Estudio de Wikidata desdel el punto de vista del enriquecimiento semántico | Study of Wikidata from the point of view of semantic enrichment

Preprint · June 2020

DOI: 10.13140/RG.2.2.11305.42082

1 author:

Kevin León-Gavilanez
Universidad de Salamanca
1 PUBLICATION **0** CITATIONS

UNIVERSIDAD DE SALAMANCA

FACULTAD DE TRADUCCIÓN Y DOCUMENTACIÓN

GRADO EN INFORMACIÓN Y DOCUMENTACIÓN

Trabajo de Fin de Grado

ESTUDIO DE WIKIDATA DESDE EL PUNTO DE VISTA DEL ENRIQUECIMIENTO SEMÁNTICO

AUTOR:

Kevin Francisco León Gavilanez

DIRECTOR:

Ángel Francisco Zazo Rodríguez

Salamanca, 2020

UNIVERSIDAD DE SALAMANCA

FACULTAD DE TRADUCCIÓN Y DOCUMENTACIÓN

GRADO EN INFORMACIÓN Y DOCUMENTACIÓN

Trabajo de Fin de Grado

ESTUDIO DE WIKIDATA DESDE EL PUNTO DE VISTA DEL ENRIQUECIMIENTO SEMÁNTICO

Study of Wikidata from the point of view of semantic enrichment

AUTOR:

Kevin Francisco León Gavilanez

DIRECTOR:

Ángel Francisco Zazo Rodríguez

Salamanca, 2020

LEÓN GAVILANEZ, Kevin Francisco

Texto (visual) : sin mediación

Estudio de Wikidata desde el punto de vista del enriquecimiento semántico = Study of Wikidata from the point of view of semantic enrichment / Kevin Francisco León Gavilanez; director, Ángel Francisco Zazo Rodríguez. – Salamanca: Universidad de Salamanca, Facultad de Traducción y Documentación, 2020.

48 p.: il. ; 30 cm

Trabajo de Fin de Grado – Grado en Información y Documentación

1. Bases de datos. 2. Fundación Wikimedia (Estados Unidos). 3. Universidad de Salamanca (España). I. Zazo Rodríguez, Ángel Francisco, dir. II. Título III. Título: Study of Wikidata from the point of view of semantic enrichment.

004.658.2

Resumen

El presente estudio trata de uno de los proyectos más novedosos de la Fundación Wikimedia: la Wikidata. Wikidata es una base de datos colaborativa que dispone de herramientas de consulta que son usadas profusamente como mecanismo para la obtención de datos, de relación con otros datos, y de enriquecimiento de información y conocimiento.

Este estudio se ha centrado en conocer el estado de la cuestión sobre aspectos generales de Wikidata (su funcionamiento, su organización, cómo se alimenta de datos, etc.), además de analizar el enriquecimiento semántico desde y hacia Wikidata desde bibliotecas y archivos. También se ha analizado la relación entre Wikidata y DBpedia.

Se incluyen estudios que utilizan estas dos herramientas en relación con su potencial de enriquecimiento semántico en bibliotecas y entornos educativos. También se incluyen recomendaciones que permitan mejorar Wikidata, desde el punto de vista de los artículos, su procedencia, las restricciones que pueden realizar y el propio enriquecimiento semántico.

Palabras clave

Wikidata; DBpedia; enriquecimiento semántico; datos abiertos enlazados

Abstract

This study deals with one of the Wikimedia Foundation's newest projects: Wikidata. Wikidata is a collaborative database that has consultation tools that are used extensively as a mechanism to obtain data, to relate to other data, and to enrich information and knowledge.

This study has focused on knowing the state of the question on general aspects of Wikidata (its functioning, its organization, how it is fed with data, etc.), as well as analyzing the semantic enrichment from and to Wikidata from libraries and archives. The relationship between Wikidata and DBpedia has also been analysed.

Studies using these two tools are included in relation to their potential for semantic enrichment in libraries and educational environments. Recommendations are also included to improve Wikidata, from the point of view of the articles, their origin, the restrictions they can perform and the semantic enrichment itself.

Keyworks

Wikidata; DBpedia; semantic enrichment; linked open data

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

A mis padres y hermanos, a quienes solo puedo expresar mi sincero agradecimiento por apoyarme y por tener esa paciencia infinita durante toda mi etapa académica que hoy culmina.

A mi querido amigo Adrián, por ser mi fuente de inspiración en la vida y por su apoyo incondicional en todo momento.

Y a mi profesor Ángel, por su acompañamiento en el desarrollo de este trabajo y por ser faro en este camino.

1. INTRODUCCIÓN

1.1. Presentación

El trabajo que presentamos para obtener el Grado en Información y Documentación se enmarca en el estudio y la revisión bibliográfica sobre uno de los proyectos novedosos de la Fundación Wikimedia, la Wikidata. Wikidata es una base de datos de conocimiento, que nos ofrece un conjunto de datos consistente y estructurado, y permite la descripción de entidades del mundo real y de relacionarlas entre ellas, permitiendo expresarlo a través de un gráfico el conocimiento.

El objetivo de este trabajo es estudiar Wikidata como base de datos colaborativa, analizar cómo desde las bibliotecas y archivos se está realizando de manera muy activa el enriquecimiento semántico de Wikidata y también analizar su relación con DBpedia, otro proyecto de Wikimedia para extraer conocimiento estructurado de la propia Wikipedia.

Este trabajo está dirigido por el profesor D. Ángel Francisco Zazo Rodríguez, del Departamento de Informática y Automática, de la Facultad de Traducción y Documentación, de la Universidad de Salamanca.

1.2. Justificación del estudio

El tema del estudio se ha elegido en función de las propuestas dadas por el director del Trabajo de Fin de Grado, siendo seleccionada el proyecto Wikidata.

Este tema nos ha presentado un gran interés, debido a que es un proyecto de la Fundación Wikimedia novedoso y está relacionado con la organización del conocimiento, por lo que nos hemos decantado por estudiar particularmente cómo funciona esta base de datos.

Wikidata, al igual que Wikipedia, supone un ejemplo de voluntarismo desinteresado muy útil para una enorme cantidad de usuarios y procesos. Wikidata dispone de herramientas de consulta que son utilizadas profusamente como mecanismos para obtención de datos, de relación con otros datos y de enriquecimiento de información y conocimiento.

1.3. Objetivo del estudio

El objetivo principal del estudio es analizar Wikidata como base de datos colaborativa, así como el enriquecimiento semántico de los datos que almacena esta base de daos, y su relación con la DBpedia.

Para llevar a cabo este objetivo principal se necesita primero establecer el estado de la cuestión centrado en aspectos generales de Wikidata (su funcionamiento, su

organización, cómo se alimenta de información, cómo se puede acceder a esa información, etc.). Un segundo objetivo es analizar el enriquecimiento semántico desde y hacia Wikidata desde bibliotecas y archivos. Finalmente, también es nuestro objetivo estudiar cómo es la relación entre Wikidata y DBpedia.

1.4. Método

El estudio se ha basado previamente en un estado de la cuestión para conocer sobre el tema de estudio y cómo es su situación actual.

Posteriormente, se ha realizado una revisión bibliográfica recuperada en las bases de datos electrónicas tales como Web of Science (WOS), SCOPUS, LISA, LISTA, Dialnet, CSIC, e IEEE Xplore, de las cuales explicaremos con más detalle posteriormente. Estas búsquedas se realizaron en un periodo que va desde noviembre de 2019 a febrero de 2020. Además, se han utilizado de manera complementaria otras fuentes de información.

Los términos que se emplearon en las búsquedas fueron: "*Wikidata*"; "*DBpedia*"; "*Enriquecimiento semántico*"; "*Semantic enrichment*". Estos términos se han empleado tanto en inglés como en español, ya que las bases de datos usadas son multilingües, internacionales y nacionales, y nos permitirá obtener resultados en ambos idiomas.

Por último, todos los documentos empleados en este trabajo se citaron y redactaron con las normas APA (*American Psychological Association*).

1.5. Estructura de la memoria

La estructura del trabajo se organiza de la siguiente forma:

En primer lugar, hablamos de Wikidata. En este apartado hablamos de Wikidata como base de datos del conocimiento, así como su modelo de datos, su calidad, su edición, y, por último, de su comunidad.

En segundo lugar, se estudia el enriquecimiento semántico de datos en Wikidata. En este apartado hablamos de la propia Web Semántica y su vinculación con Wikidata, así como las tecnologías que emplea para su funcionamiento, el *Linked Data*, el enriquecimiento semántico de datos, y, por último, se ofrece un caso práctico de datos enriquecidos de la Biblioteca Virtual de Polígrafos.

En tercer lugar, se estudia Wikidata y su relación con la DBpedia. En este apartado hablamos de los *infoboxes*, y de las diferencias que existen entre estas dos bases de datos.

En cuarto lugar, se presentan unas recomendaciones. En este apartado se aportan recomendaciones propias y de las derivadas de la revisión bibliográfica.

En quinto lugar, se muestran las conclusiones que se han obtenido una vez acabo el estudio.

Por último, y, en sexto lugar, se recoge la bibliografía y fuentes de información empleada en este estudio.

2. WIKIDATA

Wikidata es un proyecto creado en 2013 que forma parte del conjunto de iniciativas que realizó la Fundación Wikimedia, siendo Wikipedia uno de estos proyectos más conocidos a nivel mundial. Podemos decir que este proyecto se creó con intención de reunir todas las informaciones generadas en la Wikipedia y de almacenarlas en una base de datos central.

A partir de entonces, según autores como Vrandečić y Krötzsch (2014), explican que esta base de datos desde 2013 obtuvo a principios de este año más de 40.000 contribuciones y más de 3.500 usuarios activos. Un estudio posterior llevado a cabo por Ismayilov y otros (2015) recoge que Wikidata contiene más de 20 millones de artículos, 87 millones de informes, y más de 6.000 usuarios activos, por lo que podemos afirmar que este proyecto ha ido incrementando de forma exponencial gracias a las aportaciones de los contribuidores que trabajan para que esta base de datos tenga mayor repercusión en la sociedad del conocimiento.

Asimismo, creemos que uno de los puntos fuertes, a los que se debe este gran crecimiento, es que la comunidad que trabaja en este proyecto se preocupa que los datos que se ingresan en la base de datos sigan unas reglas de restricción para controlar los posibles errores y conservar una estructura uniforme en lo técnico y en relación con el contenido. Esta creencia nos la afirma Vrandečić y Krötzsch (2014) en su estudio donde explica que gracias al control de estos datos permite que Wikipedia se convierta en un recurso más preciso, útil e informativo. Con ello, podemos aportar que proporciona estabilidad, porque hay un control, debido a que presenta los datos de forma estructurado y sin errores, y continuidad, porque permite la participación para el desarrollo futuro de esta herramienta.

Además de estas aportaciones, podemos añadir otras que vienen recogidas del estudio realizado por Farda-Sarbas y Müller-Birn (2019) donde nos indica que Wikidata, además de estar diseñada para editar y almacenar datos, también permite la ventaja de consumir y reutilizar distintos datos en diversos idiomas. Gracias a que proporciona datos estructurados, estos se pueden recuperar y acceder a ellos empleando la herramienta de consulta SPARQL[1], herramienta que también se utiliza en DBpedia, siendo, en ese aspecto de recuperación de información un tanto competidor o, más bien, complementaria de Wikidata. Por último, nos explica que el diseño de esta base de daos permite proporcionar datos con referencias, debido a que trabaja con Wikipedia, y atiende a las necesidades informacionales que puedan surgirles a los usuarios y colaboradores.

[1] SPARQL o *SPARQL Protocol and RDF Query Language* es un lenguaje estandarizado para la consulta de grafos RDF, normalizado por el Consorcio W3C. Es una tecnología clave en el desarrollo de la Web Semántica.

Para finaliza, no hay que olvidar que estos datos estructurados que proporciona esta base de datos son fuente fundamental para bibliotecas y archivos, y que son usados, por ejemplo, para ayudar al enriquecimiento semántico de sus registros de autoridad. Además, podemos añadir que este enriquecimiento permitirá mejorar la búsqueda y recuperación de información (Saorín y Pastor Sánchez, 2018).

2.1. La Wikidata como base de datos del conocimiento

Wikidata se ha convertido en una base de datos de conocimiento gracias a que tiene la capacidad de ofrecernos un conjunto de datos consistente y estructurado, y, además, de describir entidades del mundo real y de relacionarlas entre ellas, permite expresarlo a través de un gráfico del conocimiento.

Antes de adentrarnos en cómo son sus datos, hemos considerado oportuno destacar los puntos esenciales que posee esta base de datos y que manifiestan la razón de ser tan utilizada por millones de usuarios. Lo demostramos a través del estudio de Vrandečić y Krötzsch (2014), y del cual hemos adaptado en relación con nuestro tema de estudio. Estas características fundamentales que muestran la esencia de Wikidata son las siguientes:

Característica	Descripción
Edición libre	Cada usuario puede editar y modificar la información salvada en la base de datos, sin necesidad de registrarse.
Comunidad	Los datos se controlan por una comunidad de colaboradores y establecen el posicionamiento de estos datos introducidos.
Pluralidad	Es posible que haya datos que se consideren conflictivos, debido a la diversidad lingüística, por lo que esta base de datos permite que coexistan entre ellos con una buena organización.
Fuentes secundarias	No solo se recopila información proveniente de fuentes primarias, sino también secundarias para permitir la contrastación de los hechos.
Capacidad idiomática	Los datos están en diversos idiomas debido a su influencia internacional y a su diseño.
Accesibilidad	Los datos se pueden usar en distintos formatos como por ejemplo en RDF o JSON. Estos datos están bajo términos legales que permitan su reutilización.
Continuo crecimiento	Debido a la participación de la comunidad, esta base de datos tiene un continuo crecimiento.

Tabla 1. Características esenciales de Wikidata. Fuente: Elaboración propia a partir de (Vrandečić y Krötzsch, 2014)

Con estas características esenciales de las que dan el enfoque a Wikidata, podemos decir que se ha convertido en una fuente de datos en crecimiento y muy importante en la

Wikimedia, así como un almacén donde se va guardando de forma libre la suma de todo el conocimiento existente.

2.2. El modelo de datos

Los datos que se almacenan en Wikidata son expresados por medio de elementos y propiedades. Por una parte, estos elementos se reflejan a través de los artículos en los que se hace referencia a la creación de cosas, resúmenes, conceptos y categorías. Por otra parte, las propiedades se encargan de relacionar estos elementos con otros, y normalmente se aprovechan para indicar hechos.

Para ejecutar estas relaciones se emplean los URI[2], que son empleados por los propios editores de la comunidad, ya que son estos los que pueden agregar etiquetas, descripciones, así como términos que puedan ser legibles para las personas, es decir, hace referencia a los idiomas que se usan en esta base de datos.

Estos editores utilizan estas relaciones para representar los datos estructurados a través de las llamadas "reclamaciones". Estos emplean la función propiedad-valor para poder recuperar y visualizar estos datos posteriormente a través de una solicitud en el sistema de consultas.

El rasgo principal como observamos en el modelo de datos de Wikidata son las reclamaciones. Estos también permiten avanzar en la representación de estos datos a través de un valor permitido en propiedades. No obstante, puede existir ciertas restricciones, en las que se reflejan las condiciones que deben cumplir la propiedad del dato. Aunque en un estudio por Piscopo, Phethean y Simperl (2017) indica que estas restricciones no son aplicables en el modelo de datos de Wikidata, debido a que se emplean únicamente para controles de calidad en los datos.

Además, cabe destacar que esta base de datos emplea los Knowledge Graphs o gráficos de conocimientos ya que en un principio Wikidata, según un trabajo por Staab y Studer (2009), se construyó en este tipo de colecciones en los que se describirían entidades y su relación entre ellas.

Por último, a modo de ejemplo mostramos una propiedad donde se muestra que el dato "Harry Potter" con identificador Q8337[3] (Figura 1), se relaciona con el elemento "novel series", dado que es una instancia de este tipo.

[2] URI o *Uniform Resource Identifier* permite identificar recursos en Internet. Véase: https://www.ecured.cu/URI
[3] Véase: https://www.wikidata.org/wiki/Q8337

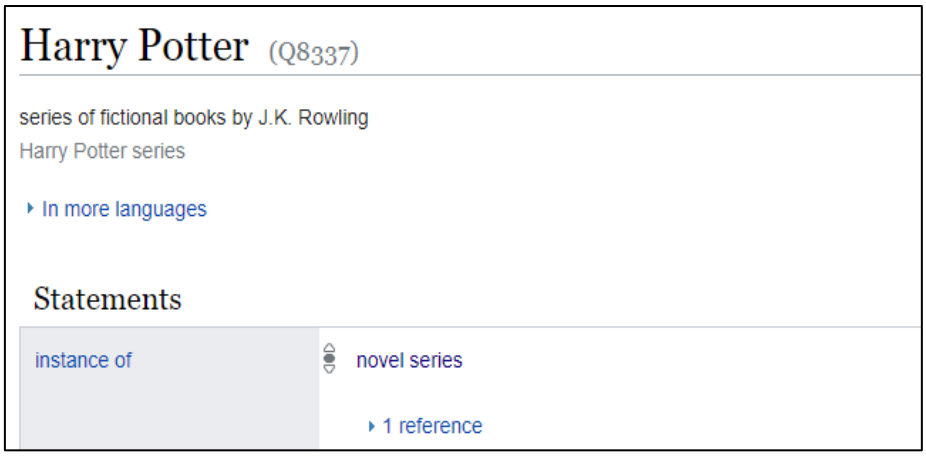

Figura 1. Propiedad y elemento. Fuente: Wikidata

2.3. Calidad de los datos

Desde que Wikidata fue lanzada en 2013, ha ido incrementando la preocupación por la calidad de los datos que almacena. Existen algunos estudios donde se muestra esta inquietud como, por ejemplo, el realizado por Brasileiro y otros (2016) donde hemos observado que existen algunos inconvenientes en las jerarquías de las taxonomías que maneja Wikidata. Esto significa en los continuos patrones de error que se generan al mal empleo de las propiedades.

En esta misma línea, otro estudio realizado por Thakkar y otros (2016) refleja que Wikidata tiene un nivel bueno con otras colecciones del conocimiento, ya que trabajaba con datos abiertos. Sin embargo, en el estudio de Färber y otros (2016) donde se realizó una comparación con otras bases de datos como, por ejemplo, DBpedia y Freebase, en relación con la calidad de sus datos, se reflejó que Wikidata tenía el mismo nivel que estas bases de datos, aunque hemos podido comprobar que destacaba en su integridad en su esquema de datos, a diferencia de los demás.

Por último, aunque se ha visto que la calidad podría estar vinculada con la propia base de datos que se emplee, hemos observado que la calidad no solo depende de la base, sino de la tipología de usuarios. Esto podemos observarlo en el estudio por Piscopo, Phethean y Simperl (2017), donde se hace un hincapié en la tipología de usuarios, ya que estos son los responsables de la buena calidad de estos datos. Además, son estos los que deben seguir las indicaciones o restricciones que la propia comunidad establece para crear y editar los datos, y así evitar pérdida de calidad en ellos. No obstante, más adelante estudiaremos cómo podemos enriquecer estos datos.

2.4. Edición de los datos

Como habíamos dicho en el punto anterior, la comunidad es la encargada de dar las indicaciones o restricciones para la creación y modificación de los datos en Wikidata.

La comunidad, de la que hablaremos en el siguiente punto, es responsable de controlar todo lo que entra en esta base de datos, así como de gestionar todos los componentes que forman Wikidata. Se ha observado que las operaciones son fáciles y no se necesita mucho esfuerzo o conocimiento. Sin embargo, esto podría ser un problema, debido a que ahí es donde podría venir uno de los problemas a la hora de crear datos porque serán los usuarios que intervengan en este proceso.

Siguiendo con la edición, al igual que Wikipedia tiene un patrón Wikidata también tiene el suyo, aunque tienen la misma tipología de tareas, según según Haythornthwaite (2009), siendo las fáciles y las complejas.

Para mostrar cómo es la edición en Wikidata, a nivel de tarea fácil, realizaremos una edición del término “Dios”, con identificador Q4500250.

Paso n.º 1: Términos en desambiguación

En este primer paso hemos observado que el término “Dios” en inglés, español, catalán y gallego estaban en un solo idioma, el español (Figura 2).

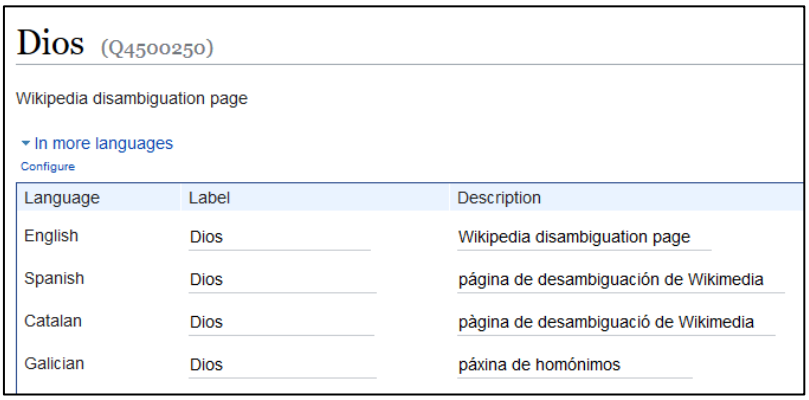

Figura 2. Término "Dios". Fuente: Wikidata

Paso n.º 2: Cambios en idioma y descripción

En este segundo paso se ha cambiado el término “Dios” a su idioma correspondiente, y descripción (Figura 3).

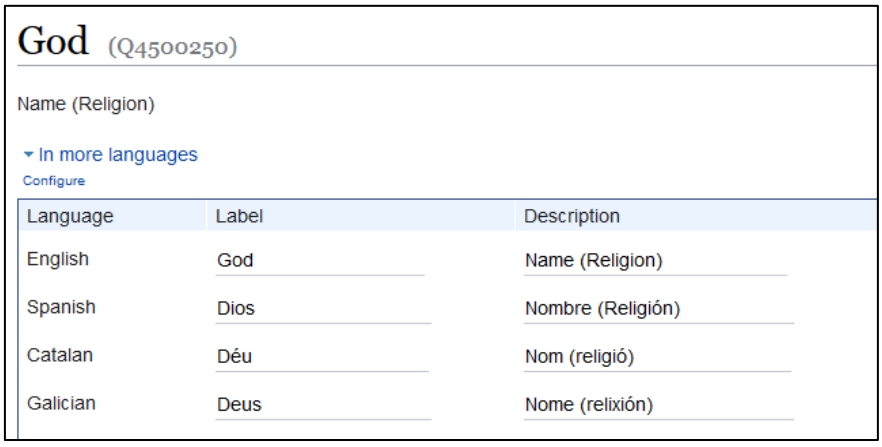

Figura 3. Idioma y descripción. Fuente: Wikidata

Paso n.º 3: Publicación de los cambios

En este último y tercer paso se ha procedido a publicar los cambios en el término “Dios” (Figura 4).

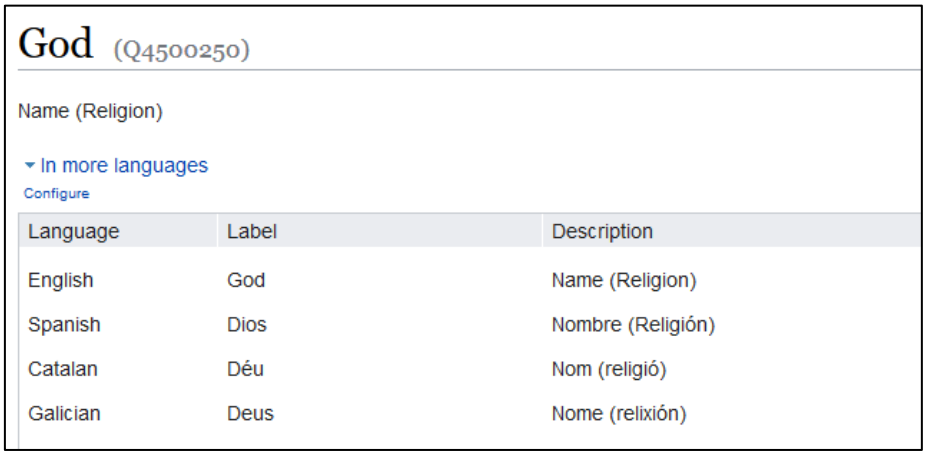

Figura 4. Publicación de los cambios. Fuente: Wikidata

Una vez concluida la edición y su posterior publicación, su aprobación dependerá de la comunidad, ya que es la encargada de revisar que los términos introducidos no producen plagio. Además, debido a que no es necesario registrarse como usuario, el sistema recoge la IP del ordenador donde se realizó la edición y lo publica en el historial (Figura 5).

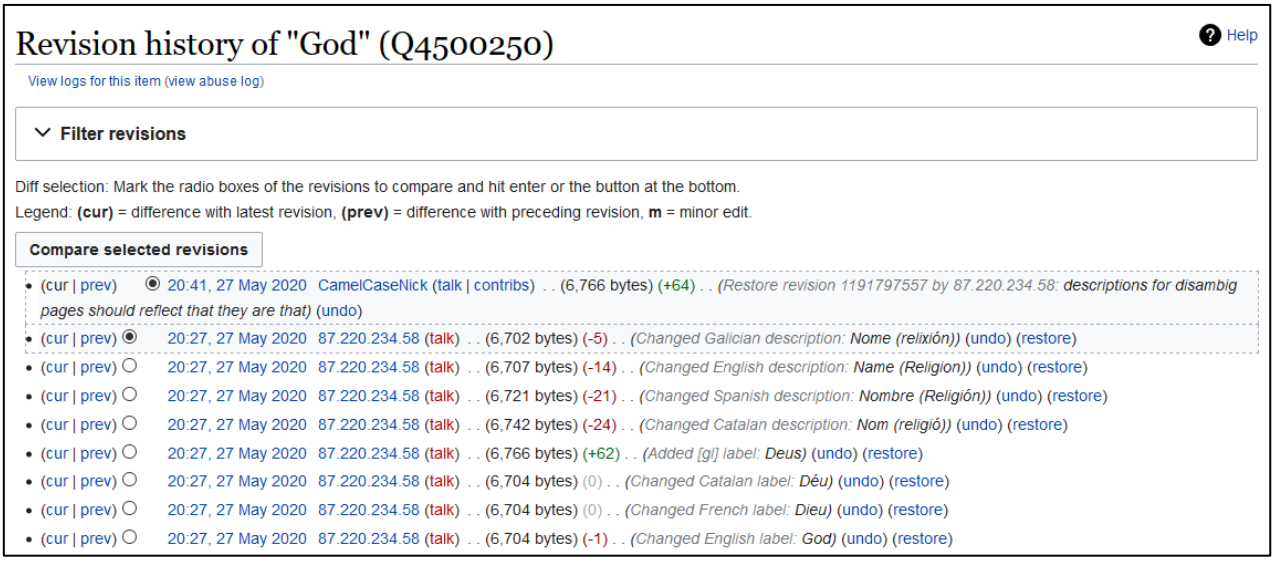

Figura 5. Historial de cambios. Fuente: Wikidata

Por último, se puede comprobar que tanto la edición como la creación es fácil de realizar por lo que cualquier usuario puede entrar y aportar su “granito de arena”, aunque esto podría ser un problema si queremos que estos datos realmente sean de calidad. Sin embargo, este trabajo de controlar la edición de términos quedará en manos de la comunidad.

2.5. La comunidad

Como habíamos dicho anteriormente, en Wikidata existe libertad para crear o editar sin necesidad de registrarse, siendo esto una de las características esenciales que forman en "alma" de Wikidata. Por tanto, los usuarios pueden ser las mismas personas o bots.

Cuando se habla de bots se consideran como los "guardianes" de controlar la calidad de los datos que se editan o crean. Estos bots son controlados por usuarios registrados siguiendo una serie de normas y políticas. Para obtener un bot, es necesario que el usuario registrado lo solicite, según explica Piscopo (2018), ya que, una vez aprobada y asignada esta herramienta por la comunidad, deberá encargarse de mantenerlo y asegurarse de que no cause ningún daño a la base de datos.

Podemos demos decir que esta herramienta es fundamental en Wikidata porque permite que exista un continuo crecimiento colaborativo gracias a la relación máquina-hombre en esta base de datos. Esta relación permite que el trabajo sea compartido, debido a que por una parte el bot permite generar una gran cantidad de datos, y, por otra, el usuario se encarga de revisarlos.

Además de estas aportaciones, podemos añadir las de Vrandečić y Krötzsch (2014) donde explican que estos bots realizan comprobaciones periódicas para evitar que existan violaciones relacionadas con la propiedad (aquella información que no está bajo licencia Creative Commons), así como también la corrección de nombres de usuarios erróneos o páginas mal redireccionadas.

Por último, en relación con los usuarios, podemos señalar que pueden contribuir de forma anónima o no, estos últimos los llamados "usuarios registrados". Estos usuarios anónimos contribuyen a que haya un incremento en la producción de datos, aunque siguiendo a Piscopo, Phethean y Simperl (2017) podemos decir que este aumento podría ser dañino en la calidad de los datos si no hay un control exhaustivo de estos.

3. METODOLOGÍA

Se comentaba que en el apartado de Introducción que para conocer el estado de la cuestión hemos realizado una revisión bibliográfica recuperada en varias bases de datos, tales como Web of Science (WOS), SCOPUS, LISA, LISTA, Dialnet, CSIC, e IEEE Xplore, de las cuales explicaremos más detalles posteriormente. Estas búsquedas se realizaron en un periodo que va desde noviembre de 2019 a febrero de 2020. También se han utilizado de manera complementaria otras fuentes de información.

Se plantearon estrategias de búsqueda empleando los siguientes términos: "*Wikidata*"; "*DBpedia*"; "*Enriquecimiento semántico*"; "*Semantic enrichment*". Se utilizaron términos tanto en inglés como en español, ya que las bases de dato empleadas son multilingües, internacionales y nacionales, y nos permitirá obtener resultados en ambos idiomas. A continuación, se detallan las acciones realizadas en las distintas búsquedas en las bases de datos, en función de la tipología de estas:

a) Bases de datos generales:

- En la base de datos **Web of Science (WOS)** se realizó una búsqueda avanzada seleccionando la colección principal de Web of Science, y buscando por el campo Titulo, aunque no se aplicó ninguna acotación en lo referente al periodo de tiempo, debido a que se quería recuperar la mayor información relacionada con nuestra temática de estudio. Posteriormente, se realizó un refinado por acceso abierto y por artículos de revista. Realizados estos cambios, se obtuvieron 10 documentos.
- En la base de datos **SCOPUS**, se realizó una búsqueda simple por Documentos, donde no se realizó ninguna acotación en el periodo de tiempo, y se refinó por el campo Título de revista, resumen y palabra clave. Tras lanzar esta búsqueda, se hizo un refinado posterior en base al acceso abierto y por artículo de revista. Por tanto, con este nuevo refinamiento se obtuvieron 29 documentos, de los cuales se excluyeron 14, debido a que no se relacionaban con nuestra temática de estudio.

b) Bases de datos especializadas en Información y Documentación:

- En la base de datos **LISA**, se realizó una búsqueda básica, donde se buscó por el campo Título del documento, por artículo como tipo de documento, y no se aplicó un periodo de tiempo concreto. Tras lanzar la búsqueda, se obtuvieron 8 resultados, pero se excluyeron 5, ya que no tenían la facilidad de consultarlos en acceso abierto.
- En la base de datos **LISTA**, se realizó una búsqueda avanzada, sin aplicar un rango de tiempo concreto. En el mismo formulario, se filtró por *Linked Full Text* para consultar el texto completo, y por artículo como tipo de documento. Tras lanzar la búsqueda nos sorprendió que sólo se recuperase 1 resultado, por lo que se

realizó una búsqueda inteligente en la que obtuvimos 225 resultados, de los cuales se excluyeron 179, ya que no se relacionaban con la temática del estudio.

c) Bases de datos nacionales:

- En la base de datos **Dialnet** se realizó una búsqueda avanzada, y por documentos. Se realizó un refinado por artículo de revista, pero sin determinar un periodo de tiempo específico. Se obtuvieron 7 resultados de los cuales no se excluyó ninguno porque se relacionaban con la temática del estudio.
- En las bases de datos del **CSIC** se realizó una búsqueda avanzada. Se realizó una búsqueda empleando el campo Título documento, y sin asignar un periodo de tiempo determinado. Tras lanzar la búsqueda, sólo se recuperaron 2 resultados, los cuales sí cumplían con la temática del estudio.

d) Bases de datos especializada en Informática y Tecnología:

- En la base de datos **IEEE Xplore** se realizó una búsqueda avanzada, en la que se buscó por el campo Titulo de documento, sin determinar un periodo de tiempo concreto. Tras lanzar la búsqueda, se realizó un refinado por acceso abierto y por artículo de revista como tipo de documento. Con la aplicación de este refinado, se obtuvieron 7 resultados, de los cuales no se excluyó ninguno, ya que trataban con la temática del estudio.

Tras las búsquedas realizadas en las diferentes bases de datos electrónicas, se muestran los resultados en la siguiente tabla:

Base de datos	Número de resultados
Web of Science (WOS)	10
SCOPUS	15
LISA	3
LISTA	46
Dialnet	7
CSIC	2
IEEE Xplore	7
TOTAL	83

Tabla 2. Resultados de búsqueda - Bases de datos. Fuente: Elaboración propia

Hemos de señalar que, aunque no obtuvimos un número relevante de resultados, hemos optado por complementarlos con otras fuentes de información, por ejemplo, tales como las Actas de 15º Congreso Internacional sobre la Web Semántica, 2016. Parte II", y la tesis "Publicación y enriquecimiento semántico de datos abiertos en bibliotecas digitales" de Gustavo Candela Romero, por la Universidad de Alicante.

Asimismo, cabe aclarar que no se tomó un periodo de tiempo específico porque la finalidad era obtener un número máximo de documentos relacionados con el tema del estudio y poder analizar la evolución sobre la temática.

Los criterios de exclusión que se aplicaron a los resultados que no se emplearon en el estudio fueron los siguientes:

- Debían recoger las materias Wikidata, Enriquecimiento Semántico, y DBpedia.
- Debían estar en acceso abierto para la consulta de los documentos.

Finalmente, en relación con los criterios de selección aplicados a los resultados recuperados estos fueron que recogían las materias Wikidata, Enriquecimiento Semántica, y DBpedia, y permitían la consulta de los documentos en acceso abierto.

4. EL ENRIQUECIMIENTO SEMÁNTICO DE DATOS EN WIKIDATA

La *World Wide Web*, o simplemente la Web, ha venido convirtiéndose en una herramienta de uso habitual en nuestra comunidad, en comparación con otros medios tan importantes, como la televisión o el teléfono, a los que supera en muchos aspectos.

La Web hoy en día es un medio excepcionalmente flexible y económico para la comunicación, el comercio y los negocios, así como para el acceso a la información y servicios, difusión de la cultura, entre otros muchos. Paralelamente al desarrollo aparatoso de la Web, las tecnologías que la hacen viable han experimentado un rápido crecimiento.

A partir de las primeras tecnologías básicas como el HTML[4] y HTTP[5], hasta nuestros días, han emergido tecnologías como JavaScript[6], PHP[7], XML[8], por mencionar algunas de las más conocidas, que permiten una web mejor, más amplia, más pujante, manejable, o más posible de mantener. Estos cambios influyen y son al tiempo influidos por la transformación de lo que entendemos por la *World Wide Web*.

La creación dinámica, el engranaje como base de datos, la mayor interactividad con el usuario, así como la idea de la Web como plataforma mundial para el desarrollo de aplicaciones, la adaptabilidad al usuario, son algunas de las tecnologías de las tendencias evolutivas más marcadas de los últimos años.

Entre las últimas tendencias que pueden resultad en el futuro de la Web a medio plazo, Castells (2005) expone que a finales de los años 90 surge el enfoque de lo que se ha dado en llamar la "*web semántica*", término establecido por Berners-Lee en 2001. Se puede decir que trata de una corriente, promovida por el propio padre de la Web y presidente del consorcio W3C, cuya finalidad es lograr que las máquinas puedan entender, y, por tanto, aprovechar lo que la Web contiene.

4.1. Web Semántica

Esta Web Semántica, siguiendo con Castells (2005), se propone representar los recursos web con representaciones entendibles no solo por personas, sino por programas que puedan asistir, representar, o sustituir a las personas en tareas frecuentes o inabarcables para un humano.

Se puede decir que las tecnologías de la Web Semántica buscan desarrollar una Web más relacionada, donde sea aún más fácil localizar, compartir e integrar información y servicios, para sacar mejor provecho de los recursos disponibles en la propia web.

[4] HTML o *HyperText Markup Language*. Véase: https://www.w3schools.com/html/default.asp
[5] HTTP o *Hypertext Transfer Protocol*. Véase: https://www.w3schools.com/whatis/whatis_http.asp
[6] JavaScript. Véase: https://www.w3schools.com/js/default.asp
[7] PHP o *Hypertext Preprocessor*. Véase: https://www.w3schools.com/php/DEFAULT.asp
[8] XML o *eXtensible Markup Language*. Véase: https://www.w3schools.com/xml/

Para hacer posible el funcionamiento de la Web Semántica es necesario la inclusión de lenguajes para la representación de ontologías[9], lenguajes de consulta, entornos de desarrollo, entre otros.

A continuación, se recogen aquellos lenguajes de representación más destacables para poder describir la Web Semántica.

4.1.1. XML

XML o *eXtensive Markup Language* es un lenguaje estándar de marcado para el intercambio de datos en la Web. XML no añade semántica al HTML, ya que implica metadatos y ontologías.

Chávez, Cárdenas y Benito (2005) indican que XML es un subconjunto del Lenguaje de Marcado Generalizado Estándar o SGML (por sus siglas en inglés de *Standard Generalized Markup Language*), y lo definen como un formato de texto diseñado para la transmisión de datos estructurados. Por tanto, al ser un subconjunto de SGML mantiene sus características de validación, estructurado y principalmente la extensibilidad, debido a que es un metalenguaje que permite representar lenguajes de marcas, tanto la definición de etiquetas como la relación estructural que existen entre ellas.

Según Chávez, Cárdenas y Benito (2005), el XML permite estructurar datos y documentos en forma de árboles de etiquetas con atributos. Además, con XML Schema, un lenguaje utilizado para describir la estructura y restricciones de contenidos en los documentos XML, se pueden acordar de antemano las distribuciones que se van a emplear, así como manipular tipos de datos primitivos y derivados.

Estos autores explican que, desde la aparición de XML en 1998, se han definido multitud de estándares para organizar información en dominios específicos como el periodismo, la enseñanza, o la medicina, entre otros muchos campos. XML es un primer paso en la dirección de mejorar hacia una representación explícita de los datos y la estructura de los contenidos de la Web, separada de su presentación HTML.

Por último, Castells (2005) añade que XML facilita una sintaxis para hacerlo posible, pero ofrece una capacidad limitada para expresar la semántica. Además, manifiesta que el modelo de datos XML consiste en un árbol que no distingue entre objetos y relaciones, ni tiene noción de jerarquías de clases.

[9] Una ontología es un sistema de representación del conocimiento que resulta de seleccionar un dominio o ámbito del conocimiento, y aplicar sobre él un método con el fin de obtener una representación formal de los conceptos que contiene y de las relaciones que existen entre dichos conceptos.

4.1.2. RDF

RDF o *Resource Description Framework* es un lenguaje para la definición de ontologías y metadatos en la Web. Para Alende (2015) este lenguaje es la piedra fundamental de la Web Semántica.

Alende (2015) explica que este lenguaje fue creado por el W3C y formalmente presentado en 1999 en el documento que especifica el estándar original[10]. Además, añade que el objetivo de RDF es definir un mecanismo para describir recursos que no hagan referencia a ningún dominio en particular, y que puedan luego ser usados para describir información de cualquier dominio.

Chávez, Cárdenas y Benito (2005) aportan que la sintaxis y estructura del RDF es similar al de los lenguajes orientados a objetos: clases y subclases, éstas se disponen en una jerarquía, y las subclases pueden adquirir propiedades de las clases, aunque también se añade la herencia múltiple, que permite la mezcla de diferentes esquemas semánticos.

Estos autores también añaden que el RDF proporciona también reglas para facilitar técnicamente la manera de explicar conceptos de modo que los ordenadores puedan procesarlo de forma rápida y proporciona un medio que posibilita la edición de vocabularios con propiedades definidas para la descripción de los recursos de una comunidad.

Siguiendo con Alende (2015), explica que RDF sigue un modelo. Este modelo es simple y flexible y permite expresar cualquier hecho, pero suficientemente estructurado para que una aplicación de software pueda operar con esta información. Por tanto, esta modelo se compone de tres partes principales:

1. **Sentencia:** cada sentencia representa una de las pequeñas partes de conocimiento en que se descompone la información. Cada sentencia toma la forma de sujeto-predicado-objeto y este orden debe respetarse siempre. El sujeto y el objeto representan dos "cosas" que existen en el mundo, reales o abstractas, denominadas *recursos*, y el predicado es la relación que los une, también denominado *propiedad*. Las sentencias pueden ser representadas por un gráfico RDF, donde el sujeto y el objeto se representan en óvalos o en cuadrados y el predicado son las flechas que los unen. La dirección de la flecha es importante. El acro siempre empieza en el sujeto y apunta hacia el objeto de la sentencia.
2. **Recurso:** el objeto y el sujeto representan algo del mundo real, que puede ser visible o abstracto. Estos recursos son identificados por unos URI (*Universal Resource Identifier*). Cualquier cosa puede tener un URI. La extensibilidad de las URI permite la introducción de identificadores para cualquier entidad

[10] Véase: https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

imaginable. El objeto también puede ser considerado como el valor de la propiedad.

3. **Propiedad:** la propiedad es un aspecto específico, característica, atributo, o relación utilizada para describir un recurso. Cada propiedad tiene un significado específico, define sus valores permitidos, los tipos de recursos que puede describir, sus relaciones con otras propiedades. Al igual que los recursos, también se identifican con las URI.

Por último, gracias a estas tres partes se crea una sentencia RDF muy potente y capaz de describir cualquier situación del mundo real. Asimismo, al ser fácilmente entendible permite desarrollar aplicaciones capaces de ejecutar consultas y obtener resultados sobre la información representada (Figura 6).

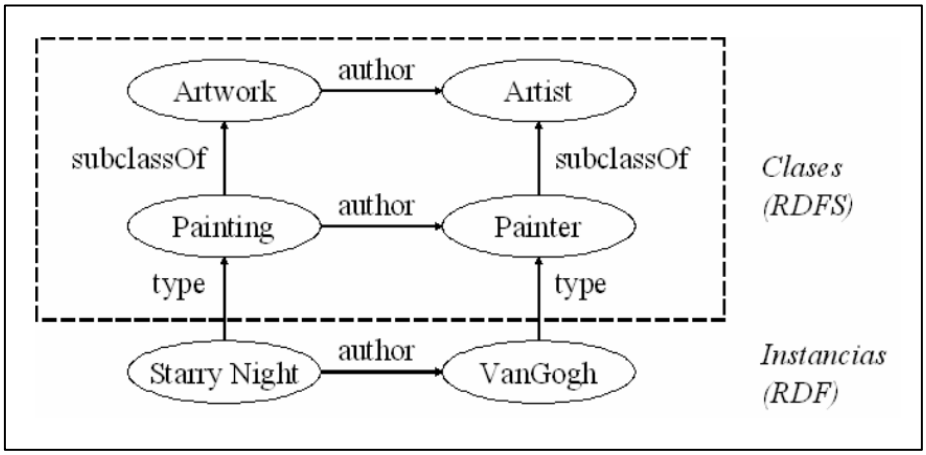

Figura 6. RDF y RDF Schema. Fuente: Imagen obtenida de (Castells, 2005)

4.1.3. RDFS Schema

El RDFS o *RDF Schema*, según Yu (2011), como se cita en Alende (2015), es un lenguaje de representación que desarrolla el conocimiento de RDF, es una extensión semántica de RDF. Puede usarse para crear un vocabulario para la descripción de clases, subclases y propiedades de recursos RDF. Además, añade que por ser un estándar provee construcciones que admiten definir clases y propiedades de un dominio específico. De hecho, estas construcciones son también clases y propiedades que son empleadas para describir las clases y propiedades de un dominio en particular.

Además, García García (2003), como se cita en Alende (2015), añade que estas RDFS no facilitan un vocabulario sobre aplicaciones orientadas a clases, sino que provee de mecanismos para explicar que tales clases y propiedades son parte de un vocabulario y de cómo se espera su relación. Además, permite que las clases puedan organizarse en forma jerárquica.

Por último, siguiendo con Alende (2015), se expone que con estas RDFS se puede definir una ontología sobre un dominio específico. Se definen las clases, la jerarquía a la que pertenecen y sus propiedades, pero n oes suficientemente expresivo, ya que es

necesario reinar el conocimiento embebido para obtener una mejor inferencia de conocimiento por parte de la aplicación que va a interpretar la ontología subyacente.

4.1.4. OWL

El OWL o *Web Ontology Language*, según Chávez, Cárdenas y Benito (2005), añade más vocabulario para describir propiedades y clases, tales como: relaciones entre clases, cardinalidad, igualdad, tipologías de propiedades más complejas, caracterización de propiedades o clases enumeradas.

Además, estos autores indican que OWL es un mecanismo para desarrollar temas o vocabularios específicos en los que asociar estos recursos, así como para definir ontologías estructuradas que pueden usarse a través de varios sistemas.

Alende (2005) añade que OWL fue desarrollado por el Consorcio W3C y que, desde su estandarización, se ha empleado para escribir ontologías en diferentes áreas de conocimiento como medicina, biología, geografía, astronomía, defensa e industrias aeroespaciales. Por último, indica que se ha convertido en el estándar de factor para desarrollo de ontologías e intercambio de datos en la comunidad científica.

4.1.5. SPARQL

El SPARQL o *SPARQL Protocol and RDF Query Language*, según Hidalgo Delgado y Rodríguez Puente (2013), es un lenguaje declarativo de consultas similares a SQL que permite realizar consultar sobre los datos en un grafo RDF. Normalmente, el resultado de las consultas puede ser un conjunto de resultados o grafos RDF.

Alende (2015) añade que el lenguaje de consulta SPARQL se basa en comparación de patrones gráficos. Estos patrones contienen triples, que son como las 3 partes fundamentales de RDF, pero con la opción de una variable consulta en lugar de un término RDF en las posiciones del sujeto, propiedad u objeto. Con ello, combinando los patrones triples obtendremos un parón gráfico básico, donde será necesario realizar una comparación exacta entre gráficos (Figura 7).

Por último, cabe destacar que este lenguaje de consultas es el utilizado en Wikidata, ya que es una base de datos del conocimiento.

Sintaxis básica de una consulta SPARQL	
Prologue (optional)	BASE <iri> PREFIX prefix: <iri> (repeatable)
Query Result forms (required, pick 1)	SELECT (DISTINCT)sequence of ?variable SELECT (DISTINCT)* DESCRIBE sequence of ?variable or <iri> DESCRIBE * CONSTRUCT { graph pattern } ASK
Query Dataset Sources (optional)	Add triples to the background graph (repeatable): FROM <iri> Add a named graph (repeatable): FROM NAMED <iri>
Graph Pattern (optional, required for ASK)	WHERE { graph pattern z}
Query Results Ordering (optional)	ORDER BY ...
Query Results Selection (optional)	LIMIT n, OFFSET m

Figura 7. Sintaxis básica de una consulta SPARQL. Fuente: Imagen obtenida de (Alende, 2015)

4.2. Linked data

La *Linked Data* es una herramienta clave de la Web Semántica porque ofrece una serie de principios para compartir datos vinculados en la Web. Esta herramienta ofrece tecnologías como las ya vistas como RDF, XML, SPARQL, entre otras, permitiendo procesar y compartir información de manera comprensible por las máquinas.

Para entender mejor esta herramienta, debemos entender los cuatro principios básicos de vinculación de datos, que estableció Berners-Lee en 2009. Por tanto, estos principios son los siguientes (Herrera-Cubides, Gaona-García, y Sánchez-Alonso, 2018):

a) Principio n.º 1: Usar URI como nombres de las cosas

Las URI son identificadores de las cosas. Si no se puede identificar una cosa, no se puede hablar de ello. Se usan para nombrar cosas en *Linked Data*. Son una versión generalizada de las URL que se emplean para localizar páginas web a través del navegador.

b) Principio n.º 2: Usar HTTP URI para que la gente pueda buscar esos nombres

Al escribir esa URI en el navegador, esta le informará que no sabe cómo manejar la situación, debido a que no es un tipo de URI que implementa. De ahí que es necesario que las URI sean resolubles en la Web, por lo que se es necesario que se use HTTP URI.

c) Principio n.º 3: Cuando se usa una URI, esta debe ofrecer información relevante

Cualquier HTTP URI se puede escribir en un navegador web y el navegador sabrá qué hacer con ella. Si el servidor remoto responde afirmativamente, devolverá una representación del recurso en diferentes formatos como RDF, entre otros. De cualquier forma, se desearía que las URI puedan resolver unas descripciones útiles acerca de lo que usted ha nombrado.

d) Principio n.º 4: Incluir enlaces a otros URL

Los datos son más útiles si se vinculan datos relacionados, documentos y descripciones. Como se utilizó HTTP URI para publicar sus datos, otras personas pueden vincular sus

datos. La capacidad de seguir estos enlaces le permite a la gente navegar por la web de datos igual que pueden navegar por la web de documentos.

Además de estos principios, Herrera-Cubides, Gaona-García, y Sánchez-Alonso (2018) explican que Berners-Lee sugirió una clasificación de 5 estrellas. En esta clasificación los recursos irán adquiriendo ciertas características con la finalidad de llegar a un recurso plenamente vinculado, de acuerdo con los principios anteriormente citados. Por tanto, esta jerarquía se compone de los siguientes 5 niveles:

1. Publique la información en la Web, en cualquier formato, bajo un tipo de licencia de datos abiertos como, por ejemplo, las Creative Commons.
2. Publique la información en forma de datos estructurados.
3. Use formatos no propietarios.
4. Use las URI para identificar las cosas, de manera que las personas puedan referenciar sus informaciones.
5. Establezca la conexión entre sus datos y otros datos, con el fin de ofrecer un contexto ampliado para las informaciones.

Cabe destacar que estos niveles presentan costos y beneficios incluyendo, por ejemplo, simplicidad de procedimientos, conversión de datos, formación de personal cualificado, entre otros.

Además, estos autores señalan que es probable que la información en la Web puede estar dispersa debido a la gran variedad de fuentes con diferente información, estructura y semántica, y esto pueda ser un problema de heterogeneidad. Por tanto, proponen una solución a través de la interoperabilidad. Esta interoperabilidad será posible haciendo que el modelo de información sea de la siguiente forma:

- **Sintáctica**, que se refiere a la diferencia en el formato de datos, de modo que pueda ser procesada e interpretada en cualquiera de las fuentes de datos.
- **Esquemática o estructural**, que se refiere a las diferencias en el modelo de datos, en los esquemas, razón por la cual debe haber alguna forma de transformación de un esquema a otro.
- **Semántica**, que se refiere a las diferencias en la definición, en el significado que se pretende dar a los términos en contextos específicos.

Por último, autores como Ding y otros (2010), como se cita en Herrera-Cubides, Gaona-García, y Sánchez-Alonso (2018), exponen que para abordar la heterogeneidad y la interoperabilidad es necesario tener en cuenta los siguientes procesos:

- **Conversión:** en primer lugar, los datos brutos se limpian y se conservan a través de la representación basada en RDF. En segundo lugar, este conjunto de datos convertido utiliza URI desreferenciables, de manera que tanto los conjuntos de datos como sus ontologías pueden ser extendidos por terceros usuarios.

- **Mejora:** se centra en extraer la semántica de valores literales en URI significativas, y enlazar conjuntos de datos asociando los URI mencionados en diferentes conjuntos de datos.

Partiendo de estas consideraciones, Wikidata se vale de esta herramienta, la *Linked Data* para poder conectarse con otros recursos y así poder enriquecer sus datos, y permitir que sus informaciones consigan ese buen nivel de calidad de datos.

4.3. El enriquecimiento semántico de datos

El enriquecimiento semántico de datos, según Candela Romero (2019), permite mejorar la interoperabilidad de los datos, así como la solución de los posibles problemas de ambigüedad que se puedan producir en estos. Además, señala que permite la descripción de entidades a través de búsquedas en diferentes conjuntos de datos almacenando la URI que identifica de forma unívoca o contextualiza las entidades.

Antes de adentrarnos a un caso práctico de enriquecimiento de datos en Wikidata, deberíamos explicar cómo se compone una entidad en Wikidata, ya que se observará en la URI del caso práctico y así explicarlo de antemano.

Candela Romero (2019) expone que Wikidata es una base de datos estructurada, abierta y colaborativa que permite almacenar entidades que componen de propiedades que poseen literales o los URI, asociados a elementos de la propia base de datos o a elementos externos. Este autor explica que una entidad en Wikidata se compone de:

- Un identificador único que comienza de acuerdo con el patrón QXXXXX, donde la X son dígitos (su URI se forma uniendo el dominio de Wikidata seguido de su identificador). Por ejemplo, el personaje Harry Potter tiene la URI https://www.wikidata.org/wiki/Q3244512.
- Una o más propiedades con sus valores, por ejemplo, país de residencia, fecha y lugar de nacimiento o lengua materna. Cada propiedad tiene su identificador único, por ejemplo, la propiedad P569 se corresponde a la fecha de nacimiento.
- Uno o más identificadores que enlazan a otros repositorios mediante propiedades explícitamente creadas, por ejemplo, la propiedad P214 se utiliza para identificar personajes en VIAF. Esta propiedad tiene valor 308752753 que se corresponde con su identificador en VIAF.

Además, añade que se puede solicitar a Wikidata la creación de nuevas propiedades. Este proceso se realiza en línea y los administradores decidirán si la propiedad debe crearse. Una vez creada la propiedad, cualquier usuario del sistema puede asociarle valores desde la interfaz. Un grupo de revisores se encargará de analizar esos datos para comprobar si son correctos; este proceso se encuentra descrito en el apartado de *Wikidata*.

4.3.1. Caso práctico: Biblioteca Virtual de Polígrafos

La Biblioteca Virtual de Polígrafos cuenta con un muy bien determinado fichero de autoridades en formato MARC21, que de forma transparente y dinámica puede transformarse según el esquema XML de Europeana Data Model (EDM), y alimentar su repositorio OAI-PMH[11]; o bien, según Thereaux (2006), como se cita en Agenjo-Bullón y Hernández-Carrascal (2018), recuperar la descripción oportuna en EDM RDF a través de la negociación de contenido.

En relación con el proceso de enriquecimiento de los registros de autoridades de polígrafos, Wikidata se ha demostrado como una fuente de datos esenciales por diferentes motivos, pero principalmente por la gran cantidad de datos estructurados, en constante mejora, por su estrecha relación con Wikipedia y otros proyectos Wikimedia y porque a su vez está relacionada con muchos otros vocabularios de valores, externos a Wikipedia, pero propias de archivos, bibliotecas y museos.

Sobre el ejemplo de la descripción de Unamuno en la Biblioteca Virtual de Polígrafos (Figura 8) y los datos disponibles en VIAF, DBpedia, Wikidata y la propia Biblioteca Virtual de Polígrafos se muestra un resumen comparativo, y no exhaustivo, de los datos que más concierten de cada una de las fuentes, pero la adaptaremos sólo a Wikidata y a la propia Biblioteca[12].

Figura 8. Miguel de Unamuno. Fuente: Biblioteca Virtual de Polígrafos

[11] OAI-PMH u *Open Archives Initiative – Protocol for Metadata Harvesting*. Es un protocolo que permite la interoperabilidad para facilitar la difusión eficiente de contenidos en Internet.

[12] Ver tabla original: Agenjo-Bullón, X., y Hernández-Carrascal, F. (2018). Registros de autoridades, enriquecimiento semántico y Wikidata. *Anuario ThinkEPI*, *12*, pág. 366. Recuperado de https://doi.org/10.3145/thinkepi.2018.61

En esta Tabla 3 se han coloreado las filas para mostrar los subprocesos del enriquecimiento semántico en el que cada tipo de dato participa principalmente. Hay que indicar que las filas en blanco se refieren a propiedades que no se han utilizado.

- **Reconciliación de datos** (color naranja claro): el nombre y los nombres alternativos se utilizan para poder crear equivalencias de nombres entre unas fuentes u otras, que es en realidad el significado del término reconciliación. Para ello, muchas fuentes *Linked Open Data* ofrecen también distintos servicios de reconciliación, cada uno con sus singularidades de funcionamiento y estructuración de los datos. Incluso la reconciliación de datos puede efectuarse por medio de la descarga directa de ficheros en formatos estructurados, desde CSV a RDF, lo que amplia notablemente el número de recursos reutilizables dado que muchos conjuntos de datos aún solo se publican en formato CSV o Excel.
- **Vinculación con fuentes *Linked Open Data*** (color azul claro): una vez que se ha establecido la correspondencia de una descripción con otra, y se está seguro de que no se trata de un falso positivo (en cuya elucidación pueden y deben interponerse otras propiedades además del nombre), es posible obtener los URI que identifican esas descripciones e incorporarlos a nuestros datos.
- **Datos adicionales que pueden extraerse** (color verde claro): son todas aquellas aseveraciones que nos puede interesar extraer. Perceptiblemente queda al criterio y conocimiento del catalogador determinar el valor que puede atribuirse a cada fuente, o a cada dato en que particular, y la conveniencia de utilizarlo o no. Es necesario conocer a fondo la estructura de datos de cada fuente y el uso que se hace de esa estructura en las descripciones, ya que pueden coexistir variantes y darse la circunstancia de que instancias de una misma clase pueden tener propiedades distintas para el mismo objetivo descriptivo. Hay que tener en cuenta que al igual que nuestros ficheros de autoridades las fuentes externas *Linked Open Data* también están en constante mejora.

Datos	Wikidata	Biblioteca Virtual de Polígrafos
URI	http://www. wikidata.org/ entity/Q185085	http://www.larramendi.es/poligrafos_y_autores/es/consulta_aut/registro.do?control=POLI20090015128
Nombre	Miguel de Unamuno	Unamuno, Miguel de, 1864-1936
Formas alternativas del nombre	56	25
Instancia de	Ser humano	-
Relaciones	-	Influido por Influye en Relacionado con (5XX)

Tabla 3. Descripción de Unamuno en la Biblioteca Virtual de Polígrafos. Datos disponibles en Wikidata y la propia Biblioteca de Polígrafos. Fuente: Adaptación propia a partir de (Agenjo-Bullón y Hernández-Carrascal, 2018)

Datos	Wikidata	Biblioteca Virtual de Polígrafos
Coautores	-	-
Género	Hombre	Hombre
País de ciudadanía / país relacionado	España	España
Fecha de nacimiento	29 septiembre de 1864	1864
Lugar de nacimiento	Bilbao	Bilbao
Fecha de defunción	31 diciembre 1936	1936
Idiomas utilizados	España y vasco	Español
Escuela filosófica	-	-
Campo de actividad	-	Filosofía, Filosofía, Poesía, Novela, Teatro
Ocupación	Poeta, filósofo, escritor, ensayista, novelista, crítico literario, profesor universitario, dramaturgo	Profesores universitarios, Filólogos, Filólogos, Poetas, Novelistas, Ensayistas, Dramaturgos, Rectores universitarios
Miembro de / Instituciones relacionadas	Real Academia Española, Sociedad de Amigos de Portugal	Instituto Vizcaino (1875-1880), Universidad de Madrid (1880-1994), Universidad de Salamanca (1891-1934)
Formación	Universidad Central	-
Biografía	Enlaces a las wikipedias	Biografía de elaboración propia y resumen extraído de Wikipedia (a través de DBpedia)
Identificadores externos vinculados	VIAF, ISNI, LC, GND, BNF, BNE, CANTIC… y 44 identificadores más	VIAF, ISNI, LC, GND, BNF, BNE, Wikidata, DBpedia y 9 identificadores más. Enlaces
Enlaces externos	52 enlaces a Wikipedias en distintos idiomas, 23 a Wikiquote, 3 a Wikisource	Wikipedia en español, English Wikipedia, Wikipédia em portugués, Viquipèdia en català, Euskarazko Wikipedia, Galipedia… 4 enlaces más, entre ellos WorldCat, Identities

Tabla 4 (cont.). Descripción de Unamuno en la Biblioteca Virtual de Polígrafos. Datos disponibles en Wikidata y la propia Biblioteca de Polígrafos. Fuente: Adaptación propia a partir de (Agenjo-Bullón y Hernández-Carrascal, 2018)

Por último, podemos decir que las técnicas de enriquecimiento semántico se favorecen de los procesos de reconciliación con diferentes vocabularios de valores, entre las que destacan ya no solo las fuentes bibliotecarias, sino también no bibliotecarias como Wikidata.

Además, su estrecha relación con las bibliotecas, los archivos y los museos le permite ser una fuente de información aportando datos, referencias y contenidos digitales de libre acceso.

Wikidata se está convirtiendo en un recurso autorizado a nivel mundial dentro de la Web Semántica.

5. WIKIDATA Y SU RELACIÓN CON LA DBPEDIA

En los últimos años, se han ido creando diversas bases de datos de conocimiento, siendo las más conocidas por su popularidad la DBpedia y Wikidata.

Lehmann y otros (2012) exponen que DBpedia ofrece una visión completa y actual de las representaciones de entidades extraídas de Wikipedia, mientras que Wikidata ofrece una variedad de sentencias de otras fuentes. Mencionan que una de las fuentes más ricas de DBpedia son las infoboxes de Wikipedia, pero que, aunque están estructurados, no son homogéneos y no están normalizados. He aquí que Wikidata tiene como objetivo rellenar automáticamente los infoboxes gracias a que es una base de datos de alta calidad que suministra datos a todos los proyectos de la Fundación Wikimedia.

Ismayilov y otros (2015) aportan que DBpedia extrae información de más de cien ediciones en idiomas de Wikipedia y de Wikimedia Commons, mientras que Wikidata es una base de conocimiento abierto y colaborativo que facilita la construcción de conocimiento estructurado para su empleo en otros proyectos Wikimedia. Por tanto, se puede decir que estos proyectos de la Fundación Wikimedia se complementan mutuamente.

Por último, cabe destacar que Wikidata proporciona información estructurada para enriquecer DBpedia y así proporcionar un valor añadido para una serie de escenarios de uso.

5.1. Los infoboxes

Los infoboxes son una herramienta importante mediante la cual se relacionan Wikidata y DBpedia. Esto se debe a que la primera suministra datos, mientras que la segunda permite enlazarlos a través *de Linked Open Data* y enriquecer a Wikipedia.

Un infobox (Wikipedia, 2020) es una tabla, con un formato prediseñado que se coloca en la parte superior derecha de los artículos de Wikipedia. Tiene como objetivo mostrar un resumen actualizado y precisos de algunos aspectos comunes que comparten los artículos, así como mejorar la navegación hacia otros artículos relacionados.

Según Alende (2015), el uso de infoboxes en los artículos no es un requisito obligatorio, aunque su utilidad permite tener un vistazo rápido de los artículos. Para incluir un infobox y saber qué partes de este hay que emplear es necesario consultar el consenso[13] determinado entre los editores de cada artículo.

Siguiendo con Wikipedia (2020), la plantilla infobox contiene datos importantes y estadísticas sobre artículos que tiene un tema en común. Por ejemplo, todas las novelas fantásticas tiene una clasificación por género, subgénero, tema, etc. El agregar un

[13] Véase: https://en.wikipedia.org/wiki/Wikipedia:Consensus

{{taxobox}}[14] en artículos relacionados con novelas fantásticas hace más fácil encontrar rápidamente dicha información y compararla con otros artículos (Figura 9). Se convierte, por tanto, en un elemento de navegación adicional a los enlaces interwikis y a la utilización de categorías.

Hay que destacar que los infoboxes no son tablas estadísticas, sino que resumen el contenido del artículo, la cual debe presentarse, para cada artículo, en el texto principal, ya que puede no ser posible para algunos lectores acceder al contenido del infobox.

Alende (2015) añade que la información en el infobox debería ser:

- **Comparable:** si un atributo común se comparte con muchos temas diferentes, es conveniente compara estos datos entre diferentes páginas. Esto implicará que el contenido deberá ser presentado en un formato estándar.
- **Concisa:** los infoboxes se deberán comprender a simple vista, para poder observar la información de manera.
- **Relevancia:** los infoboxes deberán contener información relevante.
- **Citada en alguna parte del artículo previamente:** los infoboxes deben contener fundamentalmente datos que puedan ser ampliados con referencias de fuentes fiables dentro del mismo artículo.

Así como se indica lo que debe ser, también se debe indicar lo que no debe ser la información en el infobox:

- **Extensa:** los textos largos o las estadísticas demasiados detalladas.
- **Detalles triviales:** se ha observado que un problema repetitivo es la inclusión de información trivial, haciendo que el artículo contenga datos no relevantes.
- **Banderas:** los íconos de banderas no deberían usarse en los infoboxes, ya que generan distracción innecesaria.

Autores como Yus y otros (2014), como se cita en Sáez y Hogan (2018), proponen mejorar estos infoboxes con el uso del sistema *Infoboxer*, a través de DBpedia, ya que permitiría ayudar a los usuarios a crear infoboxes sugiriéndoles los atributos más comunes y así validar el rango de estos.

Además, otro autor como Kaffee (2016), como se cita en Sáez y Hogan (2018), en un estudio posterior propone un método un método para generar de forma automática marcadores de posición para artículos de Wikipedia basados en sentencia de Wikidata. Aquí pretende que el enfoque se dirija hacia los administradores de Wikipedia, donde estos deberán generar un orden apropiado de atributos para su visualización.

Como se observa, la utilización de infoboxes varía su aplicación de una base de datos a otra según sean las funcionalidades que le sea conveniente para cada autor. Aunque Wikidata haya sido lanzada posteriormente a DBpedia, hay que destacar que gracias a

[14] Véase: https://en.wikipedia.org/wiki/Template:Taxobox

los datos estructurados que almacena ha sido piedra fundamental para Wikipedia, ya que ha permitido reforzar la calidad de datos que presenta Wikipedia, así como su enriquecimiento semántico a través de la ayuda de DBpedia con el uso del *Linked Open Data*.

Figura 9. Infobox de Harry Potter (novela fantástica). Fuente: Wikipedia

5.2. Diferencias entre DBpedia y Wikidata

A pesar de relacionarse entre sí estas dos bases de datos, estas tienen sus diferencias y se ha querido hablar de ellas tomando como referencia las aportaciones de Ismayilov y otros (2015), y Saorín y Pastor-Sánchez (2018), que se muestran a continuación:

En relación con los **identificadores (URI, IRI[15])**:

- Wikidata emplea identificadores numéricos independientes del idioma.
- DBpedia emplea identificadores legibles por humanos a partir del título de los artículos en cada idioma.

En relación con la **estructura**:

- Wikidata desarrolla su propio modelo de datos, con mayor capacidad para capturar la procedencia de la información, y permite generar diferentes serializaciones en RDF.
- DBpedia usa RD de forma nativa en su modelo de datos.

En relación con el **esquema**:

- Wikidata evita el uso directo de términos de RDF u OWL, y redefine la mayoría de ellos. Además, define una propiedad local que es similar a *rdf:type*. Existen intentos de conectar las propiedades de Wikidata con RDFS/OWL. Proporciona exportaciones alternativas de datos.
- DBpedia se basa en OWL para organizar los datos que extrae e integra las diferentes ediciones de Wikipedia.

En relación con la **curación**:

- Wikidata posee su propio entorno de edición de contenidos, WikiBase, que permite crear, editar, y depurar tanto sus datos como su estructura.
- DBpedia extrae sus datos de forma automática de la Wikipedia, y, por tanto, es un conjunto de datos de sólo lectura. Se puede decir que los editores de Wikipedia son los que mantienen esta base de datos, pero por su naturaleza semiestructurada no se pueden capturar todos los datos, y esto puede generar errores durante la extracción.

En relación con la **curación**:

- Wikidata posee su propio entorno de edición de contenidos, WikiBase, que permite crear, editar, y depurar tanto sus datos como su estructura.
- DBpedia extrae sus datos de forma automática de la Wikipedia, y, por tanto, es un conjunto de datos de sólo lectura. Se puede decir que los editores de Wikipedia son los que mantienen esta base de datos, pero por su naturaleza semiestructurada no se pueden capturar todos los datos, y esto puede generar errores durante la extracción.

[15] IRI o *Identificador de Recursos Internacionales*. Véase: https://www.w3.org/International/O-URL-and-ident.html

En relación con la **publicación**:

- Wikidata se ofrece como un potente entorno de visualización de resultados, que permite obtener mapas, diagramas o gráficos a partir de los datos.
- DBpedia, al igual que Wikidata, publican mediante técnicas basadas en *Linked Open Data*, incluyendo *datasets dumps*, URI derreferenciales y SPARQL *endpoints*.

En relación con la **cobertura**:

- Wikidata crea identificadores comunes para conceptos que existen en más de un idioma. Hay que indicar que no todos los artículos, categorías, plantillas y redirects de una edición tiene un elemento Wikidata.
- DBpedia proporciona identificadores para todos los elementos estructurados de una edición de Wikipedia. Además, incluye artículos, categorías, redirecciones y plantillas.

En relación con la **actualización de los datos**:

- Wikidata ofrece un entorno editable, y permite que los editores puedan crear, actualizar, y corregir datos sobre la marcha.
- DBpedia es un conjunto de datos estáticos y de sólo lectura que se actualiza de forma periódica. Emplea DBpedia Live para el proceso de copias locales de ediciones de Wikipedia (disponible en inglés, francés y alemán).

De estas aportaciones, cabe destacar la expuesta por Saorín y Pastor-Sánchez (2018) respecto al modelo de datos abstractos de Wikidata, ya que explican que es complejo y que suele incluir dos tipos de entidades: ítems y propiedades.

Se puede apreciar que las dos bases de datos poseen identificadores IRI únicos, por lo que tanto la definición de datos en sí, como las propiedades que las representan, y se encuentran autocontenidas en Wikidata. Por tanto, para cada propiedad de una entidad, pueden mostrarse diferentes valores, señalar su prioridad, añadir cualificadores, contener aproximaciones y asignar la procedencia de la información.

Saorín y Pastor-Sánchez (2018) señalan que Wikidata proporciona una infraestructura local que permite mejorar los otros proyectos de la Fundación Wikimedia. Además, gracias a su publicación con licencia Creative Commons 0[16] y su forma de publicación empleando *Linked Open Data* facilita su uso por terceros, aumentando así el valor principal de estos proyectos a través de la reutilización máxima de datos abiertos.

[16] Esta licencia que permite al autor de una obra original renunciar a los derechos de autor de esta y pasar así a formar parte de dominio público, por lo que cualquiera podremos hacer uso de ella libremente sin tener que pedir permiso y además con fines comerciales si deseamos. Véase: https://creativecommons.org/licenses/?lang=es

Estos autores indican que DBpedia es una infraestructura abstracta de datos abiertos por su diseño, encaminada principalmente a la reutilización y consumo de datos mediante una ontología.

Para Alende (2015), la ontología de DBpedia tiene como principal característica que es superficial y extendida sobre múltiples dominios, y fue diseñada empleando la información de los infoboxes más utilizados en Wikipedia. Esta ontología cubre más de 685 clases que forman una jerarquía en donde conceptos dentro de otros, y sobre ellas estás descriptas más de 2795 propiedades diferentes.

Por último, Ríos Álvarez (2014) añade que DBpedia se erige como un recurso fundamental dentro de la nube de *Linked Data*, debido a que funciona con un concentrador de prácticamente cualquier aplicación de datos enlazados.

Gracias a que DBpedia permite obtener información estructurada procedente de Wikipedia, ha permitido que las clasificaciones en Wikidata sean más fáciles, ya que DBpedia es un estándar de facto para tareas de clasificación.

Con DBpedia, Wikidata permite concentrar más sus datos y aportarles más enriquecimiento, ya que DBpedia permite entrelazarse con otros conjuntos de datos permitiendo así facilitar el trabajo de recopilación de información a Wikidata (Figura 10).

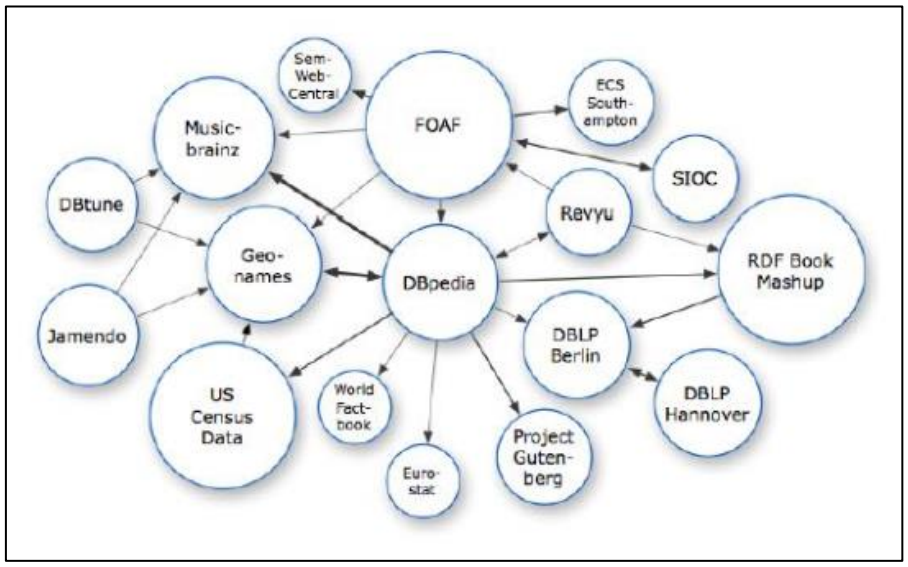

Figura 10. Conjuntos de datos entrelazados con DBpedia. Fuente: Imagen obtenida de (Ríos Álvarez, 2014)

6. RECOMENDACIONES

El proyecto Wikidata es una herramienta novedosa y fundamental para la Fundación Wikimedia, ya que es una central de datos de donde se benefician los diferentes proyectos de esta.

Hemos considerado aportar algunas recomendaciones para aumentar y mejorar su funcionamiento y añadir unas herramientas externas complementarias que podrían maximizar el potencial de esta base de datos.

Estas recomendaciones son una mezcla de nuestras aportaciones propias derivadas de la revisión bibliográfica y de aquellas tomadas de la bibliografía recuperada y analizada.

6.1. Recomendaciones en base a la calidad

Se ha observado que la calidad de Wikidata se ve afectado en función de la contribución que los editores hacen en los artículos, la procedencia de estos, las restricciones, y el enriquecimiento semántico. Por tanto, las recomendaciones propuestas en base a la calidad son las siguientes:

a) **Los artículos:** Se puede decir que los artículos son los encargados de representar las entidades en el mundo real y considerados como nociones visiblemente definidos por los editores. Se ha observado que estos artículos no suelen cumplir con el orden adecuado, debido a que las etiquetas legibles para el hombre, como son las sentencias, no están claramente definidas, y esto hace que los artículos no se agrupen dentro de una clasificación normalizada. Por tanto, se ha seguido una solución alternativa de autores como Yapinus, Sarabadani, y Halfake (2017), como se cita en Piscopo y Simperl (2019), donde establecen una escala de clasificación unívoca basada en añadir etiquetas a los elementos que van, por ejemplo, de la A la E. Dentro de esta escala se deberá consignar la integridad de un elemento, en el que deberá ser descrita como el número de sentencias pertinentes; el número de recursos utilizados para argumentar estas sentencias; un número adecuado de idiomas paras las etiquetas y descripciones; enlaces a otros proyectos de la Wikimedia; y, por último, si es posible, incluir material multimedia.

b) **La procedencia:** Se puede decir que la procedencia de la información es una de las características fundamentales que distingue a la Wikidata de otros proyectos parecidos. Esta procedencia indicada en esta base de datos permite detectar errores en los datos que almacenamos y así mejorarlos. Por tanto, se ha observado que existen carencias respecto a la procedencia de la información que se emplea en los artículos, por lo que esto impide la reutilización de los datos y la posibilidad de mejorarlos. En un estudio por Piscopo y Simperl (2019) se recoge que esto se podría solucionar si se mejorase la política de verificación de

la procedencia de datos de Wikidata. Se ha considerado que esto se podría solucionar si se realizara un control para confirmar la fuente de origen de los datos, por parte de los colaboradores de Wikidata, y asignarles un URI de estas fuentes para poder cotejar la información almacenada en la base de datos. Por último, esto permitiría que los datos se considerasen más fiables y actualizados.

c) **Las restricciones:** Se ha observado que estas restricciones permiten establecer cómo deben usarse las relaciones y propiedades cuando se añade contenido a la base de datos. No obstante, se ha comprobado que estas restricciones no se aplicaban, ya que el propio sistema no comprobaba ni registraba el acceso de los editores, por lo que podían acceder a la base de datos y añadir contenido violando estas restricciones. Por tanto, esto se podría mejorar si hubiese un control más estricto del propio sistema, como, por ejemplo, permitiendo un registro de la IP para controlar quién entra al sistema y así poder detectar posibles vandalismos y buscar los responsables.

d) **Enriquecimiento semántico:** Se ha observado que Wikidata se caracteriza por la reutilización de datos a través del empleo de los URI por lo que eso hace que sus datos se enriquezcan y actualicen cada vez más. No obstante, se ha observado que sólo se enriquece de datos a través de bases de datos y repositorios internacionales y nacionales como, por ejemplo, VIAF[17], o la Biblioteca Nacional de España, entre otros. Por lo que se ha considerado que Wikidata también emplee repositorios institucionales españoles y no sólo de la Biblioteca Nacional de España, por ejemplo, como es el caso de los repositorios universitarios, ya que esto permitiría enriquecer los datos gracias a que emplean el sistema de acceso abierto.

6.2. Recomendaciones en base a herramientas externas

Se ha recogido la aportación de Proffitt (2018) que nos muestra algunas herramientas útiles en relación con el uso potencial de los datos de Wikidata en bibliotecas, siendo las siguientes:

a) ***Inventarie*[18]:** Se trata de una herramienta que permite generar un inventario de todos los documentos de los usuarios de una biblioteca. Se basa en Wikidata y aprovecha los enlaces de Wikipedia disponibles en los elementos de la base de datos para construir biografías de autores de estos documentos, y se visualizan a través de una presentación dinámica y llamativa. Podemos decir que usa código abierto y permite la conectividad con la comunidad. Además, permite el acceso y la reutilización de datos de conocimiento abierto.

[17] VIAF o *Fichero de Autoridades Virtual Internacional* es un proyecto conjunto de varias bibliotecas nacionales, implementado y alojado por OCLC. Véase: http://viaf.org/
[18] Inventarie. Véase: https://inventaire.io/welcome

b) ***Reasonator***[19]**:** Se trata de una herramienta que recoge los datos disponibles en Wikidata relacionados con un elemento e incluye imágenes, audios y mapas. Esta herramienta permite contribuir datos a Wikidata permitiendo su enriquecimiento semántico y con ello facilitar la visibilidad de las colecciones de bibliotecas que emplean esta herramienta.

c) ***Scholia***[20]**:** Se trata de una herramienta alojada en Wikimedia Tools Labs, y es de código abierto disponible en GitHub (plataforma de desarrollo colaborativo). Esta herramienta se vale de la riqueza de los datos en Wikidata para la creación y visualización de perfiles de docentes reconocidos. Estos perfiles se organizan empleando el nombre, puesto, organización, lugar, colección, publicaciones, y tema. Gracias a estos datos el personal bibliotecario puede crear estos perfiles con los datos que recoge de Wikidata, por ejemplo, usando la información personal y publicaciones que se almacenan en ella.

d) ***Histopedia***[21]***:*** Se trata de una herramienta que se visualiza como un sitio web que usa una combinación de datos de Wikidata y Wikipedia para generar líneas de tiempo sobre la Historia de manera dinámica, y que las instituciones con fines educativos se benefician de ella.

Además de estas herramientas, hemos añadido la aportación de Zangerle y Müller-Birn (2018), de la que hemos considerado la importancia de la herramienta ***Neonion***, que proporciona un entorno para la anotación semántica de recursos textuales.

Esta herramienta presenta como núcleo una interfaz de usuario basada en un navegador intuitivo en la que se añaden de forma manual las anotaciones semánticas en un triple formato a los textos. *Neonion* presenta un modelo de conocimiento terminológico flexible basado en metadatos contextuales y en el modelo para cada anotación, como el propietario de la marca de tiempo.

Neonion se complementa con ***Snoopy*** (Gassler, 2017), un sistema de recomendación que permite a los usuarios introducir y editar información en sistemas de información semiestructurados, como es el caso de Wikidata.

Gracias a estas recomendaciones los usuarios sólo necesitan tener nociones básicas de terminología de Wikidata para sus primeras anotaciones. Estas anotaciones se guardan en *Neonion* a través de un formato basado en RDF y, posteriormente, son subidas a Wikidata con una referencia que conecta con el origen de los datos, por ejemplo, el URL de un recurso en línea o el DOI[22] de un artículo de investigación.

[19] Reasonator. Véase: https://reasonator.toolforge.org/
[20] Scholia. Véase: https://tools.wmflabs.org/scholia/
[21] Histropedia. Véase: http://histropedia.com/
[22] DOI o *Digital Object Identifier* es una forma de identificar un objeto digital como, por ejemplo, un artículo de revista electrónica, un capítulo de libro electrónico, etc. Véase: https://www.doi.org/

Con esta herramienta permite aumentar la calidad y la exhaustividad de los datos de Wikidata.

Por último, hemos considerado añadir una aportación más con la herramienta ***Mix'n'match***[23], herramienta desarrollada por Magnus Manksee que recibió el WikidataCon Award en 2019.

Esta herramienta, según Agenjo-Bullón y Hernández Carrascal (2020), permite subir catálogos de datos para encontrar automáticamente la entrada correspondiente de Wikidata. Hay que indicar que en algunos casos la correspondencia es directa, mientras que en otros es necesario que un usuario confirme una correspondencia determinada entre las posibles que la aplicación detecta. Para ello, dispone de dos modos: el modo manual y el modo semi-automático.

Además, permite crear una vinculación masiva de ficheros de autoridad de archivos, bibliotecas y museos con Wikidata.

[23] Véase: https://mix-n-match.toolforge.org/#/

7. CONCLUSIONES

Wikidata se ha convertido en un proyecto que sigue en constante auge y evolución que ha llamado la atención de profesionales e investigadores, ya que se ha observado que existen multitud de estudios que impulsan a la investigación sobre esta base de datos (Farda-Sarbas y Müller-Birn, 2019; Candela Romero, 2019; Agenjo-Bullón y Hernández-Carrascal, 2018; Sáez y Hogan, 2018; y Saorín y Pastor-Sánchez, 2018).

En relación con la aplicación de bots en Wikidata, nos ha permitido observar que para las ediciones de la información no existe la necesidad de la toma de decisiones humanas, aunque hemos de decir que a la hora de asignar las tareas de estos bots y de que estos cumplen correctamente sus tareas, las decisiones de los administradores tienen un papel importante. De hecho, los bots se encargan de realizar acciones tediosas y repetitivas consistentes en la corrección de algunos errores evidentes que se detectan en los artículos, siempre aplicando políticas de edición muy estrictas.

En relación con la Web Semántica y el enriquecimiento de datos, hemos podido comprobar que gracias a los lenguajes para la representación de esta Web es posible localizar, compartir e integrar información.

Estos lenguajes son una herramienta fundamental para que estos datos puedan ser comprendidos por las máquinas y así facilitar su trabajo, así como la gestión de estos datos.

Además, se ha observado que Wikidata emplea los URI (a través de la vinculación con fuentes *Linked Open Data*) para poder enlazarse con otras bases de datos y repositorios para enriquecer sus datos, ello permite que los datos sean más fiables, y legibles, consiguiendo que esta base de datos se convierta en un recurso autorizado a nivel mundial dentro de la Web Semántica.

Este sistema de *Linked Open Data* ha permitido que Wikidata pueda beneficiarse de conjuntos de datos a través de su relación con la DBpedia, permitiéndole enriquecer sus datos y obtener un mayor número de fuentes adicionales para corroborar la información que almacena.

No obstante, a nivel profesional se echó de menos que Wikidata no usara repositorios institucionales académicos a nivel nacional, por lo que se consideró que ello podría ser muy útil para su sistema de datos, debido a que estas instituciones disponen de acceso abierto, por lo que sus datos podrían beneficiarse de esta herramienta.

En relación con su vinculación a la DBpedia, hemos podido comprobar que ambas se complementan entre sí. Ello se debe a que Wikidata proporciona información estructurada para enriquecer a DBpedia, permitiendo añadir un valor importante para una serie de escenarios de uso. Esta relación se plasma a través de la información que

se visualiza en los infboxes de Wikipedia, donde permite comprobar el trabajo mutuo de estos dos proyectos en evolución.

Por último, a nivel personal este trabajo me ha permitido conocer el funcionamiento de Wikidata como base de datos y su aportación a la Sociedad del Conocimiento. A pesar de ser una base de datos en continuas mejoras tiene gran relevancia como recurso rico de datos estructurados lo que la diferencia de otras.

Aunque lleva sus pocos años, Wikidata ha demostrado ser una piedra fundamental para todos los proyectos de la Fundación Wikimedia, y estudiarla ha sido gratificante, permitiéndome conocer una herramienta que cada vez es más conocida y utilizada en el mundo de la investigación.

8. BIBLIOGRAFÍA Y FUENTES DE INFORMACIÓN

Agenjo-Bullón, X., & Hernández-Carrascal, F. (2018). Registros de autoridades, enriquecimiento semántico y Wikidata. *Anuario ThinkEPI*, *12*, 361-372. Recuperado de https://doi.org/10.3145/thinkepi.2018.61

Agenjo-Bullón, X., & Hernández-Carrascal, F. (2020). Wikipedia, Wikidata y Mix'n'match. Recuperado de https://listserv.rediris.es/cgi-bin/wa?A2=ind2002d&L=IWETEL&P=78#TOP

Alende, A. N. (2015). *Clasificación de las recomendaciones obtenidas del BueFinder para la propiedad semántica birthPlace*. Trabajo Fin de Grado. La Plata, Argentina: Universidad Nacional de La Plata. Recuperado de http://sedici.unlp.edu.ar/bitstream/handle/10915/47705/Documento_completo.pdf?sequence=3&isAllowed=y

Brasileiro, F., Almeida, J.P.A., Carvalho, V.A., & Guizzardi, G. (2016). Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata. In: *Proceedings of the 25th International Conference Companion on World Wide Web (WWW 2016)*. Switzerland: International World Wide Web Conferences Steering Committee, 975–980. Recuperado de https://dl.acm.org/doi/10.1145/2872518.2891117

Candela Romero, G. (2019). *Publicación y enriquecimiento semántico de datos abiertos en bibliotecas digitales*. Tesis doctoral. Alicante: Universidad de Alicante, Departamento de Lenguajes y Sistemas Informáticos. Recuperado de https://rua.ua.es/dspace/bitstream/10045/97353/1/tesis_gustavo_candela_romero.pdf

Castells, P. (2005). La web semántica. En: Bravo Santos, C., & Redondo Duque, M. (2005). *Sistemas interactivos y colaborativos en la web*. Cuenca: Ediciones de la Universidad de Castilla-La Mancha, 195-212. ISBN: 978-8-484-27352-3. Recuperado de http://cic.puj.edu.co/wiki/lib/exe/fetch.php?media=materias:castells-uclm03.pdf

Chávez, M. E., Cárdenas, O., & Benito, O. (2005). La web semántica. *Revista de investigación de Sistemas e Informática*, *2*(3), 43-54. Recuperado de http://sisbib.unmsm.edu.pe/bibvirtualdata/publicaciones/risi/n3_2005/a06.pdf

Ding, L., DiFranzo, D., Graves, A., Michaelis, J., Li, X., McGuinness, D. L., & Hendler, J. A. (2010). Data-gov Wiki: towards Linking Government Data. In: *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*. Califormia: The AAAI press. Recuperado de https://www.semanticscholar.org/paper/Data-gov-Wiki%3A-

Towards-Linking-Government-Data-Ding-DiFranzo/22a9b850e3aa6eaa67744b65fd9ea135e4eae3ce

Färber, M., Ell, B., Menne, C., Rettinger, A., & Bartscherer, F. (2016). Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web, 1*, 1-5. Recuperado de http://www.semantic-web-journal.net/system/files/swj1366.pdf

Farda-Sarbas, M., & Müller-Birn, C. (2019). Wikidata from a Research Perspective. A Systematic Mapping Study of Wikidata. Recuperado de https://arxiv.org/abs/1908.11153

García García, G. (2003). *RDF y RDF schema*. Universidad Nacional Autónoma de México. Recuperado de http://www.matem.unam.mx/~grecia/semantic_web/rdf.html

Gassler, W. (2017). *The SnoopyConcept: Leveraging Recommendations for Knowledge Curation*. *PhD thesis. Austria: University of Innsbruck, Department of Computer Science. Recuperado de* https://dbis.uibk.ac.at/sites/default/files/2018-06/gassler_diss_20170811.pdf

Haythornthwaite, C. (2009). Crowds and communities: Light and Heavyweight Models of Peer Production. In: *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences (HICSS 2009)*. United States: HICSS, 1-10. Recuperado de https://ieeexplore.ieee.org/document/4755627

Herrera-Cubides, J. F., Gaona-García, P., & Sánchez-Alonso, S. (2018). Linked Data: qué sucede con la heterogeneidad y la interoperabilidad. *Scientia et Technica, 23*(2), 230-240. Disponible en https://dialnet.unirioja.es/servlet/articulo?codigo=6643339

Hidalgo Delgado, Y., & Rodríguez Puente, R. (2013). La web semántica: una breve revisión. *Revista Cubana de Ciencias Informáticas, 7*(1), 76-85. Recuperado de http://scielo.sld.cu/pdf/rcci/v7n1/rcci09113.pdf

Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., & Hellmann, S. (2015). Wikidata Through the Eyes of DBpedia. *Semantic Web, 9*(4), 493-503. Recuperado de http://www.semantic-web-journal.net/system/files/swj1518.pdf

Kaffee, L. A. (2016). Generating Article Placeholders from Wikidata for Wikipedia: Increasing Access to Free and Open Knowledge. Bachelor Thesis. HTW Berlin University of Applied Sciences. In: Sáez, T., & Hogan, A. (2018). Automatically Generating Wikipedia Info-boxes from Wikidata. *The Semantic Web*, 1823-1830. Recuperado de http://aidanhogan.com/docs/infobox-wikidata.pdf

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., & Bizer, C. (2012). DBpedia - A Large-Scale,

Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal, 6*(2), 1-29. Recuperado de http://www.semantic-web-journal.net/system/files/swj499.pdf

Piscopo, A. (2018). Wikidata: A New Paradigm of Human-Bot Collaboration? Recuperado de https://arxiv.org/pdf/1810.00931.pdf

Piscopo, A., & Simperl, E. (2019). What We Talk About When We Talk About Wikidata Quality: A Literature Survery. In: *Proceedings of the 15th International Symposium on Open Collaboration (OpenSym 2019)*. New York: Association for Computing Machinery, 1-11. Recuperado de https://dl.acm.org/doi/10.1145/3306446.3340822

Piscopo, A., Phethean, C., & Simperl, E. (2017). What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *Proceedings og the 9th International Conference on Social Informatics* (SocInfo *2017), Part I.* Oxford: Springer, 305-322. Recuperado de https://link.springer.com/chapter/10.1007%2F978-3-319-67217-5_19

Proffitt, M. (2018). *Leveraging Wikipedia: connecting communities of knowledge*. Chicago, United States: American Library Association, 256. ISBN: 978-0-838-91732-9.

Ríos Álvarez, J. (2014). *Estudio y propuesta para enriquecimiento de información utilizando fuentes Open Data: una experiencia con videos educativos multidominio*. Trabajo de Fin de Máster. Madrid: Universidad Nacional de Educación a Distancia, Escuela Técnica Superior de Ingeniería Informática. Recuperado de http://e-spacio.uned.es/fez/eserv/bibliuned:master-ETSIInformatica-LSI-Jrio/RioAlvarez_TFM.pdf

Sáez, T., & Hogan, A. (2018). Automatically Generating Wikipedia Info-boxes from Wikidata. *The Semantic Web*, 1823-1830. Recuperado de http://aidanhogan.com/docs/infobox-wikidata.pdf

Saorín, T., & Pastor-Sánchez. J. A. (2018). Wikidata y DBpedia: viaje al centro de la web de datos. *Anuario ThinkEPI*, 12, 207-214. Recuperado de https://doi.org/10.3145/thinkepi.2018.31

Staab, S., & Studer, R. (2009). *Handbook on ontologies*. Berlin: Springer Science & Business Media. e-ISBN: 978-3-540-92673-3.

Thakkar, H., Endris, K.M., Garica, J.M., Debattista, J., Lange, C., & Auer, S. (2016). Are Linked Datasets Fit for Open-Domain Question Answering? A Quality Assessment. In: *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS16)*. United States: Association for

Computing Machinery, 1-12. Recuperado de https://dl.acm.org/doi/10.1145/2912845.2912857

Thereaux, O. (2006). Content negotiation: why it is useful, and how to make it work. W3C Blog. [Mensaje en un blog]. Recuperado de https://www.w3.org/blog/2006/02/content-negotiation/

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM, 57*(10), 78-85. Recuperado de https://dl.acm.org/doi/10.1145/2629489

Wikipedia. (2020). *Ayuda: Infobox*. Recuperado de https://en.wikipedia.org/wiki/Help:Infobox

Yapinus, G., Sarabadani, A., & Halfaker, A. (2017). Wikidata Item Quality Labels. In: *Proceedings of the 15th International Symposium on Open Collaboration (OpenSym 2019)*. New York: Association for Computing Machinery, 1-11. Recuperado de https://figshare.com/articles/Wikidata_item_quality_labels/5035796

Yu, L. (2011). *A developer's Guide to the Semantic Web*. Berlin: Springer. ISBN: 978-3-642-15969-5. Recuperado de https://doi.org/10.1007/978-3-642-15970-1

Yus, R., Mulwad, V., Finin, V., & Mena, E. (2014). Infoboxer: Using Statistical and Semantic Knowledge to Help Create Wikipedia Infoboxes. In: *International Semantic Web Conference (ISWC 2014)*. Switzerland: Springer, 405–408. Recuperado de http://ceur-ws.org/Vol-1272/paper_123.pdf

Zangerle, E., & Müller-Birn, C. (2018). Recommendation-Assisted Data Curation for Wikidata. In: *Wiki Workshop 2018, co-located with The Web Conference*. France. Recuperado de https://doi.org/10.5281/zenodo.1194790