Uso de *Wikidata* y *Wikipedia* para la generación asistida de un vocabulario estructurado multilingüe sobre la pandemia de Covid-19

Using *Wikidata* and *Wikipedia* for assisted generation of a structured multilingual vocabulary about the Covid-19 pandemic

Tomás Saorín; Juan-Antonio Pastor-Sánchez; María-José Baños-Moreno

Cómo citar este artículo:

Saorín, Tomás; **Pastor-Sánchez, Juan-Antonio**; **Baños-Moreno, María-José** (2020). "Uso de *Wikidata* y *Wikipedia* para la generación asistida de un vocabulario estructurado multilingüe sobre la pandemia de Covid-19". *Profesional de la información*, v. 29, n. 5, e290509.

https://doi.org/10.3145/epi.2020.sep.09

Artículo recibido el 18-06-2020
Aceptación definitiva: 11-08-2020

Tomás Saorín
https://orcid.org/0000-0001-9448-0866

Universidad de Murcia
Facultad de Información y Documentación
Depto. de Información y Documentación
Campus Universitario
30100 Espinardo (Murcia), España
tsp@um.es

Juan-Antonio Pastor-Sánchez ✉
https://orcid.org/0000-0002-1677-1059

Universidad de Murcia
Facultad de Información y Documentación
Depto. de Información y Documentación
Campus Universitario
30100 Espinardo (Murcia), España
pastor@um.es

María-José Baños-Moreno
https://orcid.org/0000-0001-9137-1330

Universidad de Murcia
Facultad de Información y Documentación
Depto. de Información y Documentación
Campus Universitario
30100 Espinardo (Murcia), España
mbm41963@um.es

Resumen

Se propone un método para la construcción ágil y dinámica de vocabularios controlados, especialmente para los medios de comunicación, utilizando *Wikidata* y *Wikipedia* como fuentes de información terminológica. El método se aplica a la construcción de un vocabulario sobre la pandemia de Covid-19. Para ello se propone la explotación de la estructura de items y propiedades de *Wikidata* y de los enlaces salientes y entradas de los artículos de *Wikipedia*. Mediante un proceso de definición de reglas de expansión de relaciones de *Wikidata* se ha diseñado un algoritmo en el que se parte de un conjunto de items iniciales y en sucesivas iteraciones y revisión de resultados se recopilan las declaraciones relevantes a la temática del vocabulario. El algoritmo se ha implementado en una aplicación cuyo código y resultados de recopilación del vocabulario sobre la pandemia de Covid-19 se ha publicado en un repositorio abierto. Esto permite utilizar el algoritmo tanto para verificar los resultados usando las mismas u otras reglas de expansión como para su aplicación a la recopilación de vocabularios de otras temáticas. En los resultados también se analizan los elementos recopilados en cada iteración, la propuesta de validación mediante los enlaces entrantes y salientes de los artículos, dejando como futuros trabajos la aplicación de SKOS para la representación interoperable de los vocabularios obtenidos mediante este método.

Palabras clave

Vocabularios controlados; Metadatos; Etiquetas; Palabras clave; Ontologías; Medios de comunicación; Vocabularios para medios; Web semántica; Organización del conocimiento; Emergencias; Catástrofes; Pandemias; Covid-19; Coronavirus; SKOS; *Wikidata*; *Wikipedia*.

Abstract

A method for quickly and dynamically building controlled vocabularies, especially for the media, using *Wikidata* and *Wikipedia* as sources of terminological information, is proposed. The method is applied to construct a vocabulary about the Covid-19 pandemic. For this purpose, it is proposed to exploit the structure of items and properties of *Wikidata* and links and backlinks of *Wikipedia* articles. Using a process based on the definition of *Wikidata* relationship expansion rules, an algorithm was designed, starting from a set of initial items and then being executed in successive iterations, followed by a review of the results. In this way, the *Wikidata* entities relevant to the thematic coverage of the vocabulary are collected. The algorithm has been implemented in an open-source application whose results for the Covid-19 pandemic vocabulary collection have been published in a repository. The algorithm can be used to verify the results using the same or other expansion rules or applied to compile vocabularies in other thematic areas. The results in terms of the elements collected in each iteration and the validation proposal through the links and backlinks of *Wikipedia* articles are also analyzed. The application of SKOS to achieve an interoperable representation of vocabularies obtained by this method is proposed as future work.

Keywords

Controlled vocabularies; Metadata; Tags; Keywords; Ontologies; Media; Media vocabularies; Semantic web; Knowledge organization; Emergencies; Catastrophes; Pandemics; Covid-19; Coronavirus; SKOS; *Wikidata*; *Wikipedia*.

1. Introducción

La información de actualidad está sujeta a la aparición brusca de eventos de impacto (*breaking news*) que alteran la agenda de los medios, acaparando temporalmente la atención sobre ellos y sus consecuencias y relegando al resto de noticias de actualidad sectoriales a un relativo segundo plano de atención y visibilidad. El caso de la pandemia por coronavirus de 2020 es sin duda un caso extremo de esta situación al provocar una paralización global multisectorial, afectando al resto de eventos que con regularidad ocupan la agenda pública. ¿De qué hablan los medios durante el confinamiento? Sobre el coronavirus. Se trata de un caso único en la historia de concentración temática de los contenidos con los que los medios construyen el relato colectivo. Es de interés enfocar esta situación desde el punto de vista de la indización, el etiquetado y los vocabularios para la organización del conocimiento que se usan en los medios de comunicación social. En estos medios, las palabras clave y metadatos descriptivos presentan especificidades claramente diferenciales de las del campo de las fuentes de información para investigación científica sobre el coronavirus (**Rubio-Lacoba**, 2007).

> La pandemia de Covid-19 es un caso único en la historia de concentración temática de los contenidos con los que los medios construyen el relato colectivo

El etiquetado de noticias durante la pandemia es también un caso exacerbado de cambio en el valor de discriminación de un término de indización con respecto a la colección en la que se enmarca: el principio de "*term weighting*" en recuperación de información, que establece que la relevancia de un término para representar un documento está afectada por su tasa de presencia en el conjunto de los documentos (**Baeza-Yates**; **Ribeiro-Neto**, 2011, p. 66) puede aplicarse también para entender los términos usados en la indización. En el caso que nos ocupa, cuando todos los contenidos tratan sobre el coronavirus, este término de indización pierde buena parte de su valor. En este trabajo se plantean cuáles pueden ser los mecanismos de reacción ante una avalancha de contenidos concentrados en un mismo tema, desde el punto de vista de la creación de vocabularios en el contexto de la organización temática e indexación de contenidos de la actualidad.

Es fácil recordar muchos términos que reflejan las muchas caras de la pandemia: confinamiento, mascarillas, pandemia, curva de contagio, tasa de reproducción, covid-19, patógenos, contagio, distanciamiento social, renta básica, ERTE, prevalencia… ¿Cómo organizar esta densa red de conceptos? Y sobre todo, ¿cómo hacerlo a la velocidad que exige el funcionamiento de los medios y aplicado a la organización de contenidos digitales?

Los contenidos colaborativos de *Wikidata* y *Wikipedia* podrían aprovecharse para crear un sistema de organización del conocimiento o un vocabulario controlado sobre la pandemia de la Covid-19 mediante métodos y aplicaciones de procesamiento asequibles. El proceso se abordaría, no desde un punto de vista exclusivamente científico-sanitario, sino también socioeconómico y cultural. Además, debe ser lo suficientemente rico para aportar valor a la organización de los contenidos multimedia que se producen a diario en medios generalistas y poseer una orientación multilingüe e interoperable. La propuesta incluye también la reflexión sobre los mecanismos de validación y generalización del proceso para emergencias informativas de cualquier otro signo: grandes accidentes, catástrofes naturales, atentados terroristas, conflictos armados y todo tipo de emergencias.

> Los contenidos colaborativos de *Wikidata* y *Wikipedia* podrían aprovecharse para crear un sistema de organización del conocimiento o un vocabulario controlado

2. Visión general del control de vocabularios en contenidos altamente dinámicos: la "epidemia de palabras" sobre el coronavirus

El concepto de vocabulario controlado (lenguaje documental en cualquiera de sus formas) lleva años sometido a intensas tensiones debido a las nuevas formas de organización del conocimiento en sistemas digitales y su adaptación a nuevas funciones como la navegación web, el descubrimiento de recursos o el posicionamiento en buscadores (**Lambe**, 2007; **Pérez-Montoro**; **Codina**, 2017).

El país cuenta con un largo recorrido en un vocabulario colaborativo (**Rubio-Lacoba**, 2012), que está integrado por más de 130.000 términos de indización organizados en áreas, entre las que destacan: temas, personajes, organizaciones, lugares y eventos (**García-Jiménez**; **Rodríguez-Mateos**; **Catalina-García**, 2019). Nuevos términos como "Covid-19" o "confinamiento" surgen durante la propia pandemia, y tienen que encajar con otros con mayor trayectoria como "emergencia sanitaria" o "enfermedades infecciosas". Para los medios, la construcción de un vocabulario controlado debe ser un proceso colaborativo muy dinámico entre sus diferentes redacciones y profesionales, caracterizado además por una gran necesidad de respuesta a nuevos términos y fenómenos sociales. Los conceptos estables han de convivir con otros que surgen al calor de los acontecimientos. Este vocabulario, referido por sus responsables como "árbol de conocimiento" (*El país*, 2017), es un ejemplo de sistema de organización del conocimiento que es en parte un tesauro, una folksonomía interna y que tiende a querer adoptar forma de ontología. Este sistema de etiquetado podría ser entendido como un "*organic thesaurus*", conforme a la matriz de "infraestructuras de conocimiento" de **Lambe** (2007, p. 254).

El proceso de construcción de tesauros está bien estudiado por varios autores clásicos (**Broughton**, 2006). Es un proceso exigente y prolongado en el tiempo, que tiende a conformar un instrumento de normalización terminológica en el que la estabilidad pesa más que los procesos de revisión y actualización. Más allá del conjunto básico de relaciones jerárquicas y asociativas entre conceptos en la actualidad se tiende hacia una "ontologización" que haga disponibles más tipos de relaciones entre conceptos. Para ello se procede a la formalización en tipos, subclases, propiedades y subpropiedades, lo que combinado con la correspondencia entre conceptos de diferentes esquemas, permite abordar con mayores expectativas la organización y navegación entre contenidos. Tal como afirma **Stuart** (2016): para entender mejor el proceso de construcción de vocabularios controlados se deben considerar también técnicas de construcción de ontologías.

> Para entender mejor el proceso de construcción de vocabularios controlados se deben considerar también técnicas de construcción de ontologías

Los vocabularios controlados desarrollados como proyectos de largo recorrido se amoldan bien a los recursos científicos, académicos y culturales. Sin embargo, no se adaptan de igual forma en el contexto de los medios de comunicación, donde la actualidad responde a nuevos focos de interés de la vida social, nuevos eventos y una negociación constante de nuevos conceptos y hechos. Un evento como la pandemia de Covid-19 genera una serie de acontecimientos, terminología, agentes, medidas sociales y discursos, que rompen las costuras de los sistemas previos de etiquetado de contenidos, los cuales necesitan ampliarse durante el transcurso de los propios hechos.

En la construcción de un tesauro o de una ontología, en cualquier de las diferentes propuestas metodológicas habitualmente aplicadas, se contemplan aspectos de recolección y selección terminológica y organización de relaciones. Estas tareas requieren una comprensión clara del dominio abordado y una alta cualificación. Pero en el caso de los medios de comunicación social, este trabajo puede enfocarse desde la perspectiva del "*accidental taxonomist*" (**Hedden**, 2016), puesto que es realizado por profesionales de la comunicación durante las tareas de redacción. Dado que la respuesta a una nueva situación de etiquetado y organización del conocimiento es difícilmente abordable en tiempo y forma, pueden ser valiosas nuevas estrategias eficientes de producción. Se trata de un proceso con un ritmo muy acelerado, que requiere una respuesta algorítmica y no editorial: los patrones para indexar y comprender el contenido deben automatizarse y no depender de la intervención humana para señalar nuevos temas.

Durante el confinamiento han proliferado análisis sobre los cambios en el comportamiento digital, enfocados hacia la comprensión de nuevos hábitos de consumo en circunstancias excepcionales (*GlobalWebIndex*, 2020). Los reportes sobre tendencias de búsqueda y hábitos de comportamiento digital difundidos desde la industria de los buscadores durante el período de confinamiento, permiten acercarnos a un vocabulario que refleja las necesidades de los usuarios. Ahora bien, los términos usados en estas búsquedas reflejan también preocupaciones, ansiedades y, como algunos autores han expresado, permiten a los buscadores acceder de forma masiva a la forma de pensar y sentir de millones de personas (**Galloway**, 2017). En *Google Insights* encontramos un mapa del pensamiento colectivo "basado en hechos reales" (**Sinclair**, 2020), un resumen consensuado de lo que está pasando en el pensamiento de los usuarios. La monitorización de términos de búsqueda con *Google Trends* permite explorar el vocabulario que se está construyendo en el transcurso de la pandemia, identificar picos de interés y obtener sugerencias de términos o expresiones relacionados. Para una búsqueda por "desescalada" y "confinamiento" se pueden encontrar expresiones como "Fase 0" o "salida 2 de mayo desescalada".

> Los vocabularios controlados desarrollados como proyectos de largo recorrido se amoldan bien a los recursos científicos, académicos y culturales, pero no se adaptan de igual forma en el contexto de los medios de comunicación

Estos recursos para la planificación de palabras clave permiten comprender el problema: la variabilidad puntual de la terminología aplicable a la realidad informativa. Pero también podrían aportar material para la elaborar vocabularios aplicables a la organización de contenidos que es, desde la posición de los editores de medios, otro de los retos que plantea la "emergencia informativa".

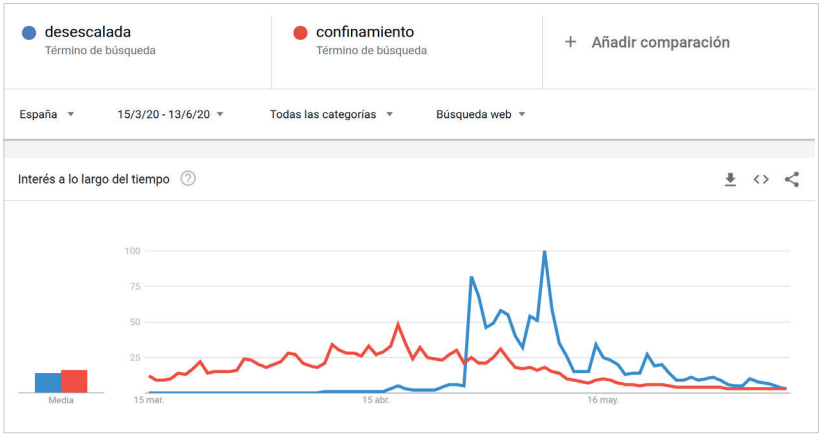

Figura 1. Comparación entre los términos "desescalada" y "confinamiento" desde el 13 de marzo hasta el 10 de junio de 2020.
https://cutt.ly/NfYVHHu

En el caso de la pandemia por Covid-19 se ha producido una concentración del foco de interés alrededor de un único tema de una forma singular. La atención que acapara un evento de impacto global, como el tsunami provocado por el terremoto del océano Índico de 2004 o el accidente nuclear de Fukushima en 2011, se ve compensada con la continuidad de la vida social: elecciones, campeonatos deportivos, concursos de televisión, estrenos de películas, noticias nacionales y locales. La crisis provocada por la pandemia ha supuesto una paralización brusca de la vida social, y ha aglutinado el grueso de la producción informativa y la atención de las personas. Si una noticia del mes de febrero sobre la propagación del virus en Italia podía ser etiquetada con "coronavirus", en un artículo del mes de abril sobre cómo afecta el confinamiento a los niños ya no es significativa esa etiqueta, puesto que todas las noticias hablan, de una manera u otra, sobre el coronavirus y sus efectos. La intensificación de la producción de contenidos tiene como efecto al mismo tiempo una mayor profundidad en el tema central (virus, contagio, epidemiología) ampliando el tema en el resto de campos.

Confinamiento, desescalada, UCI, respiradores, renta básica, distanciamiento social, reservorio, vector de contagio, curva de contagio, prevalencia, ensayo clínico, estado de alarma, *cabine fever*…, conforman un denso vocabulario que ha adquirido ante nuestros ojos, día a día, una especial consistencia para reflejar la realidad. ¿Cómo organizar los contenidos producidos durante la pandemia? ¿Cómo incorporar sentido al etiquetado de las noticias? El enfoque clásico de los vocabularios controlados se enfrenta a unas condiciones inauditas de velocidad de respuesta y concentración temática.

Un análisis de las 10.841 noticias publicadas por *El país* entre el 1 de marzo y el 18 de mayo de 2020 muestra que 6.352 (casi un 58,6%) estaban etiquetadas con "Coronavirus" y/o "Coronavirus Covid-19". En estas noticias se observa una presencia altamente significativa de unos pocos términos como "Emergencia sanitaria", "Enfermedades infecciosas", "Neumonía", etc. (figura 3). Junto a estas palabras clave existe una larga cola de más de 7.800 términos con una frecuen-

Figura 2. Ejemplos de etiquetado de noticias sobre coronavirus en *El país*.

cia menor al 1%, tales como "Estado de alarma", "Mascarillas" o "Cierre de establecimientos".

Pese a usarse 8.108 palabras clave diferentes durante el período, existe un exceso de concentración terminológica más acusado y sostenido en el tiempo que en cualquier otro acontecimiento noticioso de impacto. La figura 3 muestra los términos que con mayor frecuencia aparecen en el subconjunto de noticias indexadas con los dos términos anteriormente mencionados para el coronavirus.

Esta visión de un medio en particular es solo un ejemplo para mostrar la diferencia que existe entre el contexto de los medios de comunicación y el de las bases de datos de información científica de investigación biomédica, como los promovidos por el *Allen Institute for Artificial Intelligence*: *Covid-19 Open Research Dataset Challenge* (*CORD-19*), o el *Covid-19 Knowledge Graph*, basados ambos en técnicas automáticas de extracción y organización de conocimiento sobre publicaciones científicas.
https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
https://github.com/covid19kg

En estas fuentes de información científicas y académicas los retos son: la comprensión del propio virus, su contagio, el desarrollo de la enfermedad, el tratamiento y los datos epidemiológicos. Para los medios de comunicación supone abordar la pandemia como un fenómeno social multidimensional que requiere explicación y relato, donde intervienen tanto la comprensión de los aspectos científicos como de sus repercusiones en todas las esferas sociales.

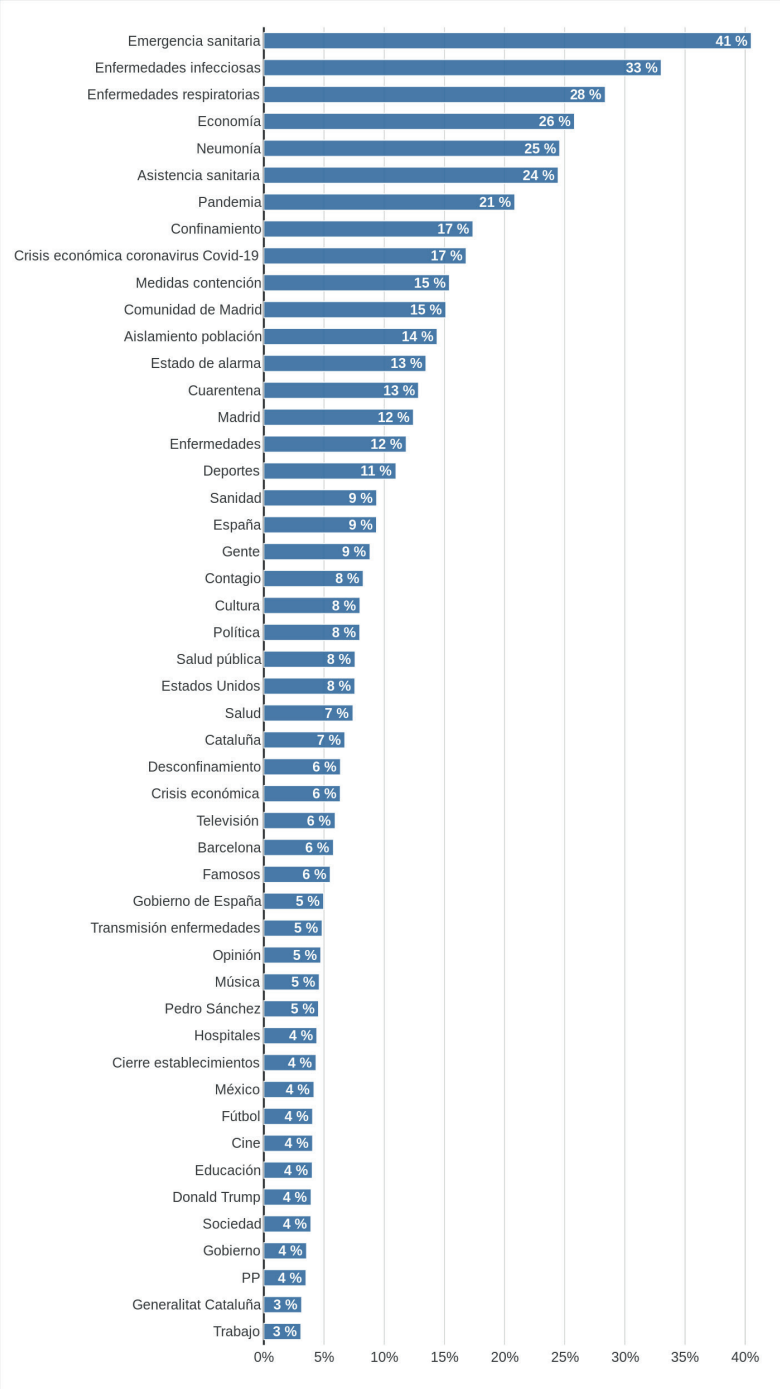

Figura 3. Distribución porcentual de las 50 palabras clave más frecuentes en las noticias publicadas en *El país* entre el 1 de marzo y el 18 de mayo de 2020, dentro del subconjunto de 6.352 noticias etiquetadas con "Coronavirus" o "Coronavirus Covid-19". Fuente: elaborado a partir de extracción del marcado semántico de la hemeroteca de la web de *http://www.elpais.com*

La situación descrita conecta con una de las características que se han hecho notar sobre *Wikipedia* como enciclopedia multidominio: la amplia cobertura que tienen los temas vinculados a la actualidad y la cultura de masas. La actividad de los "wikipedistas de actualidad" (**Saorín**, 2017) permite acceder, al mismo tiempo que transcurren los hechos, a una considerable cantidad de información semiestructurada relacionada con nuevos hechos y conceptos en construcción.

Por este motivo, este trabajo se apoya en la propia organización y producción de conocimiento presente en el proyecto de contenido colaborativo de *Wikidata* y de *Wikipedia*. Estos proyectos son una fuente de organización del conocimiento y vocabulario controlado, con contenidos semiestructurados, fácilmente accesibles para su procesamiento con diferentes tecnologías y enfoques.

Wikipedia es un recurso utilizado habitualmente en las investigaciones del campo del procesamiento del lenguaje natural (PLN), especialmente para el reconocimiento, contextualización y categorización de entidades (**Ehrmann**; **Rosset**, 2016). Esta situación adquiere una dimensión nueva a partir de la maduración del proyecto *Wikidata*. *Wikipedia* ofrece artículos textuales, en un proceso continuo de mejora y revisión editorial. En cada una de las más de 200 ediciones en idiomas distintos, los artículos son contenidos independientes, con extensión, calidad, y editores diferenciados. Estos

artículos se componen con elementos estructurales de organización interna para apartados por niveles, equivalentes a los h1, h2 de html. El principal mecanismo de navegación son los enlaces internos a otros artículos de la misma edición de la enciclopedia (*links*), conformando una densa red de relaciones unidireccionales que permiten ampliar información sobre numerosos aspectos del contexto de los artículos. También deben tenerse en cuenta los enlaces recibidos desde otros artículos (*backlinks*), que son una función complementaria nativa ofrecida por los sistemas *MediaWiki*.

> *Wikidata* amplifica las cualidades de *Wikipedia* como fuente de "conocimientos conectados" agrupando cada artículo equivalente de cada una de las *Wikipedias* en una entidad específica con un identificador único, independiente del idioma

Wikidata amplifica las cualidades de *Wikipedia* como fuente de "conocimientos conectados" agrupando cada artículo equivalente de cada una de las *Wikipedias* en una entidad específica con un identificador único, independiente del idioma. Los 145 artículos sobre Goya de las ediciones de *Wikipedia* en diferentes idiomas se corresponden con la entidad Q5432 de *Wikipedia*. A su vez, los artículos sobre el barrio madrileño de Goya se corresponden con la entidad Q3814578.

Las propiedades de cada uno de los items de *Wikidata* se describen siguiendo un modelo de datos semántico. Los items se organizan en tipos y clases, conformando una ontología colaborativa *sui generis* (**Piscopo**; **Simperl**, 2018). Mientras en *Wikipedia* la edición en cada idioma es un proyecto independiente que elabora y crea sus artículos desde el punto de vista válido para su comunidad de editores, en *Wikidata* se da forma a un banco de datos central reutilizable por todos los proyectos *Wikimedia*.

Dichas descripciones (*claims*) adoptan la forma de sujeto-predicado-objeto de manera muy similar a las tripletas RDF. *Wikidata* podría entenderse como una plataforma de publicación de datos estructurados, con una arquitectura autosuficiente, basada en la creación de declaraciones. Junto con los items, las otras entidades esenciales del modelo de datos de *Wikidata* son las propiedades. Cada propiedad representa un tipo de relación específica que puede establecerse entre dos elementos de *Wikidata* o almacenar un valor literal. Por ejemplo, la propiedad "P106: ocupación" permite indicar la profesión de un personaje, estableciendo como destino un ítem, como por ejemplo "Q33999: actor". Por su parte la propiedad "P571: fecha de creación" permite almacenar un literal (de tipo fecha en este caso). Hay otros aspectos, como las referencias, los rankings y los calificadores (similar a la reificación en RDF), que no son relevantes para este trabajo.

Paralelamente a la descripción de items del tipo nombres propios (París, 2ª Guerra Mundial) también se describen conceptos genéricos o comunes (Ciudad, Guerra). Esto permite incorporar taxonomías mediante las propiedades "P31: instancia de" y "P279: subclase de" que aportan una semántica interoperable muy valiosa. La primera permite asignar un tipo a cada ítem, lo cual a su vez permite un primer nivel de agrupación

("Q90: París" → "P31: instancia de" → "Q515: ciudad").

La segunda permite definir una taxonomía de clases y subclases:

("Q28389: guionista" → "P279: subclase de" → "Q36180: escritor" → "P279: subclase de" → "Q482980: autor").

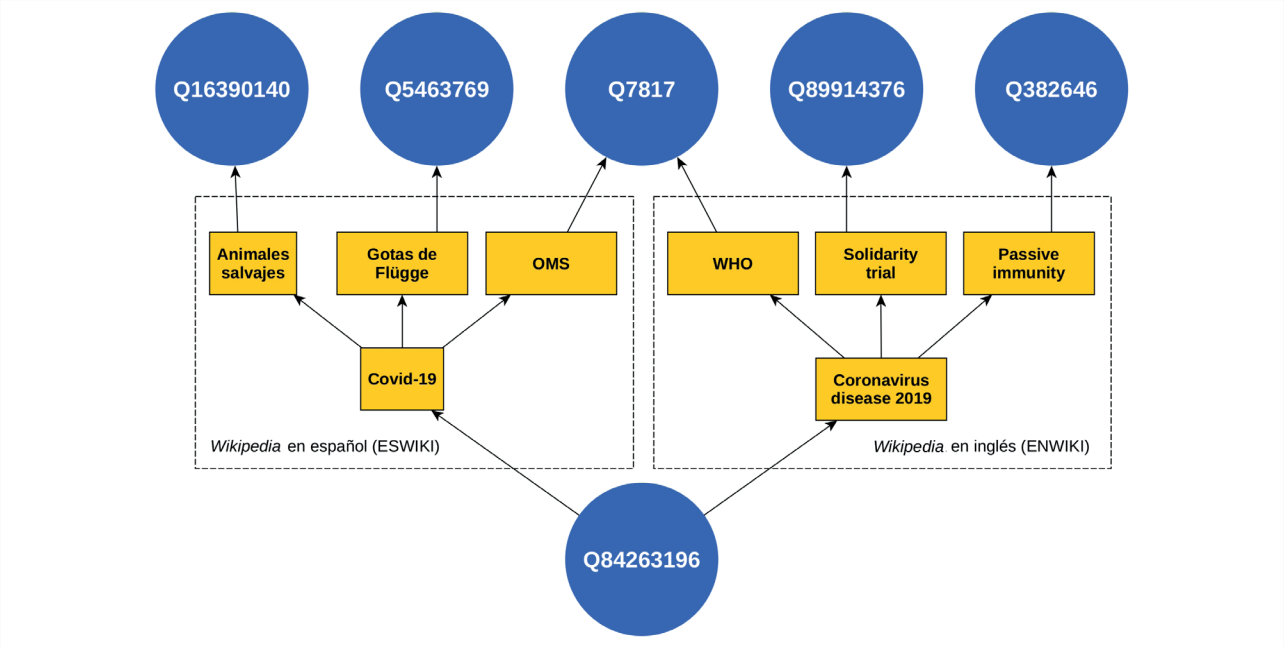

Figura 4. Ejemplos de enlaces de los artículos sobre Covid-19 en las ediciones en español (*Eswiki*) e inglés (*Enwiki*) de *Wikipedia* y sus equivalencias con las correspondientes entidades de *Wikidata*.

Mientras en *Wikipedia* tenemos enlaces (vínculos sin significado), en *Wikidata* encontramos relaciones (vínculos con un significado claramente definido). Esta doble estructura puede combinarse para extraer conceptos y términos de interés partiendo de un ítem para capturar la riqueza conceptual de la actividad de los editores de un artículo de *Wikipedia* al explicar un tema, y de los editores de una descripción en *Wikidata* al analizar un concepto.

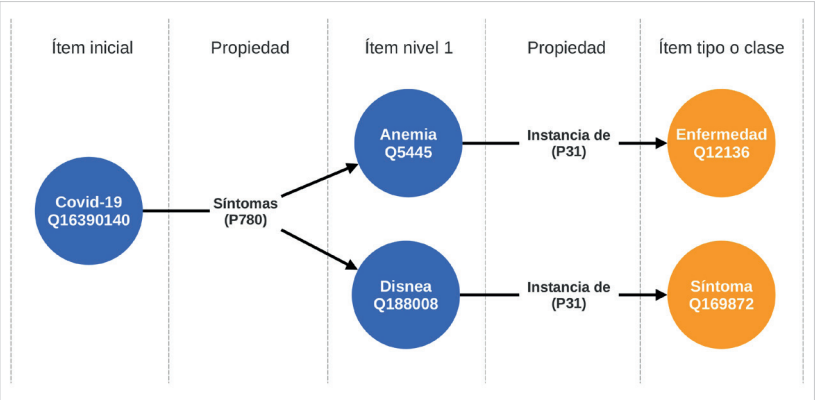

Figura 5. Esquematización de la tipología de entidades relacionadas.

Esto permite comprender mejor el significado de muchos de esos meros enlaces que encontramos en el texto de un artículo *Wikipedia*. Recorriendo primero los enlaces entrantes y los salientes de un artículo, y posteriormente las relaciones disponibles, es posible obtener un conjunto de términos estructurados relevante para un tema dado. A través de *Wikidata* se puede establecer que el enlace desde el artículo sobre Covid-19 hacia el artículo sobre Cloroquina tiene el significado de "medicamento, procedimiento o terapia usada para tratar una enfermedad" (P2176).

Debido a que casi todos los items de *Wikidata* disponen de una declaración de tipo o clase, es posible obtener con qué clase de items se relaciona un artículo de *Wikipedia*. En un primer nivel básico puede extraerse el tipo o clase de cada ítem relacionado. Mediante la propiedad con la que se relaciona con el ítem inicial, y con su tipología, se obtiene un esquema de tipos y clases, tal y como se muestra en la figura 5.

El modelo de datos de *Wikidata* dispone de una representación RDF de manera que los datos son interrogables mediante Sparql utilizando *Wikidata Query Service* (*WDQS*). Mediante consultas Sparql es posible recuperar las relaciones de un ítem con otros (incluyendo tipos y clases). De igual forma, es posible plantear una consulta inversa para obtener los items que están relacionados con uno específico. A través de *Wikipedia* y *Wikidata* es posible obtener las relaciones directas desde y hacia los items iniciales identificados para desarrollar el proceso de recopilación de los elementos del vocabulario. Sparql también ha sido utilizado para explorar la estructura de propiedades de los items iniciales, tal y como se muestra en la figura 6 (ver consulta en Anexo I, consulta 1).

Aunque este lenguaje ofrece una gran potencia y versatilidad en este contexto, la expansión de la consulta a relaciones de más de un nivel no puede generarse de forma efectiva y significativa únicamente con Sparql. Para este fin, tal y como se muestra a continuación, se ha diseñado e implementado el correspondiente algoritmo.

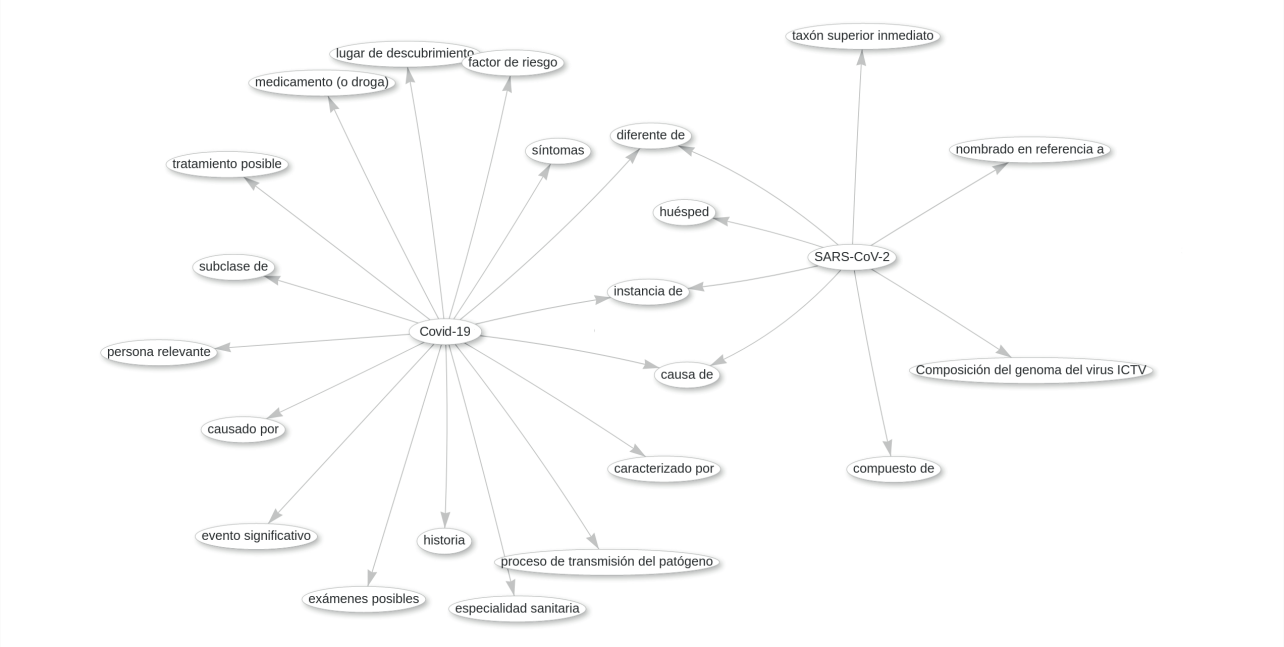

Figura 6. Grafo de relaciones que se establecen para Covid-19 y SARS-CoV2. Se excluyen las relaciones internas de proyectos *Wikimedia*. Fuente: Elaborado desde consulta Sparql en *http://query.wikidata.org*

3. Método de elaboración del vocabulario

Este trabajo propone un método para obtener de forma semiautomatizada, a partir de *Wikidata* y *Wikipedia*, un vocabulario organizado preliminar sobre un acontecimiento de actualidad. También muestra la aplicación de dicho método para el caso de la pandemia por coronavirus como fenómeno social. Para ello se contextualizan las circunstancias singulares que ha provocado este fenómeno desde el punto de vista de la terminología usada en la producción de contenidos en los medios de comunicación. No se abordan otros ámbitos de interés en el campo de la comunicación, información y documentación, como son las redes sociales, los conjuntos de datos de impacto de la pandemia, las fuentes de información bibliográfica científica o los recursos llevados a cabo para acelerar el avance de la investigación científica en este campo. Complementariamente, se sitúan los aspectos esenciales de los procesos de construcción de vocabularios y ontologías en el marco de la organización del conocimiento de recursos y fuentes de información, centrándonos en las primeras fases de recopilación y organización de la terminología.

La investigación realizada se entiende como una herramienta ágil y abierta para diversas fases del proceso de construcción de tesauros y taxonomías aplicados a la recuperación de información (**Broughton**, 2006, pp. 58-150). Este trabajo también se sitúa en el marco complementario de los procesos de construcción de ontologías para la estructuración de contextos de conocimiento, conforme a la propuesta metodológica de **Stuart** (2016). Se utilizan las que dicho autor denomina tanto "colaborativas" como "empíricas", así como la reutilización de vocabularios y esquemas que caracteriza la metodología de construcción de ontologías en el contexto de los datos enlazados (**Suárez-Figueroa** *et al.*, 2012).

En la construcción de tesauros y ontologías se dan procesos de captura de terminología o conocimiento desde documentos del campo: bibliografía, informes, etc. Es una aproximación con un importante recorrido y resultados tangibles desde el campo del procesamiento del lenguaje natural (PLN). Sin embargo, las capacidades y recursos del PLN exigen una considerable capacidad técnica que no está al alcance de muchos organismos publicadores de contenidos digitales. Por este motivo resulta de gran interés la captura de terminología inicial a partir de fuentes estructuradas abiertas, colaborativas y vinculadas a los acontecimientos de actualidad tales como *Wikipedia* y *Wikidata*.

Numerosas experiencias utilizan *Wikipedia* como fuente de validación terminológica, mediante el procesamiento del texto de los artículos, sus enlaces, las páginas de desambiguación y redirección, y las categorías. *Wikipedia* y *Wikidata* forman parte habitual del paquete de recursos utilizados en el cada vez más relevante campo de las *Named Entities Recognition* (**Nouvel**; **Ehrmann**; **Rosset**; 2016) y los grafos de conocimiento como modelos externos de comprensión del contexto para el procesamiento complejo de información semiestructurada (**Fensel** *et al.*, 2020). Es frecuente combinar las técnicas de análisis de redes para obtener agrupaciones y redes de conceptos a partir de los enlaces existentes en *Wikipedia* (**Minguillón** *et al.*, 2017).

Ambas iniciativas permiten abordar la recopilación estructurada de terminología sin recurrir a técnicas de PLN. Se trata de aprovechar la densa estructura de enlaces internos entre artículos de *Wikipedia*, reforzada con la precisa estructura de relaciones entre conceptos que aporta *Wikidata*. Los artículos en cada idioma se encuentran unificados desde el punto de vista del conocimiento factual a través de los items de *Wikidata* (**Saorín**; **Pastor-Sánchez**, 2018). A lo largo de un artículo elaborado de forma colaborativa por diferentes comunidades y actualizado conforme a la evolución de los acontecimientos noticiosos, encontramos enlaces en los que se identifican conceptos relevantes para el tema. Estos enlaces serán tratados con independencia del idioma en que están escritos y consolidados a través de la base de conocimiento de *Wikidata*, utilizada al mismo tiempo como fuente de datos factuales (instancias, acontecimientos) y datos estructurales (clases, tipologías) a modo de ontología colaborativa (**Piscopio**; **Simperl**, 2018).

Durante la pandemia se han coordinado algunos proyectos alrededor de *wikiproyectos* en varias de las enciclopedias, así como en *Wikidata*. La comunidad *Wikimedia* ha ofrecido un resumen de métricas relacionadas con la Covid-19 mostrando su cobertura, labor editorial y visitas en más de 175 idiomas. También se construyó un subconjunto de 5.209 artículos asociados con la pandemia partiendo de la exploración de artículos relacionados.

El método propuesto selecciona un reducido número de items de *Wikidata* en lugar de recopilar exhaustivamente los artículos de *Wikipedia* relacionados con un tema. Más concretamente se han seleccionado tres items iniciales de *Wikidata* para construir el vocabulario:

- la enfermedad Covid-19,
- el virus SARS-CoV-2,
- la pandemia global.

Este criterio ayuda a que la metodología sea fácilmente reproducible a bajo coste.

El algoritmo planteado considera dichos items iniciales como punto de partida para un proceso de extracción de relaciones tipadas significativas en *Wikidata*, expandiéndose a varios niveles. Se aplican procedimientos para filtrar relaciones poco o nada significativas en función de la temática y el propósito del vocabulario a construir. Así pues, la base de conocimiento de *Wikidata* adopta un doble papel:

- como recopilación de entidades tanto concretas como genéricas (individuos, enumeraciones, clases, conceptos genéricos);
- como medio de extracción de atributos que describen cada entidad.

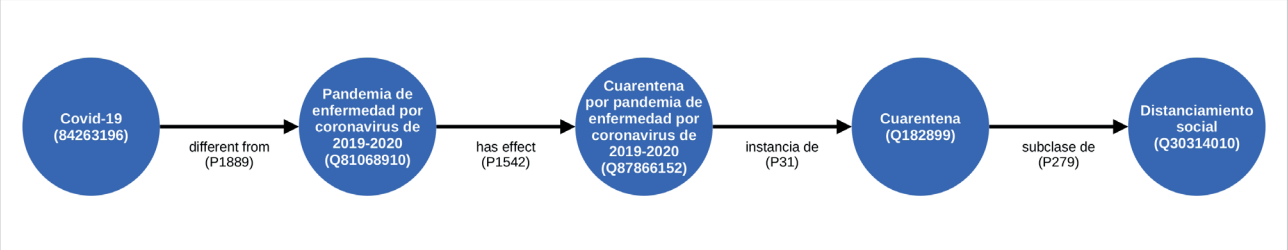

Figura 7. Cadena de relaciones de "Covid-19" hasta "Distanciamiento social".

Ambos tipos son relevantes tanto en los tesauros o esquemas de conceptos y ontologías integradas en *datasets*, como en el campo emergente de la búsqueda orientada al reconocimiento de entidades o autoridades (**Balog**, 2018). Por su parte, los enlaces no-tipados de *Wikipedia* se utilizan para una validación preliminar de items y relaciones relevantes en *Wikidata*.

El proceso permite obtener un conjunto terminológico organizado según la estructura de *Wikidata*, descargable en un formato abierto e interoperable. Aunque no ha sido tratado en este trabajo, este conjunto de datos también podría ser multilingüe, incluir sinónimos y vincularse con otros vocabularios de la web. Tanto el vocabulario generado como la implementación del algoritmo utilizado para su obtención se han publicado en un repositorio abierto.
https://github.com/j-pastor/wdkge

La publicación de los contenidos de *Wikipedia* y *Wikidata* utiliza formatos abiertos y con licencias adecuadas de reutilización. Desde el punto de vista técnico, la plataforma *Mediawiki* dispone de una API para la consulta y explotación de los datos de ambos proyectos. Por lo tanto, pueden explotarse como fuente terminológica para la construcción de vocabularios controlados, aprovechando el valor que reside en la densidad de los enlaces de *Wikipedia* y la disponibilidad de relaciones semánticas formalizadas en *Wikidata*.

Considerando la estructura de *Wikidata* resulta factible la exploración mediante la expansión de la red de relaciones, para la extracción de entidades y propiedades con el objetivo de elaborar un vocabulario estructurado. Se trataría de un proceso semiautomático con supervisión de un especialista que tomaría decisiones para limitar el recorrido a través de la red de relaciones.

La red de enlaces existente entre los artículos de *Wikipedia* puede reutilizarse con fines de revisión y validación del vocabulario. Los enlaces salientes (*links*) y entrantes (*backlinks*) de los artículos de *Wikipedia* han sido creados con criterio editorial, y en artículos de diferentes idiomas. A partir de dichos enlaces se podría construir un conjunto con los items de *Wikidata* a los que se refieren los artículos enlazados. Únicamente se consideran los enlaces de los artículos (de las ediciones relevante de *Wikipedia*) correspondientes a los items iniciales de *Wikidata* seleccionados. Podría entenderse que la existencia de un enlace hacia y desde un artículo de *Wikipedia* podría servir de apoyo para una prevalidación supervisada de las relaciones entre items que se obtengan durante el proceso de exploración de la estructura de *Wikidata*.

La propuesta mostrada en este trabajo explota únicamente los items y las propiedades que los relacionan, dejando de lado otras propiedades que definen valores o atributos. Considerando todo lo anterior, el proceso de construcción del vocabulario se llevaría a cabo en tres fases:

- Fase 1: Inicio. Además de identificar los items iniciales a utilizar como puntos de partida para la exploración de la estructura de relaciones, se establecerá el idioma en el que se recuperarán las etiquetas y se construirá un conjunto de *links* y *backlinks* a partir de las ediciones de *Wikipedia* que se utilizarán para las propuestas de validación.
- Fase 2: Iteraciones. En cada iteración se recopilan las declaraciones a partir de la lista de relaciones de los items recuperados en la iteración anterior. En la iteración inicial (iteración 0) se parte de los items iniciales identificados en la fase anterior. Al inicio de la iteración se recopila un conjunto previo declaraciones. Tras su revisión se definen las reglas de exclusión de entidades y se identifican los nodos hoja (que pasarán además a la siguiente iteración). Tras la revisión se procede a recopilar las declaraciones y realizar un marcado de las propuestas de validación utilizando el conjunto de *links* y *backlinks* recuperado en la fase inicial.
- Fase 3: Enriquecimiento. Al igual que es posible obtener en *Wikipedia* los enlaces entrantes a un artículo, en *Wikidata* también es posible obtener las relaciones que apuntan a un ítem. Aquellos items que apunten a los items iniciales identificados en la Fase 1 y que no hayan sido recuperados durante las iteraciones se incorporarán al vocabulario.

3.1. Fase 1: Inicio

El primer paso es identificar los items iniciales a partir de los cuales se realizará la exploración de la red de relaciones de *Wikidata*. En este caso se parte de tres conceptos que disponen de sus propios artículos en diversas ediciones de *Wikipedia* y, por lo tanto, de sus correspondientes items únicos en *Wikidata*:

- el virus SARS-CoV-2;
- la enfermedad Covid-19;
- la pandemia por coronavirus 2019-2020,

que corresponden con los items de Wikidata: Q82069695, Q84263196 y Q81068910, respectivamente.

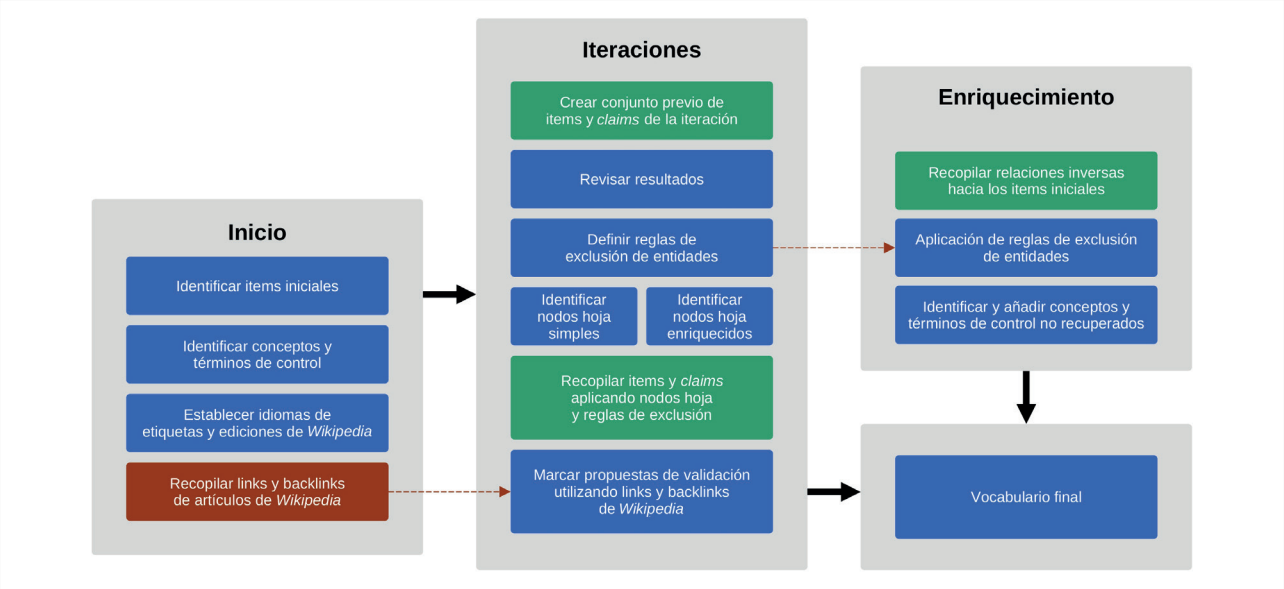

Figura 8. Esquema resumen del proceso de construcción del vocabulario.

Tabla 1. Datos editoriales de los elementos de *Wikidata* utilizados como items iniciales y sus correspondientes artículos en las ediciones en español e inglés de *Wikipedia*. Datos de fecha y número de ediciones a 27-05-2020.

Concepto	*Wikidata*			*Wikipedia* en español		*Wikipedia* en inglés	
	Id. ítem	Creación del ítem	N. de ediciones	Creación del artículo	N. de ediciones	Creación del artículo	N. de ediciones
SARS-CoV-2	Q82069695	14-01-2020	>1.000	19-01-2020	>700	09-01-2020	>3.300
Covid-19	Q84263196	02-02-2020	>1.300	11-02-2020	>700	05-02-2020	>4.700
Covid-19 pandemic	Q81068910	05-01-2020	>2.500	19-01-2020	>6.700	05-01-2020	>21.000

Wikidata puede recorrerse mediante la expansión de las relaciones entre items vinculados a varios niveles de profundidad para consultas Sparql. La expansión de relaciones también implica un crecimiento exponencial en cada paso que nos aleja del nodo de partida, y la generación de ruido terminológico. De hecho, una exploración de estos items iniciales a lo largo de, por ejemplo, cuatro niveles de profundidad, arroja más de 95.000 items distintos, relacionados a través de más de 237.000 declaraciones utilizando para ello 635 propiedades diferentes. Por este motivo es preciso establecer límites de exploración y reglas para definirlos.

En esta fase también se procede a recuperar los items de *Wikidata* correspondientes a los *links* y *backlinks* de los artículos de las ediciones de interés de *Wikipedia*. En este caso las ediciones utilizadas han sido las del español e inglés aunque hay que señalar que en principio podrían utilizarse las ediciones de *Wikipedia* en los idiomas que se desee, puesto que no se recopila el texto del enlace en un idioma concreto, sino la entidad correspondiente en *Wikidata*.

3.2. Fase 2: Iteraciones

El proceso de elaboración del vocabulario se ejecuta en sucesivas iteraciones para profundizar en la estructura de relaciones de *Wikidata*. En cada iteración el especialista debe analizar los resultados obtenidos para aplicar una serie de decisiones editoriales tendentes a reducir el volumen de términos y relaciones. Esta supervisión también permite homogeneizar el tipo de relaciones y dotar de cohesión al vocabulario verificando la especificidad/generalidad del significado de los items obtenidos.

Tales decisiones se enfocan en dos direcciones:

1) Identificar los items, propiedades o clases de items no pertinentes a la temática del vocabulario y que no deben incluirse expresamente.

2) Identificar los items, propiedades o clases de items en los que debe detenerse la exploración de la estructura de relaciones (nodos hoja).

A su vez, en relación con los nodos hoja identificados en (2) puede optarse por incluir el ítem sin más o también las propiedades que definen la clases, instancias, partes o ubicación en las se enmarque. De este modo se obtiene un equilibrio en la obtención de items adicionales que permiten contextualizar el vocabulario y por ende su aplicación resultará útil durante la descripción e indización de recursos.

Se ha procedido a descartar los items recuperados durante el proceso de exploración que contienen determinados items, propiedades o clases de items. En el caso del vocabulario sobre la pandemia de Covid-19 se han descartado las siguientes propiedades:

- P2354: lista del elemento
- P910: categoría principal del tema
- P1151: portal principal del tema
- P1424: plantilla principal del tema
- P5008: lista de interés para el proyecto *Wikimedia*
- P5125: esquema de *Wikimedia*
- P6104: mantenido por el wikiproyecto
- P7867: categoría para los mapas de este elemento
- P2293: asociación genética
- P1343: descrito en
- P769: acción farmacológica alterada por

Como puede observarse, estas propiedades están asociadas a aspectos de gestión de *Wikimedia* o a información técnica demasiado amplia o irrelevante para el ámbito temático del vocabulario.

Para identificar los nodos hoja hay que considerar que todos los items en *Wikidata* tienden a ser definidos al menos como "instancia de" (P31) o "subclase de" (P279), lo cual permite trabajar con tipos de entidades y una pseudo-ontología de clases. Además, también puede resultar interesante explotar otro tipo de relaciones de causalidad (P1542: causa de / P828: causado por), localización y ubicación (P131: situado en / P276: ubicación), etc. Considerando esto, en los casos que sea conveniente detener la exploración de la cadena de relaciones en un determinado concepto, puede resultar conveniente obtener este tipo de relaciones con sus correspondientes instancias, subclases, tipos, causas, ubicaciones, etc. Son relaciones que podrían ser útiles para dotar de contexto a los elementos del vocabulario. Utilizando WDQS y planteando la oportuna consulta Sparql (ver Anexo I, consulta 2) se ha identificado la totalidad de propiedades disponibles en *Wikidata* y tras su análisis se han identificado las que son más adecuadas para esta finalidad:

- P31: instancia de
- P131: situado en la entidad territorial administrativa
- P279: subclase de
- P361: forma parte de
- P1269: faceta de
- P1889: diferente de
- P1542: causa de
- P793: evento significativo
- P828: causado por
- P276: ubicación

Tomemos la figura 9 para mostrar la diferencia existente al definir un nodo hoja recuperando también este tipo de relaciones, en comparación con la definición de un nodo hoja simple en donde no se recuperan las mismas.

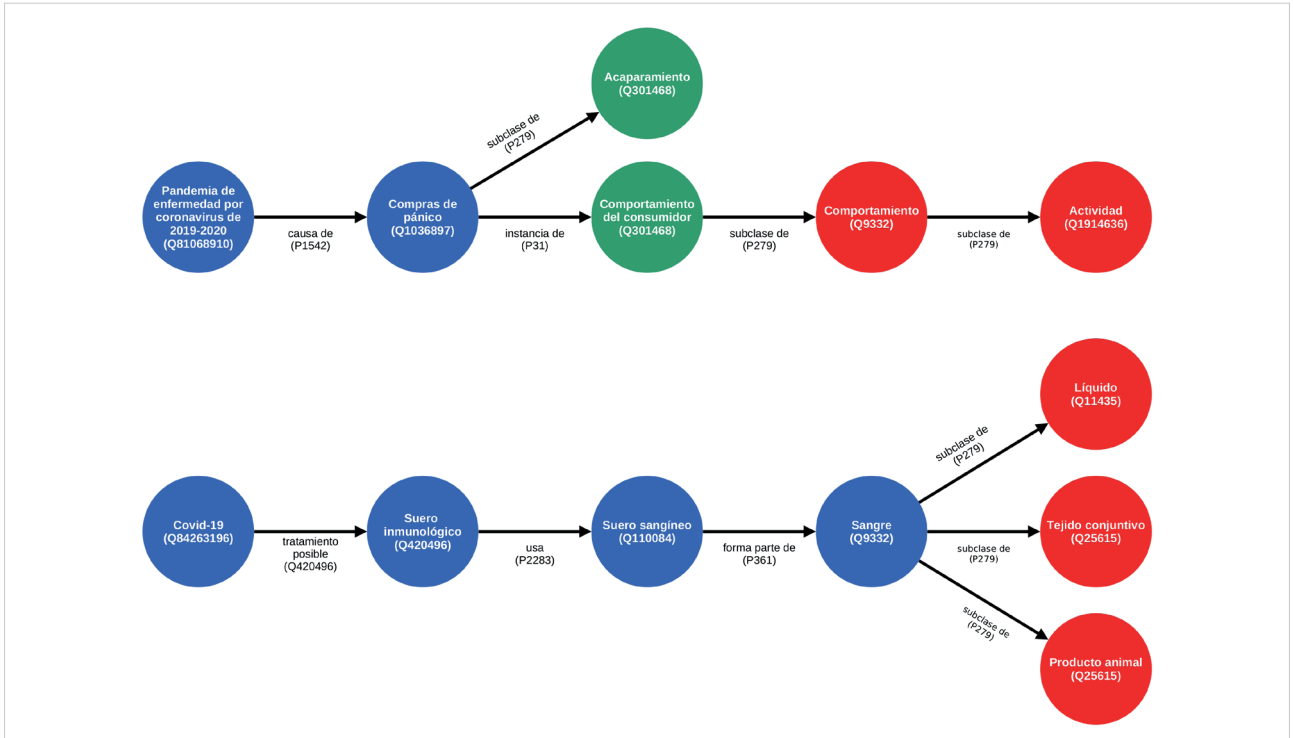

Figura 9. Comparación entre nodos hoja (arriba) y nodos hoja simples (abajo).

En el primer caso, al definir el nodo "Compras de pánico" (Q1036897) como nodo hoja, la exploración se detiene en este punto, pero también se recuperan las propiedades de contextualización junto con los items correspondientes. En el segundo caso, al definir "Sangre" (Q9332) como nodo hoja simple, la exploración finaliza inmediatamente sin recuperar ninguna otra propiedad ni los consiguientes items. Este mecanismo permite tener un control a nivel de ítem sobre las condiciones para finalizar el recorrido de la estructura de relaciones de *Wikidata*.

A veces la definición de un nodo hoja precisa de reglas más generales que las obtenidas al hacerlo a nivel de ítem individual. Para ello se pueden establecer las clases o propiedades que deben marcar dicha definición, tanto a nivel de nodo hoja con sus relaciones de contextualización como para los nodos hoja simple. Un ejemplo de items identificados como clases para definir nodos hoja son los referentes a la definición de País, Estado, Ciudad-Estado, etc.:

- Q6256: País
- Q3624078: Estado soberano
- Q15634554: Estado con reconocimiento limitado
- Q133442: Ciudad-Estado
- Q5164076: Estado constitutivo
- Q15304003: País del Reino de los Países Bajos

El uso de los items anteriores como objeto en una declaración puede realizarse a través de diversas propiedades. El criterio seguido ha sido la identificación de las declaraciones cuya propiedad coincida con alguna de las propiedades de contextualización. Esto permite definir reglas generales sin tener que especificar de forma individual todos los items que son instancia de una clase, tal y como se muestra en la figura 10.

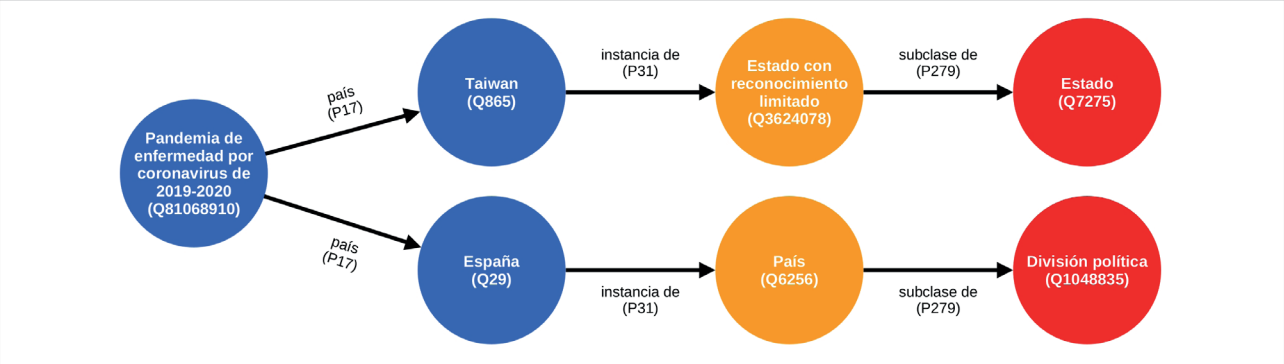

Figura 10. Definición de nodos hoja mediante clases.

En este caso, los nodos en color azul son recopilados en el vocabulario, mientras que los marcados en color rojo son descartados. Los nodos en color naranja se incluirán si los items identificados como clases se utilizan para definir nodos hoja contextualizados.

De igual forma también se han utilizado algunas propiedades para definir los nodos hoja, concretamente:

- P3342: Persona relevante
- P2975: Huésped
- P1876: Nave
- P2176: Medicamento (o droga)
- P5642: Factor de riesgo
- P780: Síntomas
- P1995: Especialidad sanitaria

Los nodos hoja serían los items objeto de las declaraciones que incluyen alguna de las anteriores propiedades.

Durante el procesamiento de cada declaración se procede a analizar los items sujeto y objeto. En el caso de que alguno de ellos se encuentre en el conjunto de items recopilados a partir de los *links* y *backlinks* de *Wikipedia* se procede a marcarlos como prevalidados. La validación final dependerá de la supervisión y oportuna decisión del editor del vocabulario.

El algoritmo básico se compone para cada ítem inicial identificado en la Fase 1 de dos partes:
- creación del grafo correspondiente a los elementos que forman parte de la lista de items iniciales;
- enriquecimiento de cada uno de dichos items.

El más complejo de ellos es el relativo a la creación del grafo cuyo pseudocódigo sería el siguiente:

```
1: PROCEDIMIENTO CREAR_GRAFO (LISTA_ITEMS_INICIALES, NIVEL_MÁXIMO)
2: ÍNDICE = 0
3: NIVEL = 0
4: COLA_DE_PROCESAMIENTO = LISTA_ITEMS_INICIALES
5: Repetir mientras NIVEL ≤ NIVEL_MÁXIMO
```

```
6:      NUMERO_ITEMS = TAMAÑO(COLA_PROCESAMIENTO)
7:      Repetir mientras INDICE ≤ NUMERO_ITEMS
7:          ITEM_ACTUAL = COLA_PROCESAMIENTO(ÍNDICE)
8:          Si ITEM_ACTUAL_SUJETO está en UNA LISTA_ENTIDADES_EXCLUÍDAS
9:          Recuperar declaraciones para ÍTEM_ACTUAL
10:         Para cada declaración recuperada de ÍTEM_ACTUAL
11:             Si (ÍTEM_ACTUAL_PROPIEDAD está en LISTA_PROPIEDADES_INSTANCIA_CLASE) y
                (ÍTEM_ACTUAL_OBJETO está en LISTA_CLASES_EXCLUÍDAS)
12:                 EXCLUIR_ÍTEM = VERDADERO
13:             Si (ÍTEM_ACTUAL_PROPIEDAD está en LISTA_PROPIEDADES_INSTANCIA_CLASE) y
                (ÍTEM_ACTUAL_OBJETO está en LISTA_EXCEPCIONES_CLASES_EXCLUÍDA)
14:                 PROCESAR_ITEM = VERDADERO
15:             Si EXCLUIR_ITEM ≠ Verdadero o PROCESAR_ITEM = VERDADERO
16:                 Para cada declaración recuperada de ITEM_ACTUAL
17:                 SI (ÍTEM_ACTUAL_OBJETO no está en COLA_DE_PROCESAMIENTO) y
                    (ÍTEM_ACTUAL_SUJETO no está en LISTA_NODOS_HOJA)
18:                     Añadir ÍTEM_ACTUAL_OBJETO a COLA_DE_PROCESAMIENTO
19:                 SI (ÍTEM_ACTUAL_SUJETO, ITEM_ACTUAL_PROPIEDAD e ÍTEM_ACTUAL_OBJETO
                    no están en LISTA_ENTIDADES_EXCLUÍDAS)
20:                     Imprimir Identificador(Q) y etiqueta de ÍTEM_ACTUAL_SUJETO
21:                     Imprimir Identificador(P) y etiqueta de ÍTEM_ACTUAL_PROPIEDAD
22:                     Imprimir Identificador(Q) y etiqueta de ÍTEM_ACTUAL_SUJETO
23:             ÍNDICE = ÍNDICE + 1
24:         EN CASO CONTRARIO
25:             Eliminar COLA_DE_PROCESAMIENTO(ÍNDICE)
26:     NIVEL = NIVEL+1
```

El algoritmo tiene en cuenta que no se vuelvan a explorar los items que ya han sido procesados previamente y finaliza cuando se alcanza el nivel máximo de exploración establecido.

3.3. Fase 3: Enriquecimiento

Durante la Fase 2 se realiza una serie de iteraciones que recorren la estructura de *Wikidata* a partir de los items iniciales. Al finalizar dicha fase se dispone de un conjunto de datos con las declaraciones que se han recuperado en función de las condiciones definidas por el editor (nodos hoja, entidades excluidas y niveles de profundidad). No obstante, algunos items no son recopilados durante la exploración debido a que no existe una ruta directa hacia ellos desde los iniciales.

Sin embargo, en *Wikidata* es posible encontrar relaciones hacia los items iniciales de igual modo que existen *backlinks* que apuntan hacia los artículos de *Wikipedia*. Estas declaraciones pueden recuperarse utilizado WDQS mediante la correspondiente consulta Sparql (Anexo I, Consulta 3).

El segundo procedimiento utiliza estas consultas Sparql para realiza el enriquecimiento de las entidades a partir de las declaraciones cuyos items objeto coinciden con los iniciales definidos en la Fase 1.

Este procedimiento también contempla la aplicación de reglas de exclusión de entidades. Tras analizar las declaraciones inicialmente recuperadas para el proceso de enriquecimiento, se identificó una serie de propiedades adicionales para la exclusión de entidades:

- P921: Tema principal de la obra
- P971: Temas asociados por la categoría
- P4224: Categoría contiene
- P509: Causa de muerte
- P1050: Condición médica
- P301: Tema principal de la categoría
- P101: Campo de trabajo
- P4844: Intervención de investigación
- P5137: Elemento para este sentido
- P703: Hallado en el taxón

El motivo para incluir dichas propiedades en las reglas de exclusión se debe a su no pertinencia en el ámbito del vocabulario. Como puede verse son propiedades utilizadas para vincular títulos de artículos de investigación sobre Covid-19 o personas enfermas o fallecidas a causa de la enfermedad. Una vez convenientemente filtrados las declaraciones innecesarias se incorporan al vocabulario.

4. Resultados y discusión

El vocabulario sobre la pandemia de Covid-19 se ha obtenido como resultado de cinco iteraciones. La primera iteración (iteración 0) recupera únicamente las declaraciones de los items iniciales por lo que no realiza ningún proceso de expansión de relaciones. Las cuatro iteraciones siguientes se basan en el mecanismo de exploración descrito anteriormente.

Al finalizar cada iteración se obtiene un conjunto de datos en formato CSV con las declaraciones seleccionadas. Se generan sendos conjuntos de datos (también en formato CSV) con los *links* y *backlinks* utilizados con el marcado propuesto

para la validación de declaraciones. Tal y como se ha indicado anteriormente, cuando el ítem sujeto u objeto de una declaración coincide con el ítem correspondiente en *Wikidata* del *link* o *backlink* extraído de *Wikipedia* se marca como "Propuesto para validación". El conjunto de datos ofrece la siguiente estructura (ver Anexo II):

- iteración en la que se obtuvo la declaración
- ítem sujeto
- etiqueta del ítem sujeto
- ítem objeto
- etiqueta del ítem objeto
- marcado de prevalidación por link
- marcado de prevalidación por backlink.

Las cinco iteraciones se realizaron doblemente. En un primer momento se exploró la estructura de relaciones sin aplicar ningún tipo de regla de exploración (iteraciones previas). Evidentemente, los resultados obtenidos suponían un excesivo número de declaraciones con sus correspondientes items y propiedades. La gran mayoría de ellos eran pocos relevantes para la temática del vocabulario o inadecuados para la finalidad del proceso de exploración. A continuación, se procedió a repetir las iteraciones, pero realizando el análisis necesario para la definición de excepciones y nodos hoja. Los resultados pueden observarse en la tabla 2.

Tabla 2. Resultados obtenidos *en* las cinco iteraciones sin aplicar reglas de exploración (prev) y tras aplicar las reglas de excepciones y definición de nodos hoja.

	Iteración 0	Iteración 1	Iteración 2	Iteración 3	Iteración 4
Declaraciones (prev)	294	29.093	249.817	1.183.000	4.170.559
Items distintos (prev)	267	15.207	110.413	434.114	1.132.774
Propiedades distintas (prev)	36	193	596	854	1.028
Proporción declaraciones por ítem (prev)	1,10	1,91	2,26	2,72	3,68
Declaraciones	279	592	1.053	1.831	2.731
Items distintos	254	454	678	1.036	1.556
Propiedades distintas	28	39	51	59	70
Proporción declaraciones por ítem	1,10	1,30	1,55	1,77	1,76

Puede deducirse fácilmente que la enorme red de relaciones de *Wikidata* influye en la gran cantidad de declaraciones e items recuperados cuando no se aplica ningún tipo de regla. En contraposición, los resultados obtenidos cuando se aplican dichas reglas suponen una reducción considerable del ruido. Un aspecto interesante es la proporción entre declaraciones procesadas y número de items distintos (elementos del vocabulario) obtenidos. De este indicador puede deducirse que la aplicación de las reglas mencionadas además de reducir el ruido incrementa la eficacia del proceso de expansión de relaciones, ya que el número de declaraciones para obtener cada elemento del vocabulario es un 39% menor en la iteración 5.

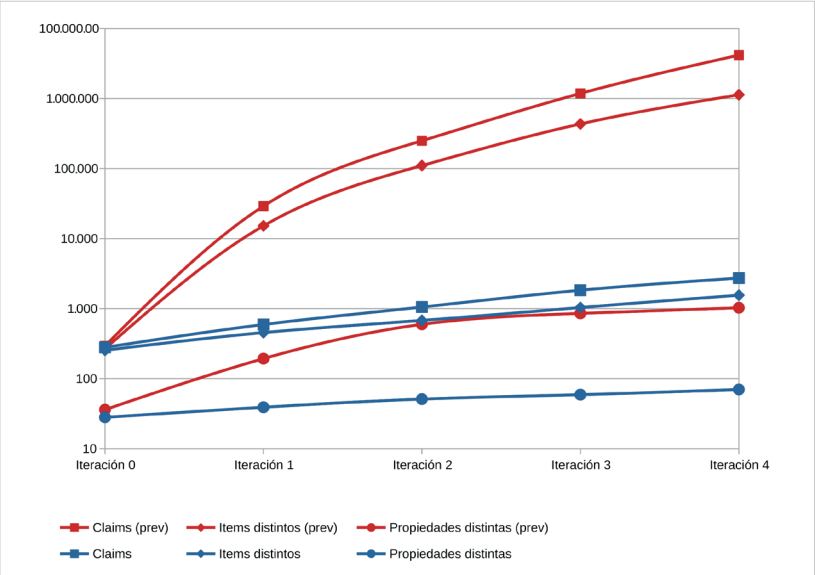

Figura 11. Incremento del número de elementos obtenidos en cada iteración sin aplicar reglas de exploración (prev) y tras su aplicación.

Visualmente y aplicando una escala logarítmica, la progresión en número de declaraciones, items y propiedades en cada iteración se refleja en la figura 11.

Respecto al uso de enlaces (*links* y *backlinks*) de *Wikipedia* debe distinguirse entre las declaraciones propuestas para su validación y los enlaces utilizados para ello. Los datos obtenidos en cada iteración, tras la aplicación de las reglas de exploración, se muestran en la tabla 3.

Tabla 3. Uso de *links* y *backlinks* en el marco de propuestas de validación de declaraciones en cada una de las cinco iteraciones.

	Iteración 0	**Iteración 1**	**Iteración 2**	**Iteración 3**	**Iteración 4**
Total de declaraciones	300	626	913	1.563	2.214
Propuestas de validación	244	419	559	810	878
% propuestas de validación	81,33	66,93	61,22	51,82	39,65
Propuestas de validación por links	239	406	532	761	832
Propuestas de validación por *backlinks*	115	218	307	460	430
Propuestas de validación por links y *backlinks*	110	205	280	411	384
Total links	2.489	2.489	2.489	2.489	2.489
Total *backlinks*	14.118	14.118	14.118	14.118	14.118
Links usados	218	271	305	388	567
Links no usados	2.271	2.218	2.184	2.101	1.922
Backlinks usados	94	135	157	226	404
Backlinks no usados	14.024	13.983	13.960	13.890	13.713

Puede observarse que aunque el número de declaraciones propuestas para su validación crece en cada iteración, su porcentaje sigue la tendencia contraria. Algo lógico considerando que el número de elementos recuperados aumenta durante el proceso de exploración mientras que se mantiene estable el número de *links* y *backlinks* utilizados (los que están presentes en los artículos de *Wikipedia* de los items iniciales). Esto refleja que existe una relación directa de las relaciones entre items de *Wikidata* y los enlaces entre artículos de *Wikipedia*. Dicha relación se debilita según se profundiza en la estructura de *Wikidata*. Sin embargo, esto no es necesariamente un aspecto negativo, sino que puede ser un factor discriminador para la validación de declaraciones.

El número de *links* y *backlinks* utilizados para la propuesta de la validación también se incrementa con el paso de cada iteración. Puede apreciarse que el número de *backlinks* utilizados en proporción al total de estos es muy bajo. Sin embargo, no se trata de un dato significativo, ya que el número de *backlinks* que se manejan es mucho mayor que el de links. En este caso, cada *link* y *backlink* referido a un ítem presente como sujeto u objeto en una declaración resulta de utilidad para el marcado de las propuestas de validación.

Además de los conjuntos de datos en formato CSV la implementación del algoritmo también genera una salida en formato *Graph Exchange XML Format* (gexf). Este formato permite almacenar los resultados en forma de un grafo etiquetado (en este caso dirigido) para que pueda ser analizado, procesado o visualizado posteriormente por cualquier software compatible. Adicionalmente también se elabora automáticamente una representación adaptada para su visualización mediante *WebVOWL* (**Lohmann** *et al.*, 2015). Esta aplicación permite ajustar el nivel de colapso de las relaciones e items de forma que es posible visualizar de forma simplificada la red obtenida durante el proceso de exploración.

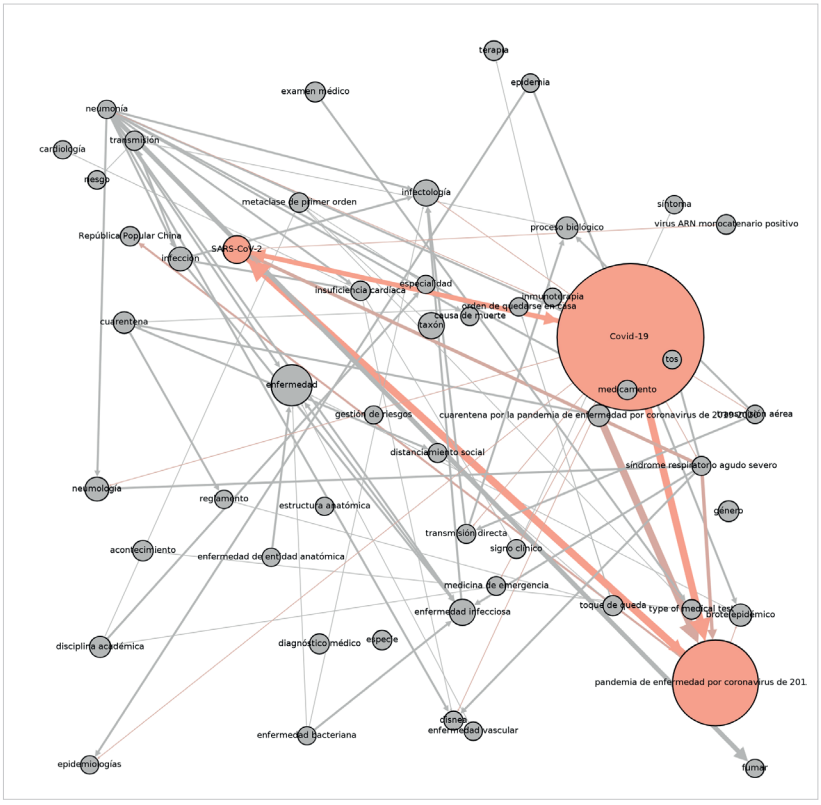

Figura 12. Visualización del grafo resultante mostrando los elementos con un grado de entrada (enlaces entrantes) igual o superior a cinco. Fuente: elaborado mediante *Gephi* (*https://gephi.org*) de exploración (prev) y tras su aplicación.

Un aspecto estudiado durante el desarrollo de esta propuesta es la "skosificación" del vocabulario obtenido. Es decir, su representación mediante SKOS, puesto que se trata del estándar utilizado para la representación de vocabularios controlados y sistemas de organización del conocimiento en el contexto de la web de datos.

De este modo podrían representarse los items como conceptos SKOS (skos:Concept). Las etiquetas podrían

representarse utilizando la propiedad skos:prefLabel, mientras que con las etiquetas alternativas se haría lo propio mediante skos:altLabel. También sería posible representar las descripciones a través de la propiedad skos:description. Sin embargo, el problema sobreviene para utilizar el conjunto de relaciones semánticas de SKOS debido a que no es posible establecer una correspondencia directa entre las propiedades recuperadas durante la exploración de *Wikidata* y las definidas en SKOS para las relaciones jerárquicas y asociativas. Sería necesario realizar un detenido análisis, que no es objeto de este trabajo, para definir el mapeado entre las propiedades de *Wikidata* y las propiedades para las relaciones semánticas de SKOS. Con dicho análisis sería posible la generación de un vocabulario controlado representado mediante SKOS.

También es importante destacar la vertiente de etiqueta alternativa y multilingüe disponible al trabajar con *Wikidata*. El modelo presentado puede fácilmente extenderse, dado el modelo de datos subyacente en *Wikidata*, para producir el mismo vocabulario en diferentes idiomas o etiquetas alternativas (tabla 4). Para cada elemento, entendido como concepto o registro de autoridad, es posible obtener las etiquetas en diferentes idiomas, así como etiquetas alternativas y descripciones.

Tabla 4. Obtención del etiquetado para descripciones y etiquetas alternativas para la entidad: "Q81068910: pandemia de enfermedad por coronavirus de 2019-2020".

Descripciones para "Q81068910: pandemia de enfermedad por coronavirus de 2019-2020"	
Consulta Sparql	**Resultado**

SELECT ?idioma ?itemdesc
WHERE {
VALUES ?item {wd:Q81068910}
?item schema:description ?itemdesc.
BIND(LANG(?itemdesc) as ?idioma)
}
	hu: világjárvány
	it: pandemia virale iniziata a dicembre 2019 nella città di Wuhan, in Cina
	lv: globāla slimības pandēmija, kas sākās Uhaņā, Ķīnā
	ms: pandemik virus yang bermula daripada pasar makanan laut di Wuhan, China
	nb: global pandemi av koronaviruset SARS-CoV-2 som startet i Wuhan, Kina i 2019
	...

Etiquetas alternativas para "Q81068910: pandemia de enfermedad por coronavirus de 2019-2020"	
Consulta Sparql	**Resultado**
SELECT ?item ?alternatelabel WHERE { VALUES ?item {wd:Q81068910} ?item skos:altLabel ?alternatelabel FILTER(LANG(?alternatelabel) = "en") }	2019-2020 SARS-CoV-2 outbreak 2019–20 coronavirus outbreak 2019–20 outbreak of Covid-19 2019–20 outbreak of novel coronavirus (2019-nCoV) 2019–2020 outbreak of Covid-19 2019–2020 Wuhan coronavirus outbreak ...

Fuente: Elaborado desde consulta Sparql en *http://query.wikidata.org*

Mayor trascendencia tiene la explotación de los términos obtenidos desde el punto de vista de los vocabularios de valores vinculados. Dado que en *Wikidata* existe un amplísimo dispositivo de referencias desde sus entidades (items y propiedades) con identificadores externos en bases de datos y registros de autoridad (tabla 5), el mapeo de los términos usados para la indización local con sus correspondientes en *Wikidata* es una práctica recomendada y abre horizontes de posibilidades que pueden explorarse en futuros trabajos.

Tabla 5. Identificadores externos registrados en *Wikidata* para los tres elementos usados en el estudio a través de las propiedades "P486: MeSH descriptor ID", "P244: Library of Congress authority ID" y "P6200: BBC News topic ID".

Entidad de *Wikidata*	***MeSH descriptor ID***	***Library of Congress authority ID***	***BBC News* topic ID**
Q81068910 (Pandemia Covid-19)	*	*	cyz0z8w0ydwt
Q82069695 (SARS-CoV-2)	C000656484	*	*
Q84263196 (Covid-19)	C000657245	sh2020000570	*

5. Conclusiones

Las prácticas y modelos de la indización aplicada en contextos muy dinámicos en la producción de contenidos diversos, como es caso de los medios de comunicación, están sometidas en la actualidad a importantes transformaciones. El modelo planteado en este trabajo ofrece una vía complementaria de explorar la ampliación y enriquecimiento de la terminología usada para organizar los contenidos digitales, y para la actualización continua de los vocabularios controlados construidos para su etiquetado tanto para la gestión del archivo interno como para la publicación web. En el caso de la publicación web, la incorporación del marcado semántico con conceptos y términos relevantes colabora en la mejora del posicionamiento y experiencia de descubrimiento de contenidos.

El mapa o red de conceptos generado a partir del trabajo colaborativo de los editores de *Wikipedia* y *Wikidata*, e implementado de forma accesible nos abre un abanico de posibilidades para el trabajo con el control de vocabularios para la indización de contenidos. No obstante, sigue exigiendo de la intervención de los profesionales de cada plataforma de contenidos para integrar los conceptos sugeridos en un tesauro o esquema de conceptos.

La formalización del aparato de etiquetado y relaciones debería normalizarse conforme al modelo SKOS, permitiendo la integración sencilla con otros sistemas de control de vocabulario. En futuros trabajos se integrará al algoritmo presentando la "skosificación" de etiquetas y, sobre todo, de las relaciones entre conceptos, así como el establecimiento de correspondencias entre recursos, uno de los mecanismos que generan conexiones externas valiosísimas en la web de datos enlazados.

Llevamos años instalados en el denominado "*metadata bypass*" (**Gartner**, 2016, pp. 93-96), un paradigma del acceso a la información en el que el papel central lo cumple el tratamiento automatizado masivo de la documentación para su recuperación. Sin embargo, no se nos puede ocultar que la aplicación de vocabularios controlados para aportar sentido a los contenidos digitales es uno de los factores que entran en juego en casi todos los sistemas de descubrimiento y recuperación de información. Un claro ejemplo es el alto valor que poseen para los buscadores los datos estructurados con los que los editores añaden semántica a sus publicaciones en la web. En los sistemas de información para su acceso existe siempre una oscilación entre la organización de información (metadatos, esquemas de conceptos, descripción de recursos) y la recuperación de información (búsqueda, algoritmos de ranking, filtrado colaborativo): cada uno de los extremos del sistema se retroalimenta de la riqueza incorporada en el otro, siendo esta una de las ideas fundamentales de lo que en el fondo constituye la disciplina de *organizing* (**Glushko**, 2016).

El trabajo de documentación, pese a las numerosísimas transformaciones derivadas de la incorporación de técnicas de inteligencia artificial y procesamiento masivo de datos, sigue aportando valor a la cadena de la información mediante formas más interoperables y heterodoxas de vocabularios controlados y esquemas de conceptos. Por eso, la capacidad de construir conjuntos terminológicos para la organización de contenidos digitales sigue siendo un ámbito crítico en la calidad de los sistemas de información. Nuestra propuesta busca incorporar la aportación en terminología y organización de conceptos que desde plataformas basadas en la inteligencia colectiva se puede obtener con bajos costes de procesamiento. En el caso de *Wikidata* y *Wikipedia*, la disponibilidad de una fuerte estructura de relaciones y enlaces, respectivamente, las hace muy valiosas como apoyo para la construcción de vocabularios controlados que necesitan disponer de capacidad de respuesta a nuevos temas, lo cual es característico del campo de los medios de comunicación, donde la actualidad se regenera continuamente y nuevos temas, entidades y puntos de vista emergen.

Existen además numerosas vías a explorar de la relación entre vocabularios controlados y *Wikidata* como "ontología de contexto". A través de la vinculación de los términos locales con los términos en *Wikidata*, puede enriquecerse la capacidad de interpretación de los datos y ampliarse considerablemente la capacidad de descubrimiento de recursos de información relevantes, patrones, agrupaciones y otras relaciones indirectas. En este trabajo nos hemos centrado en mostrar y aplicar una nueva fuente terminológica para la construcción o enriquecimiento de vocabularios controlados que, desde el punto de vista de los procesos convencionales de selección de terminología para un tesauro, ocupa una posición intermedia entre el análisis de corpus y el análisis de vocabularios. Creemos que es precisamente en los espacios de conexión entre diferentes áreas, donde existen numerosas posibilidades de innovación para un mejor tratamiento de la información.

Anexo 1. Consultas Sparql

Consulta Sparql 1: Utilizada para generar el grafo de la figura 7

```
SELECT distinct ?node ?nodeLabel ?childNode ?childNodeLabel ?rgb WHERE {
 Values ?node { wd:Q82069695 wd:Q84263196} # Pandemia COVID-19 2019-2020; SARS-CoV-2; COVID-19 disease
 ?node ?p ?i.
 ?childNode ?x ?p.
 ?childNode rdf:type wikibase:Property.
 FILTER(STRSTARTS(STR(?i), "http://www.wikidata.org/entity/Q"))
 FILTER(STRSTARTS(STR(?childNode), "http://www.wikidata.org/entity/P"))
 MINUS { ?node wdt:P31/wdt:P279* wd:Q51118821} #Exclusión items internos proyectos Wikimedia
 MINUS { ?childNode wdt:P31/wdt:P279* wd:Q51118821} #Exclusión relac.internas proyectos Wikimedia
 MINUS { ?childNode wdt:P31/wdt:P279* wd:Q15138389} #Exclusión relac. internas artículos Wikimedia
 SERVICE wikibase:label { bd:serviceParam wikibase:language "es,en". } }
```

Consulta Sparql 2: Utilizada para obtener el conjunto de propiedades utilizadas en *Wikidata*

```
SELECT DISTINCT ?property ?propertyLabel
 WHERE {
 { ?property wdt:P31 wd:Q18616576 . }
 SERVICE wikibase:label {bd:serviceParam wikibase:language "es,en".}
 } order by ?prop
```

Consulta Sparql 3: Utilizada para obtener el conjunto de declaraciones que apuntan hacia el ítem "Q84263196: pandemia de enfermedad por coronavirus de 2019-2020"

```
SELECT ?subject ?subjectLabel ?prop ?propLabel ?object ?objectLabel WHERE {
 BIND(wd:Q84263196 AS ?object)
 ?subject ?p ?object.
 ?prop wikibase:directClaim ?p .
 FILTER(STRSTARTS(STR(?object), "http://www.wikidata.org/entity/Q"))
 SERVICE wikibase:label {bd:serviceParam wikibase:language "es,en".}
```

Anexo 2. Extracto del conjunto de datos obtenido

Iteración	Ítem sujeto	Etiqueta ítem sujeto	Propiedad	Etiqueta propiedad	Ítem objeto	Etiqueta ítem objeto	*Link*	*Backlink*
0	Q84263196	Covid-19	P1542	causa de	Q344873	síndrome de dificultad respiratoria aguda	X	X
1	Q344873	síndrome de dificultad respiratoria aguda	P31	instancia de	Q12136	enfermedad		
1	Q344873	síndrome de dificultad respiratoria aguda	P31	instancia de	Q1931388	causa de muerte		
1	Q344873	síndrome de dificultad respiratoria aguda	P279	subclase de	Q767485	insuficiencia respiratoria	X	X
2	Q767485	insuficiencia respiratoria	P1995	especialidad sanitaria	Q203337	neumología	X	
2	Q767485	insuficiencia respiratoria	P31	instancia de	Q12136	enfermedad		
2	Q767485	insuficiencia respiratoria	P31	instancia de	Q1441305	signo clínico		
2	Q767485	insuficiencia respiratoria	P31	instancia de	Q1931388	causa de muerte		
2	Q767485	insuficiencia respiratoria	P279	subclase de	Q3392853	enfermedad pulmonar		
3	Q3392853	enfermedad pulmonar	P31	instancia de	Q12136	enfermedad		
3	Q3392853	enfermedad pulmonar	P279	subclase de	Q18553224	enfermedad del tracto respiratorio inferior		
4	Q18553224	enfermedad del tracto respiratorio inferior	P31	instancia de	Q12136	enfermedad		
4	Q18553224	enfermedad del tracto respiratorio inferior	P279	subclase de	Q3286546	enfermedades del aparato respiratorio	X	X
4	Q18553224	enfermedad del tracto respiratorio inferior	P461	opuesto a	Q18558209	enfermedad del tracto respiratorio superior		
4	Q18553224	enfermedad del tracto respiratorio inferior	P689	afecta	Q2859739	tracto respiratorio inferior	X	
4	Q18553224	enfermedad del tracto respiratorio inferior	P927	localización anatómica	Q2859739	tracto respiratorio inferior	X	
4	Q18553224	enfermedad del tracto respiratorio inferior	P1995	especialidad sanitaria	Q203337	neumología	X	
4	Q18553224	enfermedad del tracto respiratorio inferior	P5642	factor de riesgo	Q662860	fumar		
3	Q3392853	enfermedad pulmonar	P689	afecta	Q5	ser humano		
3	Q3392853	enfermedad pulmonar	P927	localización anatómica	Q2640512	pulmón humano		
3	Q3392853	enfermedad pulmonar	P1995	especialidad sanitaria	Q203337	neumología	X	

6. Referencias

Baeza-Yates, Ricardo; **Ribeiro-Neto, Berthier** (2011). *Modern information retrieval: the concepts and technology behind search*. Harlow, Essex: Addison-Wesley. ISBN: 978 0 321 41691 9

Balog, Krisztian (2018). *Entity-oriented search*. Cham, Switzerland: Springer Nature. ISBN: 978 3 319 93935 3
https://doi.org/10.1007/978-3-319-93935-3

Broughton, Vanda (2006). *Essential thesaurus construction*. London: Facet Publishing. ISBN: 978 1 856045650

El país (2017). "Así es el árbol del conocimiento de El país". *El país que hacemos*, 24 enero.
https://blogs.elpais.com/que-hacemos/2017/01/así-es-el-árbol-del-conocimiento-de-el-país-.html

Fensel, Dieter; **Simsek, Umutcan**; **Angele, Kevin**; **Huaman, Elwin**; **Kärle, Elias**; **Panasiuk, Oleksandra**; **Toma, Ioan**; **Umbrich, Jürgen**; **Wahler, Alexander** (2020). *Knowledge graphs: methodology, tools and selected use cases*. Cham, Switzerland: Springer. ISBN: 978 3 030374389
https://doi.org/10.1007/978-3-030-37439-6

Galloway, Scott (2017). *Four: El ADN secreto de Amazon, Apple, Facebook y Google*. Barcelona: Penguin Random House. ISBN: 978 84 16883271

García-Jiménez, Antonio; **Rodríguez-Mateos, David**; **Catalina-García, Beatriz** (2019). "Estudio sobre la indización/etiquetado y los lenguajes documentales en cinco diarios españoles". *Scire*, v. 25, n. 1, pp. 55-64.
https://www.ibersid.eu/ojs/index.php/scire/article/view/4579

Gartner, Richard (2016). *Metadata: shaping knowledge from antiquity to the semantic web*. Cham, Switzerland: Springer. ISBN: 978 3 319 40893 4
https://doi.org/10.1007/978-3-319-40893-4

GlobalWebIndex (2020). *Coronavirus research. Series 4: Media consumption and sport.*
https://www.globalwebindex.com/hubfs/1.%20Coronavirus%20Research%20PDFs/GWI%20coronavirus%20findings%20April%202020%20-%20Media%20Consumption%20(Release%204).pdf

Glushko, Robert J. (2016). *The discipline of organizing: professional edition*. Sebastopol, CA, USA: O'Reilly Media. ISBN: 978 1 491970614

Hedden, Heather (2016). *The accidental taxonomist*. Medford, NJ, USA: Information Today. ISBN: 978 1 57387 528 8

Lambe, Peter (2007). *Organising knowledge: taxonomies, knowledge and organizational effectiveness.* Oxford: Chandos Publishing. ISBN: 1 84334 227 8

Lohmann, Steffen; **Link, Vincent**; **Marbach, Eduard**; **Negru, Stefan** (2015). "WebVOWL: web-based visualization of ontologies". In: *EKAW 2014. Knowledge engineering and knowledge management. satellite events*, *LNAI 8982*, pp. 154-158. ISBN: 978 3 319 17966 7
https://doi.org/10.1007/978-3-319-17966-7_21

Minguillón, Julià; **Lerga, Maura**; **Aibar, Eduard**; **Lladós-Masllorens, Josep**; **Meseguer-Artola, Antoni** (2017). "Semi-automatic generation of a corpus of *Wikipedia* articles on science and technology". *El profesional de la información*, v. 26, n. 5, pp. 995-1004.
https://doi.org/10.3145/epi.2017.sep.20

Nouvel, Damien; **Ehrmann, Maud**; **Rosset, Sophie** (2016). *Named entities for computational linguistics*. Hoboken, NJ, USA: John Wiley & Sons. ISBN: 978 1 119268567
https://doi.org/10.1002/9781119268567

Pérez-Montoro, Mario; **Codina, Lluís** (2017). *Navigation design and SEO for content intensive websites: a guide for an efficiente digital communication*. Cambridge, MA, USA: Elsevier. ISBN: 978 0 08 100677 1

Piscopo, Alessandro; **Simperl, Elena** (2018). "Who models the world? Collaborative ontology creation and user roles in Wikidata". In: *ACM on Human computer interaction 2*, *CSCW*, article n. 141.
https://doi.org/10.1145/3274410

Rubio-Lacoba, María (2007). *Documentación informativa en el periodismo digital*. Madrid: Síntesis. ISBN: 978 84 97564595

Rubio-Lacoba, María (2012). "Nuevas destrezas documentales para periodistas: el vocabulario colaborativo del diario *El país*". *Trípodos*, n. 31, pp. 65-78.
http://www.tripodos.com/index.php/Facultat_Comunicacio_Blanquerna/article/view/38

Saorín, Tomás (2017). "Wikipedismo de actualidad. La enciclopedia escrita desde el periodismo". *Anuario ThinkEPI*, v. 11, pp. 191-199.
https://doi.org/10.3145/thinkepi.2017.35

Saorín, Tomás; **Pastor-Sánchez, Juan-Antonio** (2018). "*Wikidata* y *DBpedia*: viaje al centro de la web de datos". *Anuario ThinkEPI*, v. 12, pp. 207-214.
https://doi.org/10.3145/thinkepi.2018.31

Sinclair, Lucy (2020). *Insights de búsquedas para ayudarte a entender las necesidades de los consumidores en momentos de incertidumbre* (edición 20 abril 2020).
https://www.thinkwithgoogle.com/intl/es-es/insights/insights-de-busquedas-para-ayudarte-a-entender-las-necesidades-de-los-consumidores-en-momentos-de-incertidumbre-edicion-del-20-de-abril-de-2020

Stuart, David (2016). *Practical ontologies for information professionals*. London: Facet Publishing. ISBN: 978 1 78330 152 2
https://doi.org/10.29085/9781783301522

Suárez-Figueroa, Mari-Carmen; **Gómez-Pérez, Asunción**; **Motta, Enrico**; **Gangemi, Aldo** (2012). *Ontology engineering in a networked world*. Berlin: Springer. ISBN: 978 3 642 43235 4
https://doi.org/10.1007/978-3-642-24794-1