Enriquecimiento de entidades de Wikidata mediante un modelo de descomposición y mapeado de categorías de Wikipedia

Tomás Saorín[1][0000-0001-9448-0866], Juan-Antonio Pastor-Sánchez[2][0000-0002-1677-1059]

[1] Departamento de Información y Documentación, Universidad de Murcia, España.
tsp@um.es
[2] Departamento de Información y Documentación, Universidad de Murcia, España.
pastor@um.es

Resumen. El objetivo de este trabajo es explorar la relación entre las categorías asignadas a los artículos de Wikipedia con la descripción y metadatos generados en Wikidata. Se plantea utilizar la categorización de artículos de Wikipedia para enriquecer la descripción de entidades en Wikidata. Para ello se propone procesar los literales de las categorías mediante técnicas de procesamiento de lenguaje natural (PLN) estableciendo patrones que permitan identificar tanto propiedades como entidades o valores con los que construir declaraciones para una entidad. La secuencia de operaciones propuesta sería el siguiente: 1) Selección de un conjunto coherente de categorías, 2) Establecimiento de patrones de procesamiento de literales y asignación a propiedades y elementos de Wikidata, 3) Creación de declaraciones con cualificadores para cada categoría procesada y 4) Programación de bots para el procesamiento automático de categorías, enriquecimiento y validación de las descripciones de elementos de Wikidata. La propuesta recogida en este trabajo se centra en el uso de diferentes propiedades y entidades de Wikidata para desarrollar el punto 3. La automatización de un proceso para enriquecer y validar las declaraciones de cada elemento, permitiría aprovechar las dinámicas de edición existentes. Además ayudaría a elaborar un esquema de conceptos de más calidad, al especificarse el significado de las categorías que suponen una composición de varios términos y que en realidad resuelven necesidades descriptivas por otros medios.

Palabras clave: Wikipedia, Wikidata, Categories, Named-entity recognition, Knowledge organization

Abstract. This paper explores the relationship between the categories assigned to the Wikipedia articles with the description and metadata generated in Wikidata. It is proposed to use the categorization of Wikipedia articles to enrich the description of entities in Wikidata. For this, the literals of the categories could be processed using natural language processing techniques (NLP), establishing patterns that allow the identification of properties as well as entities or values with which to construct statements for an entity. The sequence of operations would be the following: 1) Selection of a coherent set of categories, 2) Establishment of patterns of processing of literals and assignment to properties and

elements of Wikidata, 3) Creation of declarations with qualifiers for each category processed and 4) Programming of bots for the automatic processing of categories, enrichment and validation of Wikidata element descriptions. The technique shown in this paper focuses on the use of different properties and entities of Wikidata to develop point 3. The automation of a process to enrich and validate the declarations of each element, would allow to reuse existing edition dynamics. It would also help to elaborate a scheme of concepts of higher quality, by specifying the meaning of the categories that suppose a composition of several terms and that actually solve descriptive needs by other means.

Keywords: Wikipedia, Wikidata, Categorías, Reconocimiento de entidades nombradas, Organización del Conocimiento

1 Introducción

Wikidata, el grafo de conocimiento vinculado a los proyectos Wikimedia, está suscitando un intenso interés en las comunidades sectoriales y en las investigaciones académicas, al igual que sucediera con Wikipedia (Saorín; Pastor-Sánchez, 2018). Uno de los aspectos que interesan a la comunidad GLAM, es el de la interconexión de vocabularios controlados, en donde Wikidata actuaría como un "hub of hubs" para los sistemas de control de autoridades y de terminología (Association of Research Libraries, 2019).

Una de las primeras fases de Wikidata fue la de integrar en un único elemento todos los enlaces a las diferentes versiones en cada idioma de los artículos. De este modo se construye un único elemento, con un identificador único y una URI, que representa el concepto, objeto o hecho, independientemente del idioma: cada artículo de Wikipedia se vincula con un concepto que actúa como enlace interlingüístico. La "Guerra de los 30 años" se concreta en el Q2487 , accesible desde la URI https://www.wikidata.org/wiki/Q2487, y no ya como 113 artículos en diferentes Wikipedias.

Complementariamente se realizaron procesos automatizados (bots) para extraer del texto de los artículos declaraciones válidas, de forma análoga al procesado de Dbpedia, aprovechando sobre todo la información semiestructurada en infoboxes. Cada entidad es descrita como una instancia de algo, y se le añaden datos factuales como "fecha de nacimiento", "coordenadas", "autor", "premios recibidos", etc. según la naturaleza del contenido. Estas declaraciones se realizan a través de propiedades (P) que constituyen el modelo de descripción que emerge de Wikidata.

En este trabajo queremos explorar la relación entre las categorías asignadas a los artículos de Wikipedia con la descripción y metadatos generados en Wikidata. Paralelamente a los artículos – que son el contenido principal – existe en la enciclopedia un amplio entramado de categorías, con una función secundaria de navegación y agrupación de contenidos. Las categorías suponen un tipo especial de contenido dentro de Wikipedia: son meramente términos o conceptos, y son incorporadas a Wikidata, con ciertas peculiaridades que abordaremos en este trabajo. Sobre las categorías se están desarrollando líneas de investigación, con especial énfasis en su uso para mejorar para el procesamiento de corpus documentales externos (Tramullas, 2018).

Las categorías de Wikipedia representan un esfuerzo singular en la estructuración del conocimiento histórico, cultural, científico y factual, realizado de forma colaborativa y masiva. La comunidad de editores, al mismo tiempo que elabora contenidos (artículos), trabaja en su organización conforme a un esquema de conceptos (categorías), el cual evoluciona orgánicamente junto al contenido. Es decir, no existe una planificación previa sobre la estructura de categorías, sino que se crea conforme se elaboran nuevos artículos, y están a cargo de lo que se denomina "accidental taxonomist" (Heder, 2016); personas no expertas en organización del conocimiento y cuya dedicación a la construcción de la taxonomía es parcial e irregular.

La funcionalidad de categorías en el software MediaWiki de gestión de contenidos se incorporó a en 2004 a la Wikipedia en inglés, a los tres años de rodaje del proyecto. Proporciona un índice automático de contenidos y adopta una forma esencialmente pseudojeárquica, permitiendo establecer una forma laxa de taxonomía o, más frecuentemente, un grafo. La cantidad de categorías usadas en Wikipedia es excepcional: en mayo-junio de 2018, en Wikipedia en español (eswiki) hay 403.093 páginas de categorías para 1.421.034 de artículos, y en inglés son 1.711.155 para 5.669.291 artículos. A diferencia de otros sistemas de clasificación usados en colecciones documentales (bases de datos bibliográficas, por ejemplo), donde existe un conjunto de conceptos restringido que se usan para indizar documentos primarios, y que tiende a mantenerse estable, y que crecen lentamente conforme se detectan nuevos temas a los que prestar atención, en Wikipedia las categorías están en continuo crecimiento: los artículos (el contenido) y las categorías (el etiquetado) crece al mismo tiempo: hay una ratio de unas 0,58 categorías por artículo, con variaciones según idioma. Esta circunstancia se debe en buena parte a que los contenidos de Wikipedia son, hasta cierto punto, un propio esquema de conocimiento, al ser una enciclopedia. En una base de datos puede haber miles de artículos sobre el hidrógeno, pero en la enciclopedia hay uno solo.

Las categorías de Wikipedia no son en realidad un esquema de clasificación o un vocabulario controlado para describir el contenido, sino para facilitar la navegación entre elementos relacionados. Si las entendemos como una forma de taxonomía, hemos de tener en cuenta que en el medio digital, el uso de las taxonomías para organizar conocimiento supera la limitada tarea de indexar colecciones documentales, sirviendo para hacer funcionar bancos de datos de imágenes, organizar la navegación web, potenciar el descubrimiento de contenidos a través de la búsqueda facetada en tiendas en línea, definir flujos de trabajo en un sistema de información corporativo o para describir las relaciones e interacciones entre usuarios en los medios sociales (Lambe, 2007).

Por otra parte, cuando se asignan categorías a una página de categoría, se va construyendo sobre la marcha una jerarquía de conceptos. Las páginas wiki de categorías nos informan tanto de los artículos o páginas categorizadas (P) con ellas, como de sus subcategorías (C).

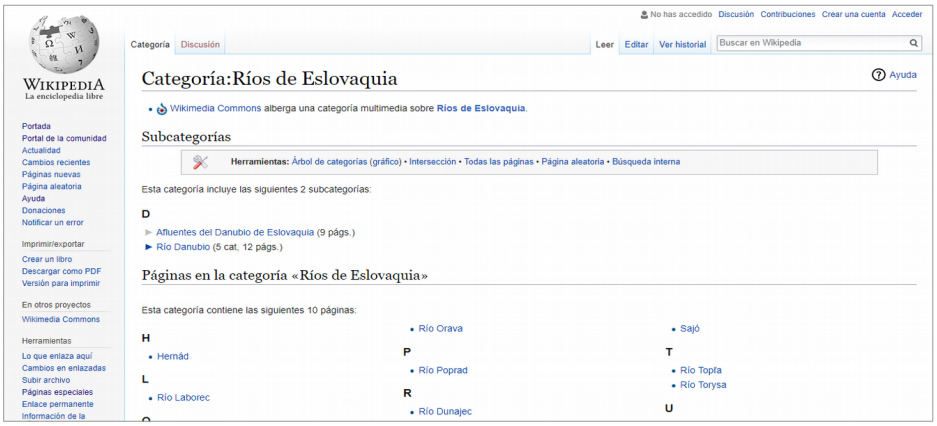

Figura 1. Ejemplo de página de la categoría “Ríos de Eslovaquia” de la edición en Español de Wikipedia donde pueden observarse tanto las subcategorías como las páginas clasificadas bajo dicha categoría. Fuente: https://es.wikipedia.org/wiki/Categoría:Ríos_de_Eslovaquia

El sistema de categorías de Wikipedia puede ser entendido como un "organic thesaurus", conforme a la matriz de "infraestructuras de conocimiento" de Lambe (Lambe, 2007, p. 245). Desde el punto de vista de los lenguajes documentales, tomando como referencia la norma UNE-ISO 25964 sobre Tesauros, resulta difícil aceptar que sea estrictamente un vocabulario controlado, puesto que los usuarios pueden crear sobre la marcha las categorías que necesiten (como en una folksonomía). En este sentido las categorías de Wikipedia podrían considerarse un esquema de clasificación enumerativo, pero no finito. Sin embargo, el etiquetado no es libre y la asignación de categorías preexistentes parece predominar frente a la creación de nuevas, dándose numerosas actuaciones de reorganización y mantenimiento del sistema de categorías. No obstante, es un vocabulario estructurado y puede ser usado para la recuperación de información. Su estructura es claramente polijerárquica. Por otro lado, se trata de un vocabulario fuertemente precoordinado.

El sistema de categorías de Wikipedia no puede considerarse un tesauro, puesto que el criterio para la partición jerárquica no mantiene el rigor en todos los casos. Su forma impura se aproxima más a lo que la norma UNE-ISO considera como taxonomía polijerárquica, que resuelven una necesidad principal de facilitar la navegación en un sitio web, son realizadas por no expertos y presentan asimetrías en su multingüismo. Adaptando la propuesta anteriormente citada de Lambe, podríamos acuñar el término “Organic taxonomy”.

Las categorías son registradas, al igual que los artículos, como entidades en Wikidata, asignándoles una tipología específica que las distingue del resto de elementos[1]. Sin embargo, no se recogen en Wikidatata ni la indización ni las relaciones jerárquicas[2]. La misma categoría en diferentes idiomas se unifica en un único elemento Wikidata, dando como resultado (a enero de 2019) 4.169.547

1 Se trata de instancias (P31) de Q4167836 (Wikimedia category) o de alguna de sus 20 subclases.

categorías diferentes en Wikidata, de las cuales sobre un 90% correspondían a las cargadas desde las más de 250 ediciones diferentes de Wikipedia, y el resto a los otros proyectos Wikimedia (Wikisource, Wikiviajes, Wikiquotes, Wikinoticias, etc.)[3]. Nuestro foco de atención son las categorías de Wikipedia, que es de donde procede el grueso de las categorías registradas en Wikidata.

Para la incorporación de las categorías a Wikidata, la comunidad del proyecto tomó la decisión de no incluir la indización, es decir, no reflejar que un artículo ha sido categorizado con determinadas categorías. Tampoco se recogen las relaciones jerárquicas entre categorías. Por el contrario, en el caso de DBpedia, sí que se optó por formalizar la relación artículos-categorías y categorías-categorías, usando, para las relaciones jerárquicas, propiedades SKOS.

Una de las razones que justifican esta decisión es que la comunidad no las considera conocimiento factual, sino operativo para construir el sitio web de Wikipedia, y que la mejor forma de caracterizar los aspectos factuales de un artículo de Wikipedia son las propiedades específicas, tales como autor, fecha de publicación, ubicación, premios recibidos, tipo/instancia, etc. Por este motivo, cualquier dato que se incluya dentro de la etiqueta de una categoría, y que también esté contemplado en una declaración a través de una propiedad, generaría una redundancia innecesaria.

Se advierte, en los debates sobre categorías en Wikipedia y Wikidata, una palpable insatisfacción con ellas, que apuntaría a una posible sustitución futura por mecanismos mejor adaptados a la función que pretenden realizar: listado y agrupación[4].

Es posible reproducir gran parte de las funciones de listado y agrupación de artículos que se realiza ahora mismo través de categorías, mediante consultas dinámicas sencillas usando “regular properties”. Esto es denominado en los debates como “category assignments” y debe entenderse en términos de “tagging”. También se argumenta que un statement de indización de un artículo no puede ser apoyado en una fuente externa, y por lo tanto no es conocimiento factual independiente, sino resultado de la actividad de la comunidad Wikimedia. También se señala el fuerte incremento de los statements para cada ítem, si se incorporaran sus categorías.

> “If the categorisation can be expressed via a property:value statement then we should add that statement, not this vague category statement. If

2 En anteriores trabajos hemos abordado la skosificación de la integración en Wikidata de las categorías (Pastor-Sánchez y Saorín, 2018).

3 Consulta SPARQL para categorías:
SELECT (COUNT(?item) AS ?total) WHERE {?item wdt:P31 wd:Q416783 }
Incluyendo las categorías tipificadas en alguna de las subclases de “Wikimedia category”, la consulta sería:
SELECT (COUNT(distinct ?item) AS ?total) WHERE {
?item wdt:P31 ?tipo. ?tipo wdt:P279* wd:Q4167836 }

4 Los puntos de vista y las citas que extraemos proceden de dos debates sobre las propuestas de “Parent category” y “Category”, archivados en:
https://www.wikidata.org/wiki/Wikidata:Property_proposal/Archive/30#Parent_category
https://www.wikidata.org/wiki/Wikidata:Property_proposal/Archive/30#category

> the categorisation can not be expressed as a statement then we almost certainly don't need it. The categorisation systems on english and other wikipedias certainly does contain a lot of information which can be harvested and converted into useful statements and we should certainly do this conversion even if it means we need to create more properties. For instance we need a "preparation method" property so we can include the statement "Chipotle:preparation method:smoked". The **raw data** however should stay in the wikipedias."

Advertimos que las categorías de la Wikipeda no son aceptadas como algo "factual" (un objeto de la realidad, que deba ser objetivado) sino como una creación adhoc dentro de la Wikipedia. No es aceptado como una clasificación de conocimiento, sino como una estructura de apoyo:

> "WP categories are a kind of classification without global overview. So instead of using an **empirical classification**, just start the work correctly."

> "**Category is no knowledge** but an arbitrary way to sort items that depends on the wikimedia project. It would not be possible to source that kind of statement and it would generate a lot of repetitions with the other statements."

> "WP categories are outdated: categories are a specific way to group articles which **is not common knowledge but particular point of view**."

> "Wikipedia use categories for classification. Wikidata use **something else**"

En cuanto a las relaciones jerárquicas, dado que en cada idioma están establecidas unas relaciones diferentes, y que el contenido declarado en los ítems (tipo, autor, , lugar, estilo, etc.) ya supone una forma de organización de conocimiento, no se refleja la categorización de categorías, que, en un sentido amplio, se corresponderían con la relación Broader Term de un Tesauro.

> "The category system is pointless because there is no unique system, but different ways to classify articles."

> "Transfering Wikipedia's categories' content as it is is not an appropriate to exploit the data. This would generate repetitions and errors and force users to develop a second ontology to define the relation between categories."

> “Not to mention the fact that this is a central point and that there is as many category hierarchy as there is Wikipedia, this would be a mess to centralize everything here”

Además, en ambos casos, se plantea como barrera la ineficiencia de “replicar” o “mimic” estos datos, que son decisiones editoriales de usuarios, en Wikidata.

> “Why do we need to reflect categories on Wikipedias? They will continue to be there and to be maintained, queryable and extractable. Why do we need to mimic them here, **would'nt they be messy and double** the work of maintaining them, a change of some wikipedia would have to be reflected here.”

Este es la situación actual implementada en Wikidata, pero creemos interesante recoger las razones de las posturas que opinan en sentido contrario: Los esquemas de categorías no son accidentes o transacciones automáticas en las wiki, sino el resultado de un esfuerzo constante y consciente por construir una herramienta de organización de ciertos aspectos de conocimiento que se consideran necesarios.

> “Categories are most **assuredly knowledge. A bit messy, lacking in organization (true) but very comprehensive**. Wikipedians are very serious about their categorization […] We also don't source Labels and Descriptions, but do you think you could live without them?”

La explotación de las categorías para generar nuevas declaraciones y nuevas propiedades puede ser una vía de futuro, pero elude la situación actual de Wikipedia en la que las categoriás cumplen una función:

> “It could/might be transferred into properly structured properties in Wikidata...maybe in 10 years time. But I'm not talking what we could have: I'm talking about what is already in Wikipedia but not reflected in Wikidata.”

También la variabilidad entre idiomas se enfoca como un valor y no como una limitación: Al igual que versión de artículo de un idioma diferente refleja una visión especifica de una comunidad, las categorías asignadas y las relaciones entre categorías manifiestan las diversidades que sobre el conocimiento existen entre grupos humanos, que es uno de los puntos fuertes de Wikipedia, compatible con la política del punto de vista neutral.

> “We'll merge the category trees of 11 languages: that's mostly a union, since category interlanguage links are only 0.35x compared to 2.2x for articles. I don't view this as "discrepancies" but as useful local views that complement each other.”

En el momento actual hay varios aspectos operativos de las categorías de Wikipedia que se ha optado por no integrar en Wikidata, por lo que podemos afirmar que están infraformalizadas: parte del esfuerzo editor no fluye hacia la base de conocimientos central. Parece evidente que naturaleza de la gestión de contenidos en Wikipedia, una vez que se ha abierto el camino a la conexión con una "central de datos", implicará abordar más tarde o más temprano que la asignación de categorías a los artículos se realice de forma integrada con Wikidata (tagging), así como la asignación de categorías a las categorías (concept schema). En la actualidad está integración podría simularse mediante el procedimiento habitual de bots que sincronizan datos de forma continua. La fuerza bruta de los bots puede suplir a la lógica de diseño de un sistema robusto de gestión de contenidos multilïngues y multiversión.

Sin embargo, nuestro trabajo parte del supuesto de que, dado que las categorías forman parte del conocimiento codificado presente ya en los artículos de Wikipedia, suponen otra vía complementaria para la extracción de conocimiento valioso. Los editores las están usando para expresar ciertas relaciones o necesidades de organización del conjunto de artículos, y por lo tanto podemos explotar su significado en relación con un artículo.

2 Metodología

El trabajo que planteamos busca aprovechar la categorización de artículos de Wikipedia para enriquecer la descripción de entidades en Wikidata. A partir del uso que se está haciendo de las categorías (listados y navegación) y por su propia naturaleza precoordinada, parece entreverse la posibilidad de extraer de ellas elementos descriptivos que pueden incorporarse como declaraciones. Es sencillo de entender con un ejemplo: si el artículo sobre Albert Einstein usa las categorías "Científicos judíos" y "Ganadores del premio nobel de física en 1921", podemos inferir y formalizar como hechos factuales que Einstein era un científico, que era judío y que ganó el premio nobel de física en un determinado año. Conocimiento de este tipo se manifiesta en las categorías usadas en los artículos, y por lo tanto es aprovechable para el enriquecimiento y validación de declaraciones en Wikidata.

Tabla 1. Comparativa de contenido entre el artículo sobre Albert Einteins en la edición en Español de Wikidata y la entidad correspondiente en Wikidata.

Categorías asignadas al artículo sobre Albert Einstein en la edición en Español de Wikipedia https://es.wikipedia.org/wiki/Albert_Einstein	Declaraciones sobre la entidad sobre Albert Einstein en Wikidata https://www.wikidata.org/wiki/Q937
• Hombres, Nacidos en 1879 • Fallecidos en 1955 • Albert Einstein • Físicos teóricos • Físicos relativistas • Físicos cuánticos • Cosmólogos • Científicos exiliados del nazismo • Nacionalizados de Suiza • Nacionalizados de Estados Unidos • Alumnado de la Escuela Politécnica Federal de Zúrich, Profesores de la Universidad Carolina • Profesores de la Universidad Humboldt de Berlín • etc.	• Instancia de (P31) Humano • Sexo (P21) Masculino • Nacionalidad (P27) Imperio Alemán; Suiza; República de Weimar; USA; Austro-Hungría • Fecha de nacimiento (P569) 14-3-1879 • Ocupación (P106) Físico teórico; Escritor científico; Inventor; Físico; Profesor universtiario… • Campo de trabajo (P101) Física teórica • Miembro de (P463) Roya Society; merican Philosophical Society, … • Premios recibidos (P166) Gold Medal of the Royal Astronomical Society (1926) … • etc.

El modelo de representación de elementos en Wikidata se basa en declaraciones (pares de propiedades y vínculos a otras entidades de Wikidata). Al igual que el modelo RDF permite incluir reificaciones. De esta forma, cada declaración puede ser matizada por otra nueva declaración (cualificadores), apoyada en una fuente externa (referencia) y ponderada (Rank). Esta forma de expandirse o anotarse o contextualizarse permite diferenciar entre el Snak (Propiedad-Valor) y el Claim (Propiedad-Valor-Cualificador).

Tabla 2. Detalle sobre una declaración de la entidad Q937 (Albert Einstein) de Wikidata. Fuente: https://www.wikidata.org/wiki/Q937

Item sujeto	**Propiedad**	**Item objeto**	**Propiedad cualificador**	**Dato**
Albert Eintein (Q937)	Premios recibidos (P166)	Gold Medal of the Royal Astronomical Society (Q753072)	Punto en el tiempo (P585)	1926

Para un determinado subconjunto de categorías en un idioma determinado, podemos procesar los literales mediante técnicas de procesamiento de lenguaje natural (PLN). Para ello, se establecerían patrones que permitan identificar propiedades, entidades o valores con los que construir declaraciones para una entidad. Muchas ramas de categorías presentan patrones reconocibles en los que es viable aplicar técnicas de reconocimiento de entidades (Named Entities Recognition) presentes en la base de

conocimiento de Wikidata, así como reconocer propiedades con las que usarlas. La secuencia de operaciones que se plantea sería la siguiente:

1. Selección de un conjunto de categorías asociadas a un dominio de conocimiento.
2. Establecimiento de patrones de procesamiento de literales y asignación a propiedades y elementos de Wikidata.
3. Creación de declaraciones con cualificadores para cada categoría procesada, que contengan las entidades reconocidas y las propiedades sobre las que aplican.
4. Programación de bots para el procesamiento automático de categorías, enriquecimiento y validación de las descripciones de elementos de Wikidata.

Este trabajo se centra en las operaciones del tercer bloque (3) y, por lo tanto solo comentaremos superficialmente lo relativo al establecimiento del subconjunto de categorías (1) o las técnicas de PNL para establecer patrones (2). El objetivo específico es presentar los mecanismos de formalización en Wikidata de los resultados obtenidos del análisis de componentes de cualquier categoría (3) y en la descripción esquemática del procesamiento por lotes que posteriormente podría realizarse mediante bots y otros procedimientos automatizados (4).

Se utilizarían al máximo los propios recursos de Wikidata como base de conocimiento, para mejorar la propia representación de las categorías y describir los componentes semánticos contenidos en la misma. Lo anterior permitiría formalizar un mecanismo para crear declaraciones en Wikidata a partir de la indización con categorías en Wikipedia.

Sobre el primer bloque (1), la selección de un conjunto de categorías asociadas a un dominio de conocimiento, existen un sinfín de posibilidades. Mediante técnicas de clustering se pueden obtener conjuntos de artículos de temática afín (Minguillón, 2017), desde los que extraer las categorías usadas y construir un conjunto para su procesamiento para la detección de patrones (2). Para el dominio de las obras literarias, una sencilla consulta a las categorías que contienen la palabra “Novelas” nos permite reconocer patrones de los que se pueden derivar con relativa facilidad información descriptiva: autor, fecha, país de publicación, género u otras características. El desarrollo de patrones de lenguaje natural es un campo muy maduro, y su aplicación sobre conjuntos de términos como las categorías es mucho más operativa que sobre textos completos. La concisión y relativa homogeneidad de las expresiones usadas hacen de esta tarea una labor que podemos considerar una “commoditiy”, abordable mediante un conjunto conocido y variable de técnicas, librerías y algoritmos que ofrecerán resultados contrastados.

Tabla 3. Muestra de categorías que contienen el término “Novelas”[5]

Item	Literal
wd:Q6185886	Categoría:Novelas de Daniel Chavarría
wd:Q5882155	Categoría:Novelas de 1851
wd:Q3919902	Categoría:Novelas
wd:Q4064950	Categoría:Novelas de 2001
wd:Q6214588	Categoría:Novelas cortas de Japón
wd:Q6267331	Categoría:Novelas de Dominique Lapierre
wd:Q6285624	Categoría:Novelas de 1786
wd:Q6285939	Categoría:Novelas de Líbano
wd:Q6285951	Categoría:Novelas ligeras adaptadas de anime o manga
wd:Q6285955	Categoría:Novelas polémicas
wd:Q6285962	Categoría:Novelas tragicómicas
wd:Q6285960	Categoría:Novelas premiadas con el Premio Goncourt
wd:Q6285944	Categoría:Novelas ejemplares
wd:Q6285949	Categoría:Novelas fantásticas de Lois McMaster Bujold
wd:Q6285953	Categoría:Novelas pastoriles

3 Resultados

Una vez identificado un patrón de extracción de conceptos de una categoría, lo que queremos destacar en este trabajo que la propia estructura de datos de Wikidata permite codificar los resultados obtenidos. El análisis lingüístico de las categorías puede realizarse con cualquier algoritmo o librerías diseñado ad hoc, pero sus resultados podrían incorporarse al propio registro de Wikidata. Este punto es esencial en nuestro planteamiento, dado que los resultados del trabajo de extracción realizado externamente a la plataforma, a menudo como proyecto de investigación, queda incorporado usando el propio modelo de datos de Wikidata: de este forma, el Item correspondiente a la categoría procesada se enriquecería con las declaraciones de la descomposición de los conceptos extraídos, quedando este conocimiento accesible en el grafo de conocimiento para ser explotado de cualquier forma que posteriormente se encuentre oportuna.

Analizaremos el encaje en nuestra prueba de concepto de las propiedades actualmente disponibles en Wikidata (más de 9000). Para nuestro propósito hemos seleccionado las propiedades P971, P4224 y P1687, que se definen del siguiente modo:

5 Estos datos del Wikidata Query Service (https://query.wikidata.org/) con la siguiente consulta SPARQL:

```
SELECT distinct ?item ?literal WHERE {
  ?item wdt:P31 wd:Q4167836 .
  ?item rdfs:label ?literal
FILTER contains(?literal,"Novelas") }
```

A. La propiedad P971 “Category combine topics” está prevista para describir los conceptos combinados en una categoría específica y actualmente se aplica sobre 648.478 categorías (16% del total de categorías de Wikidata).
B. La propiedad P4224 “category contains”, que indica que la categoría contiene elementos que son instancias de un cierto elemento y se usa sobre 583.296 categorías (14% del total).
C. La propiedad P1687 “propiedad Wikidata” se usa para relacionar un ítem con la principal propiedad con la que está relacionado y se usa en 4.440 items)[6].

(A) El uso de la propiedad P971 nos permite saber que la categoría “Novelas de Dominique Lapierre” (Q6267331) combina los Items “Novela” (Q8261) y “Dominique Lapierre” (Q1238249).

Tabla 4. Ejemplo de identificación de “topics” que definen de forma combinada la entidad de la categoría “Novelas de Dominique Lapierre” de Wikidata. Fuente: https://www.wikidata.org/wiki/Q6267331

Item sujeto	**Propiedad**	**Item objeto**
Novelas de Dominique Lapierre (Q6267331)	Category combine topics (P971)	Novela (Q8261) Dominique Lapierre (Q1238249)

De este modo, para todos los Items correspondientes a los artículos de Wikipedia indizados mediante dicha categoría se podría comprobar si se han incluído como objeto de alguna declaración los items Q8261 y Q1238249. En caso de no existir, podrían añadirse dichas declaraciones previa identificación de la propiedad más adecuada para vincular los items de los artículos con los items de autor y género. Sin embargo, no conocemos la propiedad a utilizar, puesto que, siguiendo el ejemplo anterior, en la combinación de temas del item Wikidata correspondiente a la categoría Q6267331 no se recoge la vinculación de cada uno de ellos con P50 (autor) o P136 (género literario).

De esta forma simple solo se recogen actualmente los objetos, y no las propiedades con las que pueden ser usados. Como la estructura de Wikidata permite aumentar la expresividad de las declaraciones mediante cualificadores, podríamos usarlos para expresar “este objeto puede usarse con esta propiedad”. Es necesario identificar una propiedad existente actualmente en Wikidata, que permita usarse como cualificador y que tenga un significado compatible con este propósito. Existen unas 180 específicas para ser usadas como cualificadores[7]. De todas ellas se ha optado por la propiedad P642 “Of”, entendida como “Scope of” y que puede usarse para

6 Consulta SPARQL (13-enero-2019): SELECT distinct ?item WHERE { ?item wdt:P971 ?topic } ; SELECT distinct ?item WHERE { ?item wdt:P4224 ?topic } ; Conjuntamente son usadas en 63.283 categorías.

7 Se trata de aquellas definidas con el tipo “Wikidata qualifier” (Q15720608) y que extraemos del WDQS con la siguiente consulta SPARQL: SELECT ?item WHERE { ?item wdt:P31 ?tipo. ?tipo wdt:P279* wd:Q15720608. }

cualificar el ámbito de aplicación una declaración. Esta reificación contempla los elementos necesarios para usarla en la construcción de nuevas declaraciones en los item de Wikidata correspondientes a cualquier artículo de Wikipedia indizado con esa categoría. Idealmente nos encontraríamos con unos datos que descomponen la categoría usado de ejemplo de la siguiente manera:

Tabla 6. Propuesta de descomposición de la entidad Q6267331. Fuente: elaboración propia.

Item sujeto	Propiedad	Item objeto	Cualificador	Cualificación
Novelas de Dominique Lapierre (Q6267331)	Category combine topics (P971)	Novela (Q8261)	Scope of (P642)	Géneros literarios (Q223394)
Propiedad género literario (P136)	subject item of this property (P1629)	Género literario (Q483394)	-	-
Géneros literarios (Q223393)	Wikidata property (P1687)	Propiedad género literario (P136)	-	-

Esta cadena de relaciones podría expresarse como Q6267331 > [P971 > Q8261] P642 > Q223393 > P1629 > Q483394 > P136. Este mecanismo, pese a obtener los resultados finales deseados, implica la combinación de demasiados elementos, y una lógica susceptible de inconsistencias en caso de que no se establezcan relacionas unívocas entre ítems y propiedades, o no existan tales relaciones para una propiedad determinada.

(B) El uso de la propiedad P4224 nos permite saber que la categoría “Novelas de Dominique Lapierre” (Q6267331) se usa en ítems que del tipo - son instancias de (P31) - de “Obra literaria”. Como en la propia definición de la propiedad se incluye que la relación de instanciación, no hay duda sobre el uso potencial para construir statements. Puede usarse para validar o añadir la declaración inicial que todo ítem debe tener.

(C) Por último, la propiedad P1687 nos permite saber que la categoría “Novelas de Dominique Lapierre” (Q6267331) contiene aspectos de autoría y género, al servir para guardar relaciones con propiedades, en este caso P50 y P136. Al igual que sucedía con las declaraciones de P971, nos falta un elemento para abarcar el significado completo que buscamos, es decir, quién es el autor y cuál es el género. En este caso, el uso de los cualificadores sí es eficiente, puesto que la relación a través de la propiedad P642 “of” se establecerá con ítems (Q).

Tabla 7. Propuesta de descomposición de la entidad Q6267331 en propiedades de Wikidata. Fuente: elaboración propia.

Item sujeto	Propiedad	Item objeto	Cualificador	Cualificación
Novelas de Dominique Lapierre (Q6267331)	Propiedad Wikidata (P1687)	Autor (P50)	Scope of (P642)	Dominique Lapierre (Q1238249)
Géneros literarios (Q223393)	Wikidata property (P1687)	Propiedad género literario (P136)	Scope of (P642)	Novela (Q8261)

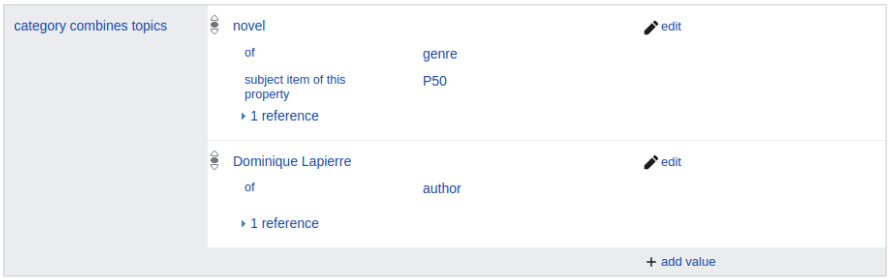

Figura 2. Modificación en Wikidata de la entidad Q6267331 aplicando la propuesta. Fuente: https://www.wikidata.org/wiki/Q6267331

En este ejemplo, para la categoría Q6267331 obtendríamos, a través de una consulta SPARQL, los siguientes pares de elementos y propiedades apropiados para construir declaraciones:

```
SELECT ?property ?propertyLabel ?object ?objectLabel  WHERE
{
  wd:Q6267331 p:P1687 ?statement .
  ?statement ps:P1687 ?property .
  ?statement pq:P642 ?object .
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en"
}
}
```

property	propertyLabel	object	objectLabel
wd:P50	author	wd:Q1238249	Dominique Lapierre
wd:P136	genre	wd:Q8261	novel

Figura 3. Ejemplo de consulta SPARQL y resultado de su ejecución en Wikidata Query Service. Fuente: elaboración propia.

4 Conclusiones

En este marco de convergencia entre metadata y taxonomías (Hedden, 2014) encontramos en Wikidata un caso de lo que Morville denominó como "hybrid metadata ecology" (Morville, 2005). El trabajo de los editores al categorizar artículos se está aportando una forma de descripción de aspectos del contenido (metadatos descriptivos), aunque sin usar un modelo de datos adecuado: las categorías actuales cumplen más un papel descriptivo que de auténtica categorización.

El valor potencial de Wikidata como fuente de datos depende de la calidad de las descripciones de sus entidades, y, por lo tanto, de la cantidad de trabajo que la comunidad de editores sea capaz de movilizar eficientemente. Dado que el apoyo de bots es uno de los factores que complementan la participación de editores humanos en Wikipedia, en el caso de Wikidata se ofrecen aún más posibilidades de enriquecimiento y validación de contenidos, al tratarse de datos formalizados. El trabajo existente en la categorización de artículos en Wikipedia permite ser explotado en dos sentidos diferentes: para añadir nuevas declaraciones a los elementos, y para validar las ya existentes.

Aunque el sistema de categorías de Wikipedia está sometido a discusión y responde a necesidades coyunturales de organización y navegación de contenidos, no puede olvidarse que se trata de un trabajo directo de los editores de la enciclopedia. Este trabajo materializa conocimiento en relación con los artículos, y puede aprovecharse para mejorar y contrastar los datos almacenados en Wikidata. Esto responde al interés primario de la comunidad y respeta la construcción colaborativa de conocimiento por editores responsables.

Las categorías de Wikipedia, en su estado actual, no son el mecanismo óptimo de organización del conocimiento. Sin embargo, su utilidad para aumentar el alcance y capacidad del conjunto de datos semántico de Wikidata ofrece una nueva visión de ellas, como nexo para una transición hacia un grafo de conocimiento exhaustivo de todo lo editado hasta el momento en Wikipedia. De esta manera puede plantearse su desaparición o transformación sin pérdida de todo el trabajo realizado por miles de editores en todos los idiomas. El trabajo en un idioma se generaliza como conocimiento univeral al volcar conocimiento en un grafo común.

Hemos descrito un mecanismo operativo para codificar los conceptos y propiedades implícitos en una categoría, que usa elementos ya disponibles en la infraestructura de Wikidata. Se valida funcionalmente el uso de propiedad P4224 (B) para vincular con ítems a través de la propiedad P31 y de la propiedad P1687 (A) para especificar las propiedades inferidas al procesar los literales de una categoría y, a través de cualificadores vía la propiedad P642, indicar el elemento objeto de la relación. Se descarta, por el contrario, el uso de la propiedad P971 (A) dado que no especificarse las declaraciones con construirse cualificadores que apunten a propiedades. No obstante, para una mayor rigurosidad, se considera recomendable la creación de una nueva propiedad específica denominada “Category combines properties” cuyo dominio sean ítems instancia de “Wikimedia category” y su rango “Properties”.

Estamos ante un proceso automatizable que, a partir de las categorías asignadas a cada artículo en cualquier idioma, enriquecería y validaría las declaraciones de cada elemento. Sea cual sea nuestra opinión de las categorías, la realidad es que se miles de editores las vienen usando en los artículos de Wikipedia, y que por lo tanto se puede aprovechar una dinámica de edición que ya existe para extraer un valor secundario de ella de forma automatizada. Al mismo tiempo, serviría para obtener un esquema de conceptos de más calidad, al especificarse las unidades de significado de las categorías que suponen una composición de varios conceptos. En caso de una futura sustitución de las categorías por otros mecanismos más óptimos de navegación y agrupación de artículos, el conocimiento generado el usarlas para etiquetar artículos se habría reutilizado para ampliar y validar el grado de conocimiento de Wikidata.

Referencias

Association of Research Libraries (2019) “ARL White Paper on Wikidata: Opportunities and Recommendations”, abril de 2019

Hedden, H. (2016). The accidental taxonomist (2a). Information Today.

Lambe, P. (2007). Organising knowledge: taxonomies, knowledge and organizational effectiveness. Oxford: Chandos Publishing.

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. Recuperado de: https://doi.org/10.3145/epi.2017.sep.20.

Morville, Peter (2005) Ambient findability: What We Find Changes Who We Become. Sebastopol: O’Reilly.

Pastor-Sánchez, J.A. & Saorín, T. (2018) Proposal for the integration of the semantic structure of Wikipedia categories using SKOS. 15th International ISKO Conference (OPorto, 2018). Recuperado de: http://hdl.handle.net/10760/38627

Saorín, T. & Pastor-Sánchez, J.A (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

Tramullas, Jesús; Sánchez Casabón, Ana-Isabel; Garrido, Piedad. Wikipedia categories in research: towards a qualitative review of uses and applications. En: in: Fernanda Ribeiro, Maria Elisa Cerveira (Coords.) Challenges and Opportunities for Knowledge Organization in the Digital Age, pp. 490 – 498, Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal