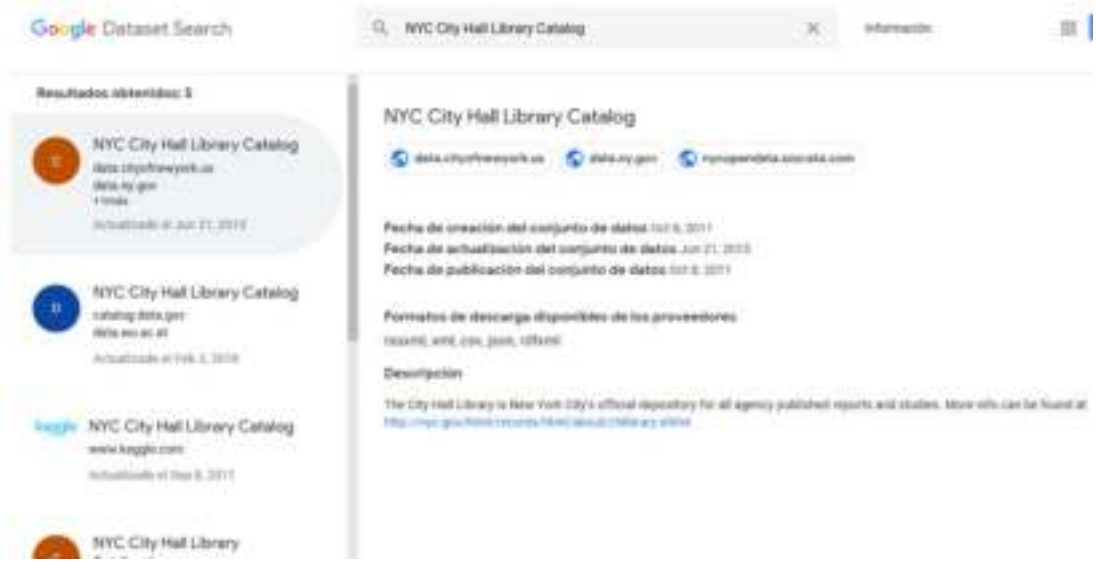


Fig. 5 - Búsqueda de Dataset de Google.

Si no existiera una dataset para la entidad de información, nuestra recomendación para las bibliotecas sería utilizar DCAT,[38] que provee de un amplio espectro para el manejo de información de datos bibliográficos (Fig. 6).

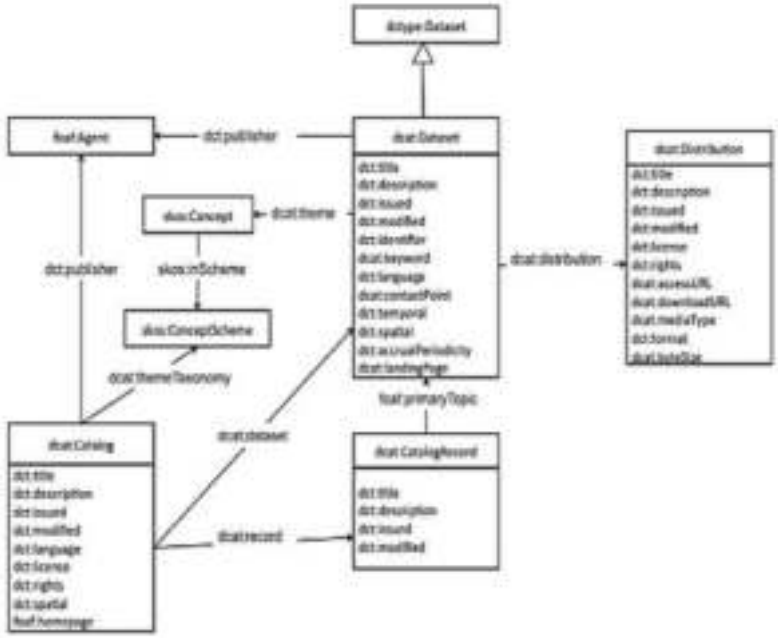


Fig. 6 - DCAT.

1.5. Transformación del RDF.

Se busca eliminar los errores que se dan en el proceso de publicación del grafo RDF. Los metadatos del conjunto de datos, en algunas ocasiones, no están dispuestos junto con los publicados, están incompletos o no pueden tratarse debido a la forma en que se generaron. Este proceso puede realizarse usando *LOD Laundromat*.[39,40] Esta herramienta elimina los grafos duplicados, los errores de sintaxis y los nodos vacíos. Los datos vinculados se generan usando un conjunto de reglas que tiene como base R2RML y RML. Estos lenguajes están construidos teniendo en cuenta la capacidad de procesamiento de la máquina, y evita que los usuarios puedan comprender las reglas descritas dentro de ellos, lo que impide que se gestionen las anotaciones deseadas para corregir el orden de los datos. En esta metodología se necesita describir las reglas necesarias para los bibliotecarios o expertos en información. Proponemos escribir estas reglas con YARRRML, como se muestra a continuación:[41]

```
{
2
"persons": [
3
{"firstname": "Jose", "lastname": "Martí"
},
```


La tarea de transformación del RDF de *bibframe* requiere constatar que las instancias están publicadas en la web. Para esto utilizamos la plataforma LOD 4ALL https://lod4all.net/index.html, una herramienta que facilita el *browsing* y el acceso a datos bibliográficos que ya están en formato html publicadas en *linked data*. Además, provee un entorno de desarrollo para que las plataformas usen *linked data*.

1.6. Enlazado de datos.

El enlazado de datos es la etapa donde se generan enlaces a los grafos RDF previamente construidos, lo que facilita el enriquecimiento de los grafos publicados. Es una tarea donde se generan enlaces del tipo *owl:sameAs* entre el grafo RDF obtenido y los grafos RDF de la base de datos que se desee enlazar. Para generar los enlaces se utiliza la herramienta *Silk*,[42] la cual proporciona un lenguaje declarativo de alineación. Se reconoce que Silk proporciona acceso a *SPARQL Endpoints* remotos para realizar el enlazado.

1.7. Validación.

Con la validación de los datos del grafo debemos acometer dos tareas que son básicas en los procesos de enlazado:

a) Validación de las sintaxis:

La validación de las sintaxis se realiza para que el proceso de enlazado sea heterogéneo. La herramienta que estamos proponiendo para este paso de la metodología se denomina *Serd*.[43] Es una Librería en lenguaje C que proporciona la evaluación de la sintaxis RDF que admite la lectura y la escritura de *Turtle, TRiG, NTriples y NQuads*. *Serd* es adecuada para aplicaciones de rendimiento crítico o limitadas en recursos; por ejemplo,

la serialización de conjuntos de datos muy grandes como los de las bibliotecas, protocolos de red o sistemas integrados que requieren dependencias mínimas y una implementación liviana.

b) Validación de la implementación:

La validación se realiza usando el lenguaje SHACL

https://w3c.github.io/data-shapes/data-shapes-test-suite/

1.8. Generación de la base de conocimientos bibliográficos.

El proceso de la confección de la base de conocimiento conlleva el uso de *Virtuoso*.[29,44] *OpenLink Virtuoso* es un servidor universal de *CROSS PLATFORM* que implementa las funciones de servidor web, de archivos y de base de datos junto con *Native XML Storage* y *Universal Data Access Middleware* como una solución de servidor único. Incluye soporte para estándares clave de internet, web semántica y acceso a datos, tales como: XML, XPATH, XSLT, SOAP, WSDL, UDDI, WebDAV, SMTP, SQL-92, ODBC, JDBC y OLE-DB. *Virtuoso* es compatible actualmente con los siguientes sistemas operativos: Windows 95/98/NT/2000, Linux (Intel, Alpha, Mips, PPC), Solaris, AIX, HP-UX, Unixware, IRIX, UNIX digital, DYNIX / PTX, FreeBSD, SCO, Mac OS X. *Virtuoso* es un proveedor de bases de datos de última generación y de alto rendimiento para la era de la computación distribuida. Proporciona acceso a sus fuentes de datos gestadas por proveedores.

Fase 2: Desarrollo de los mecanismos de recuperación de información.

Construir un mecanismo de recuperación de información eficiente obliga a que el grafo sea capaz de recuperar la información que necesita el usuario sin ruidos, y que los mecanismos de inferencia sean lo suficientemente coherentes. Además, en esta etapa se debe diseñar las consultas formales, las consultas semánticas y su mecanismo de visualización.

2.1. Diseño del sistema de navegación.

La primera parte del sistema de navegaciones es manejar la forma en que han de verse los datos en *Pubby*.[45] *Pubby* es un frontend de datos enlazados (*Linked Data Frontend*) que proporciona vistas HTML sobre los recursos existentes en un grafo RDF almacenado en un almacén de tripletas. Su funcionamiento se basa en la reescritura de URI y el manejo de la negociación de contenidos mediante redirecciones del protocolo HTTP. La herramienta es de código abierto bajo la licencia de Apache 2.0. *Pubby* fue integrada a la plataforma BM2LOD con la finalidad de asegurar la visualización de los grafos RDF generados por las herramientas anteriores y almacenadas en *Virtuoso*.

2.2. Diseño de consultas semánticas.

Una vez descrito el sistema navegacional, los autores van a describir en este acápite las consultas de mayor nivel semántico basadas en el modelo de consultas sobre la semántica de SAPARQL, reconociendo que la semántica navegacional podría facilitar el proceso.[8] Las consultas formales son las que se establecen a nivel de catálogo y no median en ellas los grafos de conocimiento; por eso no se analizan en este trabajo.

En el diseño de consulta de grafos hay que tener en cuenta algunas problemáticas según *Arnout* y *Elbassunoi*:[26]

- Datos incompletos:

  Los grandes grafos de conocimiento en el terreno de las bibliotecas pueden contener una gran cantidad de información, ya que se valen del texto libre asociado a los campos de anotación, lo que permite ampliar el nivel de inferencia de la búsqueda en caso donde no se conocen exactamente cuáles son las instancias asociadas a la consulta. Es en este escenario donde aparece el concepto de triple-patrón. Triple patrón *Sea U* es un conjunto de URI; L es un conjunto de literales, y X es un conjunto de variables. La consulta del patrón es un conjunto de patrones triples:

$\{q_1\ q_2 \ldots q_n\}$Se cuando $q_i \in \{U \cup X\} x \{U \cup X\}\ x \{U \cup L \cup X\}$ para $1 \leq i \leq n$

?x editor ?y

?x autor ?y

- Consulta flexible:

  A pesar de que las consultas expresadas en tripletes tienen una semántica estructurada y expresiva, estas solo implementan la concordancia booleana; por tanto, es vital definir en el triplete de búsqueda mecanismos de consulta flexibles para permitir una búsqueda más efectiva de datos RDF que se mezclan con un algoritmo de semántica navegacional.

A continuación se presentan casos para ilustrar los aspectos anteriores:

Caso 1:

Cuando una de las dos constantes (URI o literales) de la consulta no existe en el gráfico de conocimiento.

? x autor Christie_ Agatha

? x editor Christie_ Agatha

El resultado de la consulta sería de la siguiente manera:

? x autor ?z [Christie_ Agatha]

? x editor ?z [Christie_ Agatha]

Caso 2:

Cuando uno o más patrones triples no coinciden en el grafo de conocimiento y, sin embargo, todas las constantes están en la consulta estamos en presencia de una consulta extendida de triple patrón. Una consulta extendida de triple patrón es un grafo con múltiples aristas. En el caso de una triple consulta de patrones, uno o más puntos del grafo pueden estar asociados con una palabra

clave algo que no se pueden expresar utilizando patrones triples. Por ejemplo, se puede utilizar el siguiente patrón extendido para obras escritas y editadas por la misma persona:

? x director Tabío_Juan Carlos
? x productor Mendoza_Izquierdo_Miguel

La consulta que ilustra esta formulación es la siguiente:

? x director Tabío_Juan Carlos
? x ?y Mendoza_Izquierdo_Miguel [productor]
? x director Tabío_Juan Carlos
? x productor ?y [Mendoza_Izquierdo_Miguel]

Caso 3:
El complemento a una consulta no puede ser respondido porque no existe información en el grafo a pesar de que cada triplete aparece individualmente en el grafo. Esta consulta se denomina triple patrón con cero resultados:

x protagonizada por Beatriz_Valdés
x director Juan_Piñera

Esta consulta no produciría ningún resultado cuando se ejecuta porque no existe ninguna película en el grafo de conocimiento que fuera dirigida por *Juan Piñera*. Cada uno de ellos se dividirá en un resultado individual, protagonizada por *Beatriz Valdés* y películas dirigidas por *Juan Piñera* respectivamente. En este caso se generan consultas relacionadas con la temática:

? x protagonizada por? y [Beatriz _Valdés]

? x director Juan_Piñera

? x ? y Betariz_Valdés [protagonizada]

? x director Juan_Piñera

? x protagonizada por Beatriz_Valdés

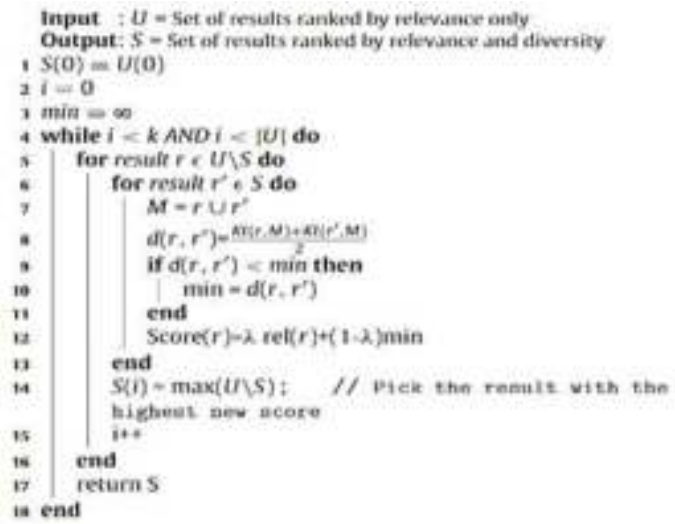
? x? y Juan_Piñera [director]

? x protagonizada por Beatriz_Valdés

? x director? y [Juan Piñera]

Orden de los resultados:

Los gráficos de conocimiento RDF producen muchos resultados, por lo que es importante ordenarlos en *ranking*. Esto es vital cuando los gráficos de conocimiento RDF y las consultas de los tripletes se extienden a los textos asociados a las descripciones bibliográficas y cuando se implementan consultas relacionadas. Para este fin, se debe tener alguna noción de relevancia o importancia para el usuario final observada en algoritmo que se describe para el orden de los resultados de la consulta de jerarquía:[8]



Diversidad de resultados:

El orden de los resultados en el *ranking* facilita que los más relevantes sean agrupados. Los mejores resultados tienden a ser homogéneos, lo que dificulta la exploración del gráfico de conocimiento. La diversidad de resultados puede desempeñar un papel importante en la información que los usuarios recuperan. A este tipo de consulta se le denomina triple patrón extendido con resultados nulos. A continuación se muestra el esquema que se realizaría si se decidiera conocer las películas protagonizadas por *Beatriz Valdés* y dirigidas por *Carlos Herrera*, así como el cuadro que ilustra la forma en que el sistema estructuraría la consulta.

? x protagonizada por? y [Beatriz _Valdés]

? x director Herrera_Carlos

Cuadro 1 - Consulta de múltiples resultados

Sujeto	Predicado	Objeto
Amor en concreto	Protagonizado Dirigida	Betariz Valdés Diego Risquez
Manuela Sáenz	Protagonizado Dirigida	Beatriz Valdés Diego Risquez
La bella del Alhambra	Protagonizado Dirigida	Beatriz Valdés Enrique Pineda Barnet
La Muerte	Dirigida	Carlos Herera
Soul	Dirigida	Carlos Herrera
Capablanca	Protagonizado Dirigida	Beatriz Valdés Manuel Herrera

2.3. Visualización de las consultas.

Las consultas se han desarrollado usando http://en.lodlive.it/. Es un producto que permite embeber las páginas y los recursos web para realizar consultas sobre ellos.[46,47] *LodLive* facilita que los recursos publicados en el grafo de conocimiento puedan ser navegables. *LodLive* está compuesto por un

complemento *jQuery (lodlive-core.js*), un mapa de configuración *JSON* (*lodlive-profile.js*), una página HTML, algunas imágenes (*sprites*) y algunos otros complementos públicos de *jQueri* (Fig. 7).

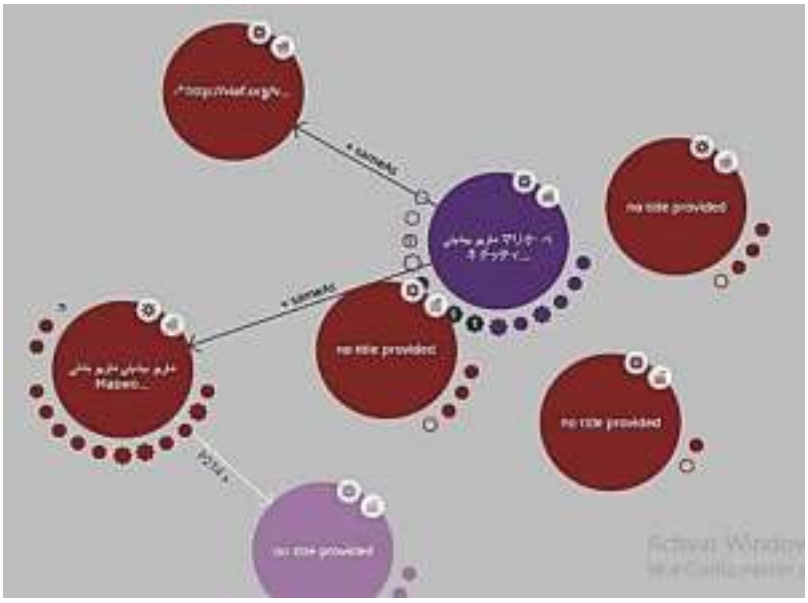


Fig. 7 - Visualización de los resultados.

Evaluación de la metodología ANCORP

La evaluación de la efectividad de esta metodología ANCORP se realiza con el objetivo de determinar la eficiencia de la recuperación de la información dentro del grafo de conocimiento propuesto. Esta evaluación se realizó mediante el uso de los *datasets* de la Biblioteca Nacional de España. Transformamos los datsets de esta biblioteca en *Bibframe*; generamos consultas sobre los datos seleccionados y evaluamos la precisión, la sensibilidad y el *f-scores* (Cuadro 2).

Cuadro 2 - Consultas semánticas

Necesidad de información	**Consulta**
Consultas de triple patrón con muchos resultados	
Libro escrito y editado por una persona	? m escrito ?x ? m editado ?x
Poesías escritas por poetas hipanoamericanos sin que concurse su país de origen	?x generonovela_hispanoamericana ?x paísdenacimiento ?p
Consultas extendidas de triple patrón con muchos resultados	
Autores de obras literarias escritas sobre la guerra civil española	?m escritores? c [guerra civil española]
Escritores cubanos que nacen el día del Natalicio de José Martí	?b escritores ?x[cubanos] ?x fecha de natalicio?a ?x José _Martí
Libros escritos por *Josué Lezama Lima* y prologados por *José Miró*	? m escritor José_Lezama_Lima ?m prologuista José _Miro
Consultas de triple patrón extendido con cero resultados	
Corriente literaria que influencia a los poetas modernistas cubanos	?w nacidosCuba ?w actividad?o[poeta] ?winfluencedBy?x
Películas musicales protagonizadas por *Lola Flores*, dirigidas y producidas por *Pedro Almodóvar*	?mprotagonosta?x[musicales] ?mprotagonizadas?y ?xproducidas?y ?mdirectorPedro_Almodovar

La tabla muestra los valores VP, FP, FN, precisión, *Recall* y *F-Measure* del modelo de consultas para datos en *bibframe*. Los resultados en cuanto a la

precisión son superiores a 0,95 en todas las consultas. En cuanto a la sensibilidad y *F-Measure*, los valores para algunas consultas son mayores o iguales que 0,98; por tanto, se consideran muy buenos estos resultados. En algunas consultas estos resultados comienzan a disminuir propiciado principalmente por la forma de construcción del fichero. Esto afecta la recuperación de información. Es necesario señalar que, aunque puede suceder, la mayoría de los errores que se presentan en el trabajo con consultas semánticas se deben a errores tipográficos (Cuadro 3).

Cuadro 3 - Resultado de las consultas

Necesidad de información	Precisión	Recall	F-Score
Consultas de triple patrón con muchos resultados			
Libro escrito y editado por una persona	0,96	0,90	0,93
Poesías escritas por poetas hispanoamericanos sin concurse su país de origen	0,98	0,93	0,98,5
Consultas extendidas de triple patrón con muchos resultados			
Autores de obras literarias escritas sobre la guerra civil españolas			
Escritores cubanos que nacen el día del Natalicio de José Martí	0,98	0,98	0,98
Consultas extendidas de triple patrón con muchos resultados			
Libros escritos por Josué Lezama Lima prologados por José Miro	0,98	0,86	0,91
Consultas de triple patrón extendido con cero resultados			
Corriente literaria que influencia a los poetas modernistas cubanos	0,97	0,70	0,81
Películas musicales protagonizadas por Rosario Flores dirigidas y producidas por *Pedro Almodóvar*	0,98	0,86	0,91

Conclusiones

Los grafos de conocimiento se han utilizado en muchos de estos dominios; sin embargo, en la actividad informacional son escasos los usos de estas herramientas, ya que las propuestas son generadas desde el punto de vista computacional, lo que es una barrera ante el gremio de la documentación.

La metodología propuesta integra dos fases que permiten la solución de dos problemas básicos en las Ciencias de la Información. Por un lado, maneja una lógica de integración de datos, las formas de migración y los procesos de limpieza de grafos; por otro, mejora la calidad de la recuperación de la información y la visualización de los resultados, y brinda un producto final con alto valor agregado.

Los resultados de las consultas en cuanto a sensibilidad, precisión y *recall* son altos. Hay que tener en cuenta que todos superan más del 0,8. Los problemas de instancias repetidas y de documentos mal descritos han generado ruido en los procesos y disminuyen la eficiencia de la búsqueda con triple patrón extendido con múltiples y 0 resultados.

La recuperación de información es muy eficiente con el grafo de conocimiento gestado con ANCORP al construirse un amplio marco de relaciones a nivel de grafo que facilitan conectar todos los elementos que subyacen en la consulta semántica. La precisión en el experimento presenta resultados sobre 0,95, y la sentencia primera es la menos precisa con 0,96 por ser igualmente menos semántica y más ambigua.

El recobrado es una medida directamente proporcional a la precisión. El aumento de una lleva a la disminución de la otra; sin embargo, el valor más bajo de recobrado es 0,86, lo que indica que un bajo porcentaje de los documentos que estaban en los registros RDA no fueron recuperados con la consulta. Estos niveles, aunque son muy buenos, pueden mejorarse si se mejoran las técnicas de incrustación del grafo y se perfecciona la semántica del grafo para disminuir la ambigüedad. La medida *F-Measure* indica que la

armonía entre precisión y recobrado es positiva y facilita una recuperación de información satisfactoria, lo que prueba la eficacia del procedimiento ANCORP.

Referencias bibliográficas

1. Fan M, Zhou Q, Zheng TF, Grishman R. Distributed representation learning for knowledge graphs with entity descriptions. Pat Recogn Let. 2017;93:31-7.
2. Faralli S, Finocchi I, Ponzetto SP, Velardi P. CrumbTrail: An efficient methodology to reduce multiple inheritance in knowledge graphs. Knowl Bas Syst. 2018;151:180-97.
3. Faralli S, Panchenko A, Biemann C, Ponzetto SP, editors. Linked disambiguated distributional semantic networks. International Semantic Web Conference. Springer; 2016.
4. Qiao B, Fang K, Chen Y, Zhu X. Building thesaurus-based knowledge graph based on schema layer. Clust Comp. 2017;20(1):81-91.
5. Maia A, Lopes JB, Martins P, Pessoa T. Authoring tools as instruments for a new approach of educational planning. In: Chova LG, Martínez AL, Torres IC, editors. INTED: 9th International Technology, Education and Development Conference. INTED Proceedings; 2015. p. 5149-58.
6. Suárez-Figueroa MC. NeOn Methodology for building ontology networks: specification, scheduling and reuse [Tesis Doctoral]. Universidad Politécnica de Madrid; 2010.
7. Lesnikova T, David J, Euzenat J, editors. Cross-lingual RDF thesauri interlinking. 10th International Conference on Language Resources and Evaluation; 2016.
8. Shekarpour S, Marx E, Auer S, Sheth AP, editors. RQUERY: Rewriting Natural Language Queries on Knowledge Graphs to Alleviate the Vocabulary Mismatch Problem. Association for the Advancement of Artificial Intelligence; 2017.
9. Qian X, Hu Y, Pan JC. Research of Chinese Word Knowledge Graph Based on SLPA Algorithm. DEStech Transactions on Engineering and Technology Research. 2017.
10. Jana A, Goyal P. Can Network Embedding of Distributional Thesaurus be Combined with Word Vectors for Better Representation? ArXiv preprint; 2018.

11. Nie B, Sun S. Knowledge graph embedding via reasoning over entities, relations, and text. Fut Gener Comp Syst. 2019;91:426-33.
12. Silva VS, Handschuh S, Freitas A, editors. Recognizing and justifying text entailment through distributional navigation on definition graphs. Thirty-Second AAAI Conference on Artificial Intelligence; 2018.
13. Martínez-Fernández S, dos Santos PSM, Ayala CP, Franch X, et al. Aggregating empirical evidence about the benefits and drawbacks of software reference architectures. ACM-IEEE International Symposium on Empirical Software Engineering and Measurement; 2015. p. 154-63.
14. Zhu G, Iglesias CA. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. Exp Syst Appl. 2018;101:8-24.
15. Wang Z, Zhang J, Feng J, Chen Z, editors. Knowledge graph and text jointly embedding. Proceedings of the conference on empirical methods in natural language processing; 2014.
16. Wang Z, Zhang J, Feng J, Chen Z, editors. Knowledge Graph Embedding by Translating on Hyperplanes. Association for the Advancement of Artificial Intelligence; 2014.
17. Guo S, Wang Q, Wang B, Wang L, Guo L, editors. Semantically smooth knowledge graph embedding. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015.
18. Lin Y, Liu Z, Sun M, Liu Y, Zhu X, editors. Learning entity and relation embeddings for knowledge graph completion. AAAI; 2015.
19. Jia Y, Wang Y, Jin X, Cheng X. Path-specific knowledge graph embedding. Knowl Bas Syst. 2018;151:37-44.
20. Chen J, Chen Y, Zhang X, Du X, Wang K, Wen JR. Entity set expansion with semantic features of knowledge graphs. J Web Sem. 2018;52:33-44.
21. Tonon A, Catasta M, Prokofyev R, Demartini G, Aberer K, Cudre-Mauroux P. Contextualized ranking of entity types based on knowledge graphs. Web Sem Sci Serv Agen World Wide Web. 2016;37:170-83.
22. Dou J, Qin J, Jin Z, Li Z. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage. J Vis Lang Comp. 2018;48:19-28.

23. Achichi M, Lisena P, Todorov K, Troncy R, Delahousse J, editors. DOREMUS: A graph of linked musical works. International Semantic Web Conference; 2018.
24. de Boer V, Melgar L, Inel O, Ortiz CM, Aroyo L, Oomen J, editors. Enriching media collections for event-based exploration. Research Conference on Metadata and Semantics Research; 2017.
25. Shan Y, Li M, Chen Y. Constructing target-aware results for keyword search on knowledge graphs. Data Knowl Engin. 2017;110:1-23.
26. Arnaout H, Elbassuoni S. Effective searching of RDF knowledge graphs. J Web Sem. 2018;48:66-84.
27. Mantle M, Batsakis S, Antoniou G. Large scale distributed spatio-temporal reasoning using real-world knowledge graphs. Knowl Bas Syst. 2019;163:214-26.
28. Rodríguez Perojo K, Leyva Mederos AA, Senso Ruíz JA. Marco procedimental para facilitar la interoperabilidad en el contexto de la Biblioteca Virtual en Salud de Cuba: el modelo Ontomed. Rev Cubana Inform Cienc Salud. 2016;27(4):456-73.
29. Southwick SB. A guide for transforming digital collections metadata into linked data using open source technologies. J Libr Metad. 2015;15(1):1-35.
30. Cofield MC, Marchock A, Melanson D, Ringwood A. BIBFRAME beginnings: educating ourselves for the linked data future. UT Faculty/Researcher Works; 2017.
31. Jin Q, Hahn JF, Croll G. BIBFRAME transformation for enhanced discovery. Libr Resour Techn Serv. 2016 [acceso: 28/03/2021];60(4). Disponible en: http://hdl.handle.net/2142/90248
32. Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Transact Knowl Data Engin. 2012;25(5).2.
33. Christen P, editor. An open source data cleaning, duplication and record linkage system with a graphical user interface. Nevada, EE.UU.: Proceedings of the 14th ACM SIGKDD International; 2008.
34. Baxter R, Christen P, Churches T, editors. A comparison of fast blocking methods for Record Linkage. Washington DC. ACM KDD'03 workshop on data cleaning, Record linkage and object consolidation; 2003.

35. Christen P, Goiser K, editors. Quality and complexity measures for data linkage and duplication quality measures in data mining. Berlin Heidelberg: Springer-Verlag;
2007.
36. Bilenko M, Mooney RJ, editors. Adaptive duplicate detection using learnable string similarity measures. Washington, DC: IX ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2003.
37. Batini C, Scannapieco M. Data and Information Quality Cham. New York: Springer International Publishing; 2016.
38. Klímek J. DCAT-AP representation of Czech National Open Data Catalog and its impact. J Web Sem; 2018.
39. Beek W, Rietveld L, Bazoobandi HR, Wielemaker J, Schlobach S, editors. LOD laundromat: a uniform way of publishing other people's dirty data. International Semantic Web Conference; 2014.
40. Rietveld L, Beek W, Hoekstra R, Schlobach S. Meta-data for a lot of LOD. Sem Web. 2017;8(6):1067-80.
41. Heyvaert P, De Meester B, Dimou A, Verborgh R, editors. Declarative Rules for Linked Data Generation at Your Fingertips! European Semantic Web Conference; 2018:
42. Hidalgo-Delgado Y, Senso JA, Leiva-Mederos A, Hípola P. Gestión de fondos de archivos con datos enlazados y consultas federadas. Rev Esp Docum Cient. 2016;39(3):145.
43. Michalek T. Implementation of Parser for RDF Data Files [Tesis]. Checoslovaquia: Universidad Ostrava; 2016.
44. García A, Linaza MT, Franco J, Juaristi M. Methodology for the publication of linked open data from small and medium size DMO. Information and Communication Technologies in Tourism; 2015. p. 183-95.
45. Bizer C, Jentzsch A, Cyganiak R. State of the LOD Cloud. Berlín: Public Web-page; 2011 [acceso: 28/06/2012]. Disponible en: http://lod-cloud.net/state/
46. Morato J, Sánchez-Cuadrado S, Ruiz-Robles A, Moreiro-González JA. Visualización y recuperación de información en la web semántica. El Profes Inform. 2014;23(3):2.

47. Gómez-Romero J, Molina-Solana M, Oehmichen A, Guo Y. Visualizing large knowledge graphs: A performance analysis. Fut Gener Comp Syst. 2018;89:224-38.

Conflicto de intereses

Los autores declaran que no existe conflicto de intereses.

Contribución de los autores

José A. Senso Ruiz: Conceptualización, metodología, *software*, redacción.

Amed Leiva Mederos: Metodología, *software*, redacción.

Yorbelis Rosell León: Conceptualización, metodología, supervisión, redacción, revisión y edición.

Ania Hernández Quintana: Metodología, supervisión, redacción, revisión y edición.

[a]Cadenas ocultas de *Markov*.