Revista Española de Documentación Científica
46(3), julio-septiembre 2023, e366
ISSN-L:0210-0614. https://doi.org/10.3989/redc.2023.3.2013

ESTUDIOS / *RESEARCH STUDIES*

Un canon literario universal basado en datos enciclopédicos multilingües: propuesta de un método de medición de obras literarias usando datos cuantitativos obtenidos de Wikidata y Wikipedia

Juan Antonio Pastor-Sánchez*, Tomás Saorín**, María-José Baños-Moreno***

Departamento de información y documentación. Universidad de Murcia
*Correo-e: pastor@um.es es | ORCID iD: https://orcid.org/0000-0002-1677-1059
**Correo-e: tsp@um.es | ORCID iD: https://orcid.org/0000-0001-9448-0866
***Correo-e: mbm41963@um.es | ORCID iD: https://orcid.org/0000-0001-9137-1330

Recibido: 20-09-22; 2ª versión: 10-11-22; Aceptado 16-11-22; Publicado: 07-07-23

Cómo citar este artículo/Citation: Pastor-Sánchez, J. A., Saorín, T., Baños-Moreno Odilo, M.J. (2023). Un canon literario universal basado en datos enciclopédicos multilingües: propuesta de un método de medición de obras literarias usando datos cuantitativos obtenidos de Wikidata y Wikipedia. *Revista Española de Documentación Científica*, 46 (3), e366. https://doi.org/10.3989/redc.2023.3.2013

Resumen: La investigación descrita en este artículo tiene como objetivo verificar la viabilidad de usar Wikidata y Wikipedia como fuente para identificar un canon literario universal. Para ello, primero se sitúan ambos proyectos de la fundación Wikimedia en el contexto de los datos sobre obras literarias. La metodología utilizada se basa en la construcción de un conjunto de datos a partir de datos específicos sobre obras literarias recuperados de Wikidata y de las ediciones de Wikipedia en todos los idiomas. Se analiza la profundidad de descripción de los ítems de obras literarias en Wikidata y su presencia y nivel de elaboración de los correspondientes artículos en Wikipedia. Mediante K-means se identifican tres clústeres de obras literarias a partir de los cuales se identifican un conjunto de obras que pueden conformar un canon literario universal. Se propone una métrica denominada Wiki3DRank que permite seleccionar y ordenar las obras literarias analizadas. El estudio aborda también algunos aspectos de la distribución por idiomas, así como un análisis específico de las asimetrías en su distribución temporal entre obras clásicas y obras contemporáneas. El artículo incluye una sección de discusión con reflexiones sobre los resultados obtenidos y concluye proponiendo Wikidata y Wikipedia como una fuente complementaria valiosa para la elaboración de cánones literarios tanto globales como de idiomas específicos.

Palabras clave: Canon literario; obras literarias; Wikidata; Wikipedia; Wiki3DRank.

A universal literary canon based on multilingual encyclopedic data: Proposal of a method for the ranking of literary works using quantitative data obtained from Wikidata and Wikipedia

Abstract: The research described in this article aims to verify the use of Wikidata and Wikipedia as a source to identify a universal literary canon. Both Wikimedia Foundation projects are placed in the context of data on literary works. The methodology used is based on the construction of a dataset from specific data on literary works retrieved from Wikidata and Wikipedia editions in all languages. The depth of description of the items of literary works in Wikidata and their presence and level of elaboration of the corresponding articles in Wikipedia are analyzed. The authors use K-means to define three clusters of literary works that allow the identification of a set of works that can be used to create a universal literary canon. Wiki3DRank is proposed as a metric that allows the literary works analyzed to be selected and ranked. The study deals with the analysis of the language of literary works and their presence in Wikipedia, their temporal distribution. The article includes a discussion section with reflections on the results obtained and concludes with the proposal to use Wikidata and Wikipedia as an alternative source for the elaboration of both global and language-specific literary canons.

Keywords: Literary canon; literary works; Wikidata; Wikipedia; Wiki3Drank, Ranking.

1. INTRODUCCIÓN

Este trabajo parte de una pregunta en principio muy sencilla: ¿podrían usarse Wikidata y Wikipedia como fuente para identificar un canon literario universal? El canon literario es entendido como una selección cultural fuertemente afectada por el punto de vista del grupo de poder que lo establece. Por lo tanto, está sometido a contestación desde las posiciones diferentes que han emergido desde periferias geográficas, identitarias y culturales, que buscan ampliar la visión del canon literario occidental popularizado por el crítico literario Harold Bloom, o el presente en los libros de texto escolares y en los programas de estudios superiores. Además, cualquier canon tomado como referencia, no es inmutable y está sujeto a un interminable proceso de atención, olvido y recuperación a lo largo de siglos, épocas y décadas. Al ser el canon una construcción cultural cambiante ¿Podría usarse la actividad autónoma y no planificada de la comunidad de editores de Wikidata y Wikipedia para obtener otro punto de vista complementario? Son comunidades implicadas en la redacción y categorización de artículos en todos los idiomas y en la definición de datos descriptivos de todo tipo. Apoyados en la idea del punto de vista neutral, trabajo descentralizado y multilingüe, el ecosistema Wikimedia podría ser un candidato para poder obtener resultados no mediados directamente por ningún autor, academia, nación o grupo de interés.

Los estudios sobre cobertura temática en Wikipedia han girado sobre diversos campos, como el de la ciencia, las biografías, patrimonio cultural, cultura de masas o la actualidad social (Hill y Shaw, 2020; Reznik y Shatalov, 2016; Minguillón y otros, 2017). Sin embargo, no existe una buena y amplia panorámica de la participación de Wikipedia en el conocimiento de las obras literarias o de las obras impresas. Es un terreno cubierto tradicionalmente por los catálogos de biblioteca, las obras de referencia de historia de la literatura y el libro, las revistas de crítica literaria u orientación lectora, o los repertorios bibliográficos. Además, desde la puesta en marcha en 2012 de Wikidata se dispone de una infraestructura para almacenar de forma estructurada los datos estructurados sobre artículos de Wikipedia. Existe un activo movimiento interesado en establecer los procedimientos para usar Wikidata también como base de datos bibliográfica multipropósito: referencias en la propia Wikipedia, análisis bibliométrico, repertorio universal, etc. En definitiva: se percibe un creciente interés e interrelación entre el universo del libro y los proyectos Wikimedia.

Considerando lo anterior se plantea como hipótesis que Wikipedia y Wikidata pueden utilizarse de forma conjunta como fuentes de datos para construir un canon literario. En consecuencia, este trabajo establece una serie de objetivos y una metodología de trabajo para determinar los datos necesarios que deben extraerse, los procesos para realizar tal extracción y el modo en el que deben utilizarse para definir un indicador que permita identificar y ponderar aquellas obras que deben formar parte de dicho canon.

2. DATOS ENCICLOPÉDICOS COLABORATIVOS SOBRE LOS LIBROS Y EL CANON LITERARIO

La omnipresencia de Wikipedia como fuente de información multidominio es un lugar común en los estudios sobre producción colaborativa de contenido (Reagle y Koerner, 2020) y sobre prácticas de uso de información digital. Wikipedia ha alcanzado un altísimo grado de notoriedad y presencia en nuestra vida cotidiana. La enciclopedia online es relevante no solo por su volumen de contenido generalista y local, sino por el lugar que ocupa en las prácticas cotidianas de uso de la red para obtener información, incluido su uso inadvertido como componente de las respuestas que nos proporcionan buscadores y asistentes (Haider y Sundin, 2019).

Una importante cantidad del contenido de Wikipedia está dedicado a los objetos culturales y su contexto: monumentos, cuadros, teatro, autores, discos, libros, películas, esculturas, etc. Sobre este contenido se ha identificado un marcado componente local, puesto que cada comunidad cultural tiene un acervo diferente, vinculado al idioma o el territorio (Miquel-Ribé y Laniado, 2018). Los autores lo denominan *Cultural Context Content* y lo calculan en un 25% en las principales enciclopedias. El proyecto *Wikipedia Diversity Observatory* indica, en su apartado *Topical coverage*, que entre el 1-2% de los artículos de las principales Wikipedias corresponden al tema genérico de "libros"[1]. En este marco, Wikipedia es una fuente relevante de información y recomendación sobre obras literarias, teniendo en cuenta además que no se ofrece un discurso único, puesto que cada comunidad idiomática elabora los artículos sobre obras literarias incorporando sus propias diferencias culturales (Jemielniak y Wilamowski, 2017). Pese al excepcional tamaño de la Wikipedia en inglés, y que a menudo se la contempla como una "catch-all encyclopedia", existen considerables brechas de contenido entre ediciones, especialmente en los contenidos de carácter local (Miquel-Ribé, 2019). Muchas de las grandes obras literarias y del pensamiento, que forman parte del canon cultural y las tradiciones históricas, han merecido la elaboración

de detallados artículos enciclopédicos. Antes de adentrarse en el tratamiento que reciben las obras literarias en Wikipedia, conviene señalar que Wikipedia no es un mero catálogo de referencias de libros, sino que los que aparecen deben ser entidades "notables" con una relevancia enciclopédica suficiente.

Wikipedia tiene una clara tendencia a prestar mayor atención a los fenómenos de la cultura de masas y a su constante producción de novedades. Esto se refleja, para el caso de los libros, en una importante atención a las obras literarias notables recientes y no solo a la literatura clásica y consagrada, a la que en este trabajo denominamos "canon literario universal". Los artículos sobre libros en Wikipedia presentan una gran variabilidad en extensión y tratamiento. Suelen incluir un resumen breve del argumento, explicar las condiciones de escritura y edición, hablar de los personajes, estilo, técnica literaria y repercusión en la época. También suelen contener una ficha descriptiva (*infobox*) que presenta sus datos bibliográficos esenciales, enlaces a bibliotecas digitales para acceder al texto completo de las obras de dominio público y un sistema de categorización.

En el contexto del canon literario es posible observar que existe una mayor cobertura de los autores frente a las obras. Los estudios sobre personas son un enfoque frecuente en investigaciones sobre Wikipedia desde la óptica del análisis de redes (Hube y otros, 2017). Sin embargo, no siempre hay un artículo específico en Wikipedia sobre cada una de las obras de estos grandes autores, aunque sí sea frecuente encontrar información básica (normalmente una lista enumerativa) sobre sus obras principales. También es posible encontrar artículos sobre los propios universos de ficción: personajes, objetos y lugares de ficción.

Los artículos de la enciclopedia corresponden habitualmente al nivel abstracto de Obra (Work) conforme a la conceptualización del modelo de referencia bibliotecario LRM-FRBR. La correcta modelización de los niveles Obra-Expresión-Manifestación es una tarea que interesa a la comunidad bibliotecaria implicada en los datos abiertos y enlazados (Lemus-Rojas y Pintscher, 2018), para que Wikipedia y Wikidata sean un espacio de información bibliográfica más preciso. Por otra parte, la propia definición de qué es una obra literaria es un concepto abierto. En un sentido muy amplio e histórico se entiende como "belles-lettres", incluyendo el ensayo y las obras de pensamiento, y en un sentido más moderno como la ficción creativa (Damrosch, 2009: 6). Aunque cada artículo en cada enciclopedia es un contenido individual, editado y revisado por su propia comunidad de editores, a través de la base de conocimiento Wikidata se encuentran interconectados, de forma que existe una única entidad para representar una obra y vincularla con los artículos en los idiomas en los que exista.

La relación entre Wikipedia y el canon literario no ha sido estudiada específicamente. Se encuadra en las líneas de estudios sobre la literatura en los que se pone el foco en el "sistema literario", o "campo literario" siguiendo la terminología de Bourdieu (1995), y que busca conocer más su impacto y recepción a lo largo del tiempo, y menos su calidad literaria intrínseca. El estudio de las reseñas y críticas publicadas en revistas y suplementos literarios, la presencia de autores y obras en monografías, diccionarios y enciclopedias literarias es una de las metodologías usadas para estudiar el campo literario. Por otra parte, la corriente de estudios "Distant reading" (Moretti, 2013), aborda el estudio de la literatura ampliando el conjunto de fuentes y datos habituales. De esta forma se aprovecha la accesibilidad a la mayor parte de la producción literaria de los últimos siglos, permitiendo el procesamiento de grandes volúmenes de datos de la actividad literaria, incluyendo el análisis informatizado de los propios textos completos. En este sentido, Wikipedia y sus artículos, en cada uno de los idiomas en los que se despliega, es una fuente de datos amplia, dinámica. La exploración de nuevas fuentes interesantes es de interés como punto de partida para definir y comprender las dimensiones de un canon, así como los criterios para estudiarlo (Algee-Hewitt y otros, 2018). En el caso de Wikipedia, contamos además con un espacio de un tamaño muy amplio, pero claramente delimitado y, sobre todo, marcado y codificado con claridad, en formatos fácilmente procesables y con APIs y sistemas de consulta parametrizados, en especial al contar con la información estructurada en Wikidata.

El conglomerado de más de 250 Wikipedias en distintos idiomas está alineado con el campo de estudio de la "*World literature*" (Damrosch, 2009). Esto permite ampliar el foco desde un canon occidental con fuertes sesgos, hacia otro más amplio y global. También permite ir más allá del "*translated canon*", en donde existe un sesgo muy fuerte hacia las lenguas con grandes mercados editoriales, como el inglés, francés o el español, etc. (Venuti, 2008). Como hemos mencionado anteriormente, los estudios sobre Wikipedia son conscientes del "*culture gap*" entre ediciones para contenidos locales y culturales (Miquel-Ribé y Laniado, 2021). Por lo tanto, para explorar el canon global y en cada idioma, de acuerdo con Wikipedia, será necesario partir de las ediciones en cada idioma para obte-

ner datos que reflejen su verdadera naturaleza de fuente diversa.

Cada edición Wikipedia para cada idioma funciona de forma independiente, representando las elecciones de sus editores y su contexto. Sin embargo, Wikidata es una base de datos común, producida al mismo tiempo por editores en todos los idiomas. Es un único proyecto cuyo objetivo es la creación es un grafo de conocimiento producido colaborativamente por editores de cualquier idioma. Wikidata tiene un alcance universal, y modela los diferentes ámbitos del conocimiento mediante la creación colaborativa y supervisada de propiedades. Integra tanto los datos sobre las instancias (Charles Chaplin; Estadio Azteca; Monte Everest), como las propiedades para establecer relaciones y recoger datos (Fecha de nacimiento; Aforo; Coordenadas), como las clases, subclases y el vocabulario controlado para describirlas (Actor; Estadio de fútbol; Montaña). Con respecto al libro, existe un wikiproyecto en el que se acuerdan metadatos y pautas de descripción y otros aspectos de interés para su descripción[2].

Se han realizado numerosas propuestas para la evaluación automática de aspectos de calidad de los contenidos de Wikipedia basados en métodos cuantitativos, que constituyen por sí mismas un subcampo de estudio sobre Wikipedia (Nielsen, 2019). Unos explotan las métricas del análisis de redes, usando los enlaces entre artículos y el grafo resultante. Otros usan las métricas propias disponibles para el contenido de los artículos: número de palabras, número de referencias, extensión, enlaces entrantes, etc., complementados con el estudio de la actividad de los editores, reputación y redes de colaboración. Del mismo modo sucede en Wikidata, con investigaciones para establecer la calidad y completitud de los datos (Shenoy y otros, 2022). Las métricas automáticas sirven de medición indirecta de la "calidad esperada" o probabilidad de calidad, en realidad, credibilidad (Claes y Tramullas, 2021). Se trata de un campo que genera investigación aplicada, uno de cuyos casos, el sitio web WikiRank[3] ilustra con claridad la posibilidad de establecer rankings de artículos segmentados por tipos de contenido, mediante indicadores agregados que denominan "*popularity*", "*Authors' Interest*" (AI) y "Citation Index" (Lewoniewski y otros, 2019). El trabajo más conocido sobre ranking es el de Skiena y Ward (2014) en el que se comparan personajes históricos diferenciando entre celebrity (popularidad actual) y gravitas (popularidad consolidada).

3. OBJETIVOS Y METODOLOGÍA

Wikidata es un grafo de conocimiento que utiliza su propio modelo de datos compatible con RDF. Sus elementos principales son ítems con un identificador único cuya designación comienza por la letra "Q". Por ejemplo, el libro "Cien años de soledad" de Gabriel García Márquez es el elemento Q178869, aunque está vinculado a 74 artículos en diferentes Wikipedias (español, japonés, italiano, ruso, etc.). A su vez, cada ítem se describe mediante propiedades cuyas designaciones comienzan por la letra "P". Las propiedades definen relaciones entre elementos o se refieren a valores literales (cadenas, números, fechas). Por ejemplo, del libro mencionando se declara que tiene como autor (P51) al elemento Q5878 (el escritor García Márquez) y que su fecha de publicación (P577) es 1967. Wikidata no tiene clases definidas explícitamente diferenciadas del resto de los elementos. En cambio, algunos elementos desempeñan tal papel de clase al enmarcarse en una taxonomía de clases y subclases conectadas a través de la propiedad P279 (subclase de). La pertenencia de los ítems a las clases se realiza mediante la propiedad P31 (instancia de). Esta circunstancia permite, hasta cierto punto, entender a Wikidata como una "ontología colaborativa", que no solo contiene datos primarios, sino una suerte de esquema formalizado de organización del conocimiento (Piscopo y Simperl, 2018). Dentro de cada ítem existe una sección denominada "Identificadores", que definen conexiones con registros y bases de datos externas de todo tipo, como, por ejemplo, con el sistema internacional de control de autoridades VIAF (Bianchini y Sardo, 2022).

A partir de las consideraciones hasta ahora expuestas, se propone reutilizar aquellos datos disponibles, tanto en los contenidos enciclopédicos de Wikipedia como en la base de conocimiento estructurado de Wikidata, para construir un procedimiento que permita definir un canon literario. Por lo tanto, para demostrar la hipótesis planteada en la introducción del trabajo, se establecen los siguientes objetivos generales:

- Identificar el conjunto de datos enciclopédicos relativos a obras literarias de todas las épocas en cualquier idioma.
- Validar un procedimiento analítico automático para establecer agrupaciones y ranking de obras literarias con cobertura en cualquiera de las diferentes ediciones de Wikipedia.
- Identificar medidas representativas del impacto de cada obra literaria en el ecosistema Wikimedia.
- Analizar la distribución temporal de las obras del canon literario desde el punto de vista de su publicación o, en su defecto, creación.

El método para obtener el conjunto de datos se ha desarrollado en cuatro fases:

- Primera etapa: Determinación del ítem que desempeñaría el papel de clase a partir de la cual recuperar los ítems de las obras literarias.
- Segunda etapa: Construcción de un conjunto de datos.
- Tercera etapa: Agregación de ciertos datos del conjunto de datos.
- Cuarta etapa: Análisis de los resultados de agregación.

Tanto el conjunto de datos obtenido, los datos agregados, como los scripts en Python y Orange Data Mining están disponibles para su consulta y reutilización pública[4].

En la primera etapa se tomó el ítem "Obra literaria" (Q7725634) como clase de partida para la exploración de Wikidata. De este modo se recuperan aquellos ítems relacionados con dicha clase mediante la propiedad P31 (Instancia de). La taxonomía de clases usada para el universo bibliográfico es amplia y con significativas imprecisiones en sus jerarquías y aplicación. Se recuperaron tan solo los elementos con asignación directa a esta clase. Se tomó la decisión de no considerar las taxonomías derivadas de los elementos Q471 (libro) ni Q47461344 (Obra escrita), aunque son usadas para instanciar un número considerable de ítems del campo literario, para minimizar el riesgo de recuperar resultados alejados del foco del trabajo, que habrían requerido procedimientos muy minuciosos de validación.

En la segunda etapa se construyó el conjunto de datos. Solo se recuperaron los elementos sobre obras literarias que tienen un artículo escrito sobre ella en alguna Wikipedia de cualquier idioma. Este criterio de relevancia o notabilidad permitió extraer información solo de obras en las que se identifica un esfuerzo editorial y no únicamente la existencia de meros datos en Wikidata. Dada la estrecha interrelación entre ambos proyectos, enciclopedias y base de conocimiento, la mayor parte de los ítems de Wikidata también pertenecen a alguna Wikipedia.

Se utilizaron consultas SPARQL en Wikidata Query Service (WDQS) que permitieron obtener:

- Identificadores de todos los ítems definidos como instancias del ítem "Obra literaria" (Q7725634) con una o varias correspondencias en ediciones de Wikipedia, así como una lista de todas las propiedades y declaraciones utilizadas para la descripción de cada uno de dichos ítems. Este trabajo, utiliza la denominación "obra literaria" para referirse a cada uno de los ítems recuperados.
- Los idiomas en los que se escribieron las obras literarias recuperadas.
- URL de los artículos en Wikipedias de diferentes idiomas de los ítems recuperados. La denominación "sitelink" se refiere a cada una de dichas referencias.
- La fecha de publicación o concepción de las obras.
- El título identificativo de cada obra en español e inglés y, en su defecto, en el idioma original.
- Un listado completo de todas las propiedades de Wikidata, distinguiendo aquellas utilizadas en la sección de identificadores (propiedades ID).

Además de WDQS se ha utilizado el servicio Xtools de Wikimedia[5] consultado desde scripts Python para automatizar las consultas. Mediante la correspondiente API de este servicio se recuperó información estadística sobre la estructura de cada uno de los artículos correspondientes a los ítems recuperados. Así pues, de cada artículo de Wikipedia, se han obtenido los datos correspondientes al número de palabras, referencias, número de ediciones, fechas de creación y modificación, enlaces externos, etc.

Fue necesario realizar un proceso de consolidación de datos. Por ejemplo, no todos los ítems recuperados incluían declaraciones explícitas relativas al idioma de la obra (P407) o la fecha de publicación (P577). Sin embargo, en algunos casos, esta información se ha podido obtener extrayendo el idioma en el que se encuentra el título original de la obra (P1476) y la fecha de concepción de ésta (P571). En el conjunto de datos final se indican las propiedades utilizadas para obtener estos datos.

En la tercera etapa se procedió al procesamiento del conjunto de datos para obtener resultados con datos más agregados. Se desarrolló un script en Python para la agregación y la obtención de medidas estadísticas.

Para los ítems de cada obra se agregaron los siguientes datos a partir del conjunto de datos previamente generado:

- Identificador "Q" en Wikidata, en el espacio de nombres o prefijo "wd:".
- Idioma original de la obra.
- Etiqueta o título identificativo de la obra.
- Fecha de publicación o concepción de la obra.
- Número total de Wikipedias en las que el ítem tiene presencia con su correspondiente artículo (N_{Wikis}).
- Número total de declaraciones: en este caso se ha distinguido entre propiedades ID y el

resto de las propiedades utilizadas en las declaraciones (N_{Props}).
- Número total de palabras utilizadas en todos los artículos correspondientes al ítem en las diferentes Wikipedias (N_{Words}), calculado a partir de los datos recuperados de Xtools.

De igual forma, para cada idioma se agregaron o calcularon los siguientes datos:

- Código estándar de identificación del idioma. Se han agrupado las diferentes variaciones regionales de un mismo idioma.
- Número de ítems recuperados de las obras escritas en ese idioma.
- Media aritmética del número de Wikipedias en las que tienen presencia los ítems de las obras del idioma en cuestión.
- Media aritmética del número de declaraciones con propiedades no ID.
- Media aritmética del número de palabras de los artículos en Wikipedia correspondientes al ítem de la obra.

Para finalizar esta etapa se procedió a generar una matriz de idiomas/Wikipedias que representa el número de artículos sobre obras literarias de un determinado idioma que tiene presencia en cada una de las diferentes ediciones de Wikipedia. No obstante, estos datos no se han explotado en este trabajo.

En la cuarta etapa se analizaron los datos obtenidos mediante la herramienta Orange Data Mining[6]. Dicho análisis comenzó con la representación de la distribución normalizada de los ítems de cada obra en función de los valores de N_{Wikis}, N_{Props} y N_{Words}. Se realizó un clustering de los ítems mediante el método K-means (Hartigan y Wong, 1979; Arthur y Vassilvitskii, 2007). El número de clústeres se determinó mediante la puntuación obtenida a través del método Silhouette (Rousseeuw, 1987).

Tras analizar los resultados obtenidos y estudiar la distribución de N_{Wikis}, N_{Props} y N_{Words} se procedió a calcular un indicador que combinara las tres variables. Este indicador, denominado Wiki3DRank, permitiría ordenar los ítems de las obras literarias con una distribución normalizada que considerara los tres factores establecidos en los objetivos de la investigación: *presencia en Wikipedias, profundidad de descripción en Wikidata y extensión de los artículos en Wikipedia*. Una vez hecho esto se comprobó que los resultados de Wiki3DRank eran coherentes con los obtenidos en el proceso de clustering.

4. RESULTADOS

En primer lugar, se presentan los datos relativos a la pregunta planteada de cuáles y cuántos podrían ser, según la actividad de las comunidades Wikimedia, las obras literarias que compondrían un canon universal, delimitando un subconjunto de entre las obras literarias recuperadas para nuestro dataset. En segundo lugar, se analizan aspectos sobre las literaturas en cada idioma. En tercer lugar, se realiza una presentación de la distribución temporal de las obras del canon.

4.1 Canon literario universal a partir de los datos de Wikipedia: difusión y esfuerzo editorial

Este trabajo establece la presencia de un artículo sobre la obra literaria en alguna Wikipedia como condición indispensable para considerar un ítem relevante. Por lo tanto, el dataset resultante incluye un total de 107.434 ítems de Wikidata[7], definidos como instancias (P31) de la ítem-clase "Obra literaria" (Q7725634). Sin considerar dicha condición de vinculación el total de ítems asciende a 192.236. Esto implica que se descartaron más de un 44% de ítems que pueden considerarse como meros "registros de catálogo" y no entidades con la suficiente relevancia o notabilidad para requerir un artículo enciclopédico explicativo. Este hecho señala cierta tendencia a usar Wikidata como base de datos bibliográfica de propósito general, como WikiCite.

Se ha considerado la distribución de los ítems de las obras literarias en función del número de Wikipedias en las que aparece (N_{Wikis}) el número de declaraciones en Wikidata (N_{Props}) y el total del número de palabras de sus artículos en la enciclopedia (N_{Words}). En la Tabla I se detallan algunos indicadores estadísticos para cada variable. La mayor dispersión de valores (C_v) se da para N_{Words} y $N_{Wikis.}$ Las tres variables tienen una distribución con una fuerte asimetría positiva (Coeficiente de Asimetría de Fisher) y un alto grado de curtosis. Gran parte de los ítems del conjunto de datos tienen valores bajos en cada una de las variables, que señala a una gran bolsa de obras literarias con una baja presencia en Wikipedias, menor profundidad en la descripción y artículos más breves.

El análisis de la correlación entre las tres variables (Tabla II) refleja una correlación entre N_{Wikis} y $N_{Words.}$ Esto es obvio: a mayor número de ediciones de Wikipedia en las que un ítem de Wikidata tiene un artículo equivalente, mayor es el número total de palabras del conjunto de dichos artículos. Este análisis también refleja que la menor correlación se produce entre N_{Props} y N_{Words}, es decir entre la descripción y el artículo, entre datos y texto.

Pese lo anterior, existen otros ítems cuyos valores para alguna de las variables (e incluso las tres)

Tabla I: Análisis estadístico de N_{Wikis}, N_{Props}, N_{Words}.

Variable	Media	Mediana	C_v	Mínimo	Máximo	Asimetría	Curtosis
N_{Wikis}	1,964	1	2,026	1	140	11,248	191,148
N_{Props}	5,946	5	0,756	1	276	8,942	294,509
N_{Words}	849,19	198	4,006	0	168.391	18,453	537,757

Tabla II: Índices de correlación de Pearson y Spearman entre N_{Wikis}, N_{Props} y N_{Words}.

Pearson	N_{Wikis}	N_{Props}	N_{Words}	Spearman	N_{Wikis}	N_{Props}	N_{Words}
N_{Wikis}	-	0,529	0,839	N_{Wikis}	-	0,334	0,412
N_{Props}	0,529	-	0,494	N_{Props}	0,334	-	0,236
N_{Words}	0,839	0,494	-	N_{Words}	0,412	0,236	-

están por encima del resto. Estos datos permitirían verificar la hipótesis de este trabajo, puesto que los ítems con valores más altos que el resto permitirían identificar las obras que destacan y que podrían formar parte del canon literario. Es decir, los ítems de las obras que forman parte del canon literario tendrían una mayor presencia en diferentes ediciones de Wikipedia, un mayor nivel de descripción en Wikidata y un mayor grado de elaboración de los artículos respecto al resto de obras.

¿Qué número de obras compondría ese grupo selecto de obras universales muy destacadas? Se utilizó el algoritmo K-means++ para agrupar los ítems en clústeres que permitieran identificar las obras de un posible canon literario. Los resultados del método Silhoutte indicaban la posibilidad de usar K-means++ para obtener dos o tres clústeres. La aplicación de K-Means++ con dos clústeres identificó 1.008 ítems. Esta cifra podría resultar excesiva para la idea de un canon literario como lista de obras abarcable de forma fácil para una persona o "para llevarse a una isla desierta", aunque quizá no tanto para hacer un inventario selecto de la cultura escrita universal desde hace más de tres milenios. Por este motivo se amplió la aplicación de K-means, realizando los correspondientes cálculos hasta con siete clústeres.

En función del tamaño del clúster superior para cada iteración de K-means++, se evaluó el nivel de coincidencia con N_{Wikis}, N_{Props} y N_{Words}. La variable con mayor ratio de coincidencia es N_{Words}. Sin embargo, el conjunto de los ítems de las obras que debían formar parte del canon era diferente en función de la variable utilizada. Por este motivo se procedió a reducir la dimensionalidad utilizando dos métodos. El primero de ellos fue el método PCA (Ding y He, 2004) calculado a partir N_{Props} y N_{Words} puesto que son las variables con menor correlación. También se ha definido y calculado un indicador, al que se ha denominado Wiki3DRank, como la agregación

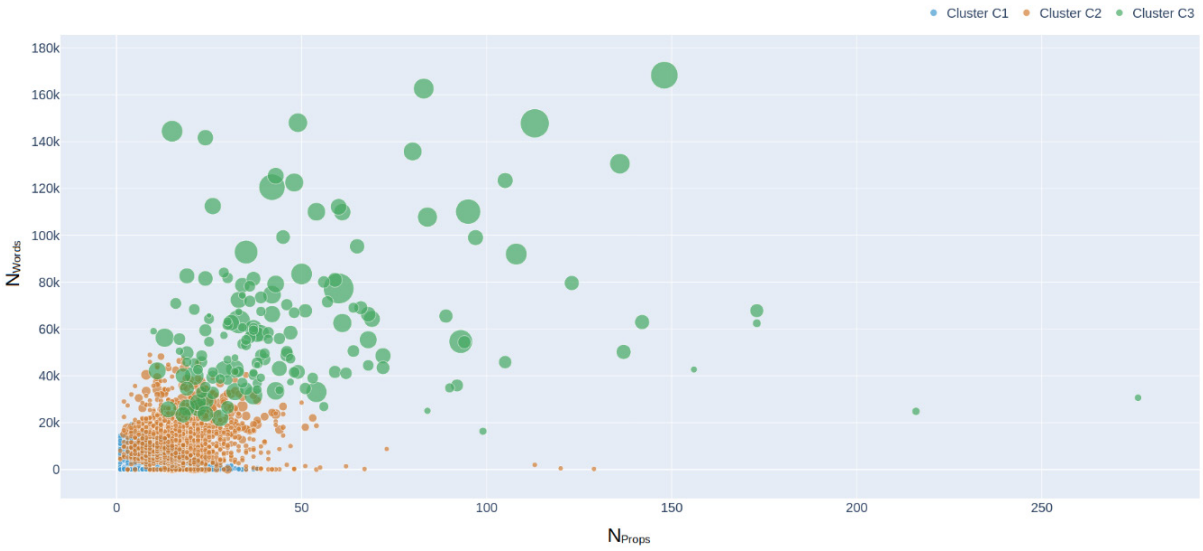
Tabla III: Análisis estadístico de Wiki3DRank.

Variable	Media	Mediana	C_v	Mínimo	Máximo	Asimetría	Curtosis
Wiki3DRank	7,556	7,745	0,365	1,386	21,874	-0,014	0,610

Tabla IV: Ratios de coincidencia para las diferentes iteraciones de K-means++ (entre paréntesis el número de ítems coincidentes). Fuente: elaboración propia.

S_n	Silhoutte	Ratio de coincidencia				
		N_{Wikis}	N_{Props}	N_{Words}	PCA	Wiki3DRank
1.008	0,909	0,869 (876)	0,499 (503)	0,802 (808)	0,882 (889)	0,927 (934)
163	0,827	0,822 (134)	0,595 (97)	0,822 (134)	0,822 (134)	0,939 (153)
152	0,493	0,822 (125)	0,559 (85)	0,849 (129)	0,822 (125)	0,934 (142)
74	0,499	0,676 (50)	0,608 (45)	0,824 (61)	0,676 (50)	0,919 (68)
65	0,493	0,615 (40)	0,600 (39)	0,846 (55)	0,615 (40)	0,908 (59)
36	0,457	0,472 (17)	0,556 (20)	0,750 (27)	0,472 (17)	0,833 (30)

Figura 1. Representación de los tres clústeres principales de obras literarias. Distribución de N_{Props}/N_{Words} (el tamaño de los elementos representa N_{Wiki}).



de la transformación logarítmica de cada una de estas variables (Shatnawi, 2015). Para cada ítem Wiki3DRank se calcularía como:

$$Wiki3DRank = \log(1+N_{Wikis}) + \log(1+N_{Props}) + \log(1+N_{Words})$$

Esta ecuación, cuyo cálculo es muy sencillo, integra N_{Wikis}, N_{Props} y N_{Words} en un único indicador con una distribución relativamente normalizada (Tabla III).

En cada iteración *n* de K-means++, se estimó la coincidencia entre el conjunto de ítems del clúster C_n (clúster superior) y el subconjunto delimitado entre el intervalo [1, *Sn*] de cada uno de los rankings establecidos por N_{Wikis}, N_{Props}, N_{Words}, *PCA y Wiki3DRank.* En función del número de elementos coincidentes y el tamaño del clúster superior (S_n) se calculó una ratio de coincidencia (ver Tabla IV). *Wiki3DRank* alcanza las mayores ratios de coincidencia en cualquier iteración, siendo el más alto el correspondiente a la iteración con tres clústeres. También puede observarse que de las tres componentes que definen un ítem, N_{Words} es más representativa que N_{Wikis} o N_{Props} respecto a la coincidencia con los resultados de K-means++.

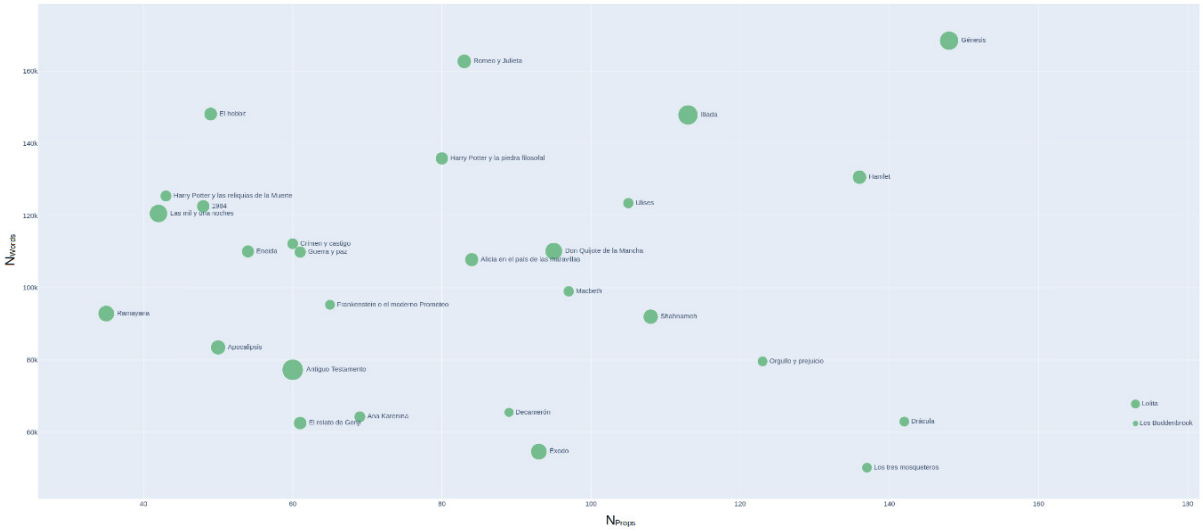
Considerando estos datos, se ha optado por usar K-means++ para obtener tres clústeres. El tamaño de C_1 es de 105.100 ítems, C_2 (al que denominamos clúster secundario) contiene 2.171 ítems y C_3 (clúster principal) incluye 163 obras. En consecuencia, podría interpretarse que el clúster principal contiene los ítems de aquellas obras candidatas a ser consideradas Canon Literario Universal. C_1 podría denominarse "producción bibliográfica", un vasto conjunto de libros y obras con mayor o menor fortuna, de impacto más local y atención escasa. El clúster secundario C_2 lo forma un conjunto, relativamente abarcable, de obras que representan en cierto modo la clase media de la literatura: obras con notoriedad en un conjunto de idiomas y con niveles de atención enciclopédica variables. En la Figura 1 se visualizan claramente los tres clústeres. Cada obra literaria se representa en un diagrama de dispersión con respecto a los ejes N_{Words} y N_{Props}, mientras que el tamaño de cada elemento se determina mediante N_{Wikis}.

A modo de ejemplo, se muestran los datos de una obra de cada clúster en la Tabla V:

Tabla V: Ejemplo de obra de cada clúster.

Ítem	Título	Clúster	Wiki3DRank	N_{Wikis}	N_{Props}	N_{Words}
Q8275	Ilíada	C3	21,5304	132	113	147.831
Q220331	Ben-Hur	C2	17,0640	28	27	31.712
Q27223	Babel-17	C1	13,3556	11	10	4.782

Figura 2: Selección de las 30 obras del clúster C3 con mayor Wiki3DRank.



Además de la agrupación en clústeres, la métrica *Wiki3DRank* ofrece un valor para operaciones de selección y ordenación. De este modo es posible ordenar los ítems del conjunto de datos, independientemente del clúster al que pertenezcan y dentro de él. La Figura 2 representa una selección de las cincuenta primeras obras de C_3. En el anexo final es posible consultar un listado completo de las obras de este clúster de "clásicos universales de todos los tiempos".

En las primeras posiciones de C_3 encontramos obras fácilmente reconocibles como: *El Génesis, La Ilíada, Hamlet, Romeo y Julieta, Don Quijote de la Mancha, Shahnameh, Ulises, Harry Potter y la piedra filosofal, Alicia en el país de las maravillas, Lolita, Macbeth, Orgullo y prejuicio*, etc. Aplicando estas mismas métricas se puede extraer el canon literario escrito en cualquier idioma, como, por ejemplo, la tabla VI recoge las diez primeras obra de todos los tiempos escritas en italiano.

4.2 La literatura clásica y actual: cánones literarios locales y el peso de la tradición

Los datos disponibles permiten abordar diferentes facetas de los estudios de los fenómenos culturales, como la perduración de ciertas obras a lo largo del tiempo, la representación equilibrada de los diferentes idiomas o la permeabilidad del canon global y los cánones en cada idioma entre diferentes ámbitos lingüísticos. En este trabajo presentamos tan solo los datos básicos de la distribución por idiomas del canon considerado global y analizamos con algo más de detalle su distribución temporal según la fecha de producción o publicación.

Tabla VI. Selección de las obras con mayor ranking escritas en italiano.

Item	Title	Cluster	*Wiki3DRank*	N_{Wikis}	N_{Props}	N_{Words}
Q16438	The Decameron	C3	8,583	64	89	65521
Q8065468	The Adventures of Pinocchio	C3	8,151	67	49	41696
Q172850	The Name of the Rose	C3	8,123	53	41	58536
Q131719	The Prince	C3	8,081	72	19	82696
Q48922	Orlando Furioso	C3	7,971	35	34	74398
Q1053313	Jerusalem Delivered	C2	7,681	32	35	40388
Q808428	Gospel of Barnabas	C3	7,356	34	10	59060
Q1645493	Lives of the Most Excellent Painters, Sculptors, and Architects	C2	7,165	31	23	19048
Q914235	Hypnerotomachia Poliphili	C2	7,069	24	16	27636
Q641651	Six Characters in Search of an Author	C2	8,583	64	89	65521

La faceta idiomática se abordó desde dos vertientes: el idioma de las obras y las ediciones de Wikipedia en las que estaban presentes. El idioma de cada obra se determinó con dos mecanismos: de forma expresa mediante la propiedad P407 o (en algunos casos) extrayendo el idioma del título original. Para la literatura en cada idioma se contabilizó el número de obras y se calcularon los valores medios de N_{Props} y N_{Words}. En la Figura 3 puede observarse la dispersión de cada idioma en función de las medias de N_{Props} y N_{Words}. El tamaño de los elementos se define en función del total de obras de cada idioma según dichas medias. Esta primera aproximación permite observar el cuidado puesto en el contenido enciclopédico para cada literatura; las obras en inglés son mayoritarias y además con un alto grado de descripción. También puede observarse un elemento con la etiqueta "<none>" referido a aquellas obras en cuyos ítems de Wikidata no existen datos de idioma. Estas obras son numerosas (39.465) pero como puede verse sus ítems de Wikidata tienen un bajo nivel tanto de descripción como de edición de sus correspondientes artículos en las diferentes ediciones Wikipedia. Del resto de idiomas destacan el español, francés, japonés, ruso y alemán. Cabe destacar el caso del latín, sánscrito y griego clásico con un bajo volumen de obras, pero con numerosas declaraciones descriptivas y con artículos extensos en las distintas ediciones de Wikipedia.

En relación con la fecha de las obras literarias seleccionadas, los datos obtenidos permiten trazar un panorama sobre la época a la que pertenecen las obras que forman parte del canon literario universal o local. A partir de los datos obtenidos podemos analizar la distribución temporal de las obras. Es necesario señalar que para un gran número de ítems de obras no se dispone de información sobre la fecha de publicación o de creación. Únicamente 61.702 ítems (algo más de un 57%) incluyen alguna propiedad para obtener esta información. La mayoría de los datos se obtuvieron a partir de la propiedad P577 (fecha de publicación) y únicamente un 2,2% mediante la propiedad P571 (fecha de creación). Los resultados, agrupados por siglos, pueden verse en la Tabla VII.

Más del 87% de los ítems que dispone de algún tipo de fecha (un 50% del total de ítems del conjunto de datos) tienen una fecha de publicación o creación correspondiente a los siglos XX y XXI. Los datos pueden agruparse o analizarse de forma más detallada. La Figura 4 muestra una distribución del Wiki3DRank por siglo y también de todos los ítems de las obras publicadas o creadas en el Siglo XX por año. También se muestra el clúster al que pertenece cada obra. Como parece razonable, se recogen pocas obras de la antigüedad remota que no tengan cierta relevancia (clúster principal y secundario). En general, podemos ver que los datos tienen una distribución temporal variada, con un acento, para el siglo XX, en sus años centrales.

El filtrado por idioma es otra interesante posibilidad que ofrece el estudio de los datos temporales.

Figura 3: Distribución de idiomas con un mínimo de 100 obras, basada en la media de N_{Props} (eje x) y N_{Words} (eje y). El tamaño de los puntos representa el número de obras literarias en el idioma.

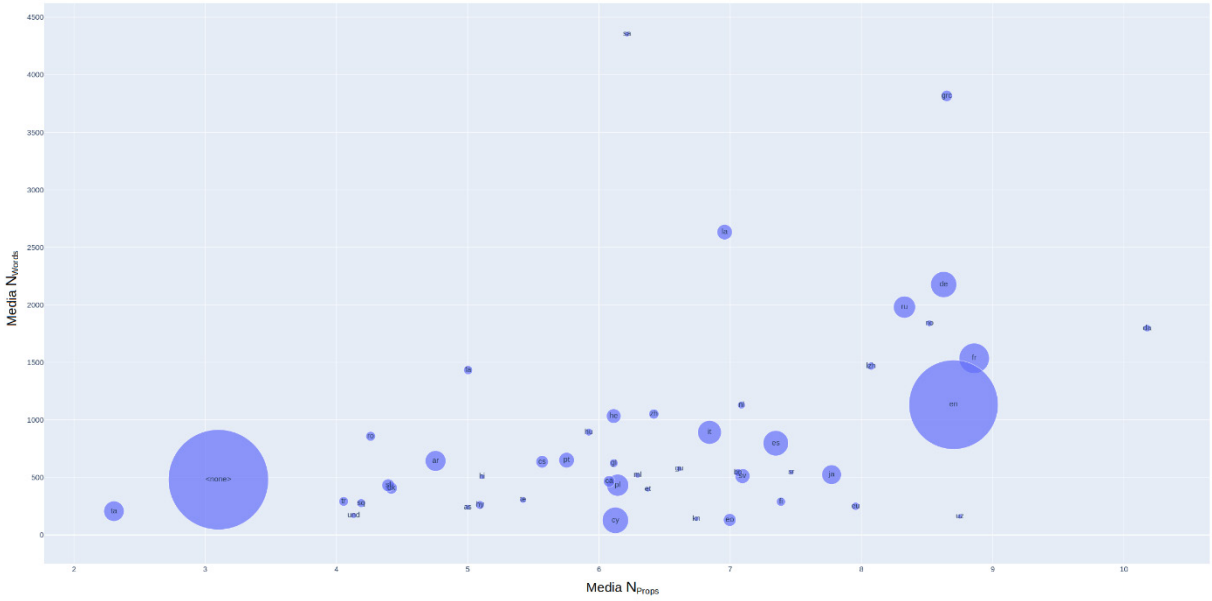
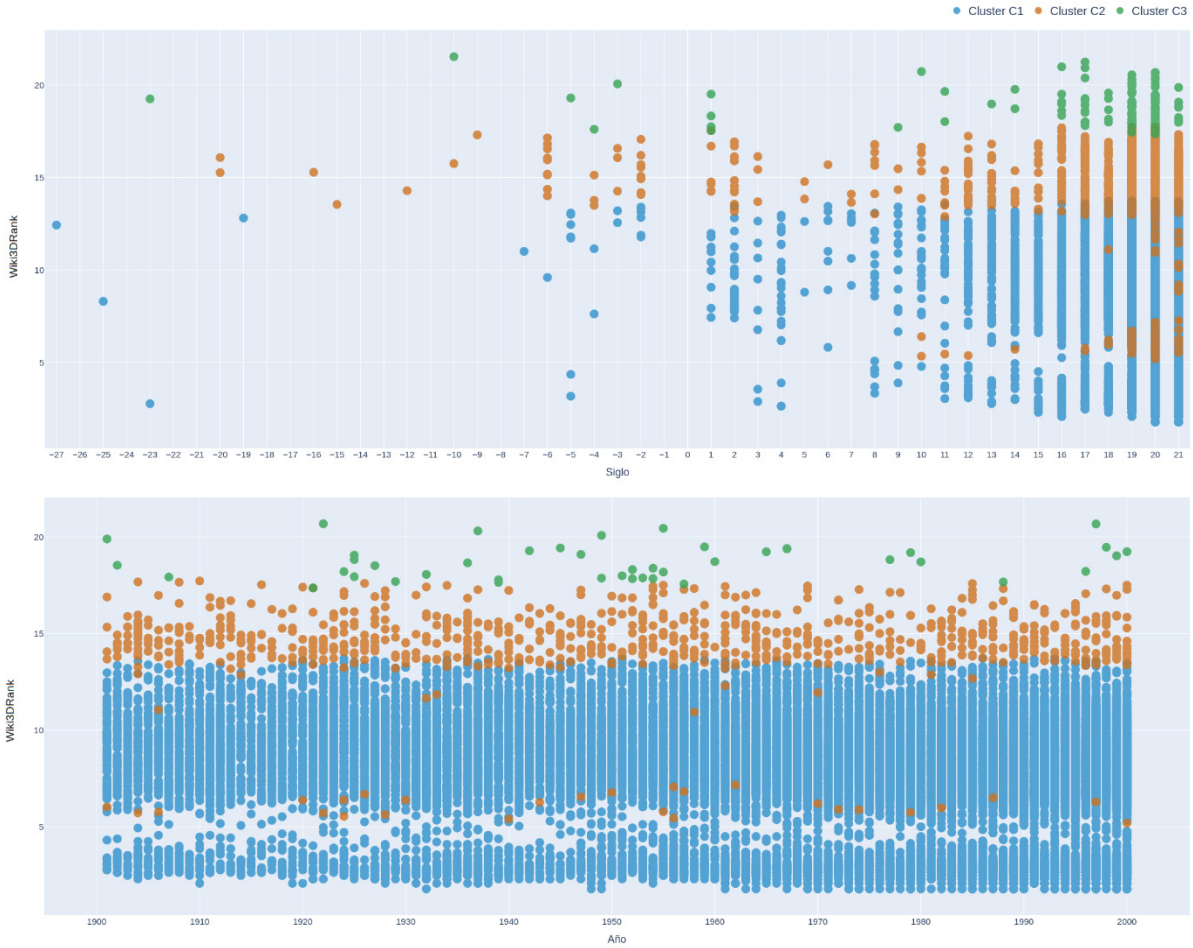


Tabla VII: Distribución temporal por siglos del número de ítems y Wiki3DRank. Fuente: elaboración propia.

Siglo	Ítems	C1	C2	C3	Total Wiki3DRank	% ítems	% con fecha	% Wiki3DRank	Ratio Wiki3DRank
21	2.319	22.095	217	7	171.355,06	24,87	40,32	35,01	7,68
20	1.836	30.990	802	44	256.594,02	35,47	57,51	52,42	8,06
19	5.174	4.675	54	45	48.627,96	5,77	9,35	9,93	9,4
18	745	665	75	5	7.076,54	0,83	1,35	1,45	9,5
17	510	423	75	12	5.118,64	0,57	0,92	1,05	10,04
16	405	364	35	6	3.764,97	0,45	0,73	0,77	9,3
15	127	115	12	0	1.231,98	0,14	0,23	0,25	9,7
14	109	100	7	2	1.065,45	0,12	0,2	0,22	9,77
13	109	90	18	1	1.113,1	0,12	0,2	0,23	10,21
12	78	57	21	0	817,74	0,09	0,14	0,17	10,48
11	39	29	8	2	421,13	0,04	0,07	0,09	10,8
10	31	22	8	1	351,85	0,04	0,06	0,07	11,35
9	20	17	2	1	217,65	0,02	0,04	0,04	10,88
8	24	17	7	0	250,53	0,03	0,04	0,05	10,44
7	8	6	2	0	98,69	0,01	0,01	0,02	12,34
6	9	8	1	0	103,88	0,01	0,02	0,02	11,54
5	4	2	2	0	50,04	0	0,01	0,01	12,51
4	25	25	0	0	219,4	0,03	0,05	0,05	8,78
3	11	8	3	0	110,54	0,01	0,02	0,02	10,05
2	39	28	11	0	421,77	0,04	0,07	0,09	10,81
1	19	9	6	4	256,07	0,02	0,03	0,05	13,48
-2	16	6	10	0	229,05	0,02	0,03	0,05	14,32
-3	6	2	3	1	92,72	0,01	0,01	0,02	15,45
-4	6	2	3	1	78,76	0,01	0,01	0,02	13,13
-5	9	8	0	1	100,65	0,01	0,02	0,02	11,18
-6	10	1	9	0	150,79	0,01	0,02	0,03	15,08
-7	1	1	0	0	11,01	0	0	0	11,01
-9	1	0	1	0	17,31	0	0	0	17,31
-10	2	0	1	1	37,28	0	0	0,01	18,64
-12	1	0	1	0	14,29	0	0	0	14,29
-15	1	0	1	0	13,54	0	0	0	13,54
-16	1	0	1	0	15,28	0	0	0	15,28
-19	1	1	0	0	12,81	0	0	0	12,81
-20	2	0	2	0	31,34	0	0	0,01	15,67
-23	2	1	0	1	22,02	0	0	0	11,01
-25	1	1	0	0	8,3	0	0	0	8,3
-27	1	1	0	0	12,43	0	0	0	12,43

Figura 4: Distribución por siglos (arriba) o años del Siglo XX (abajo) y Wiki3DRank de los ítems según su fecha de creación o publicación.



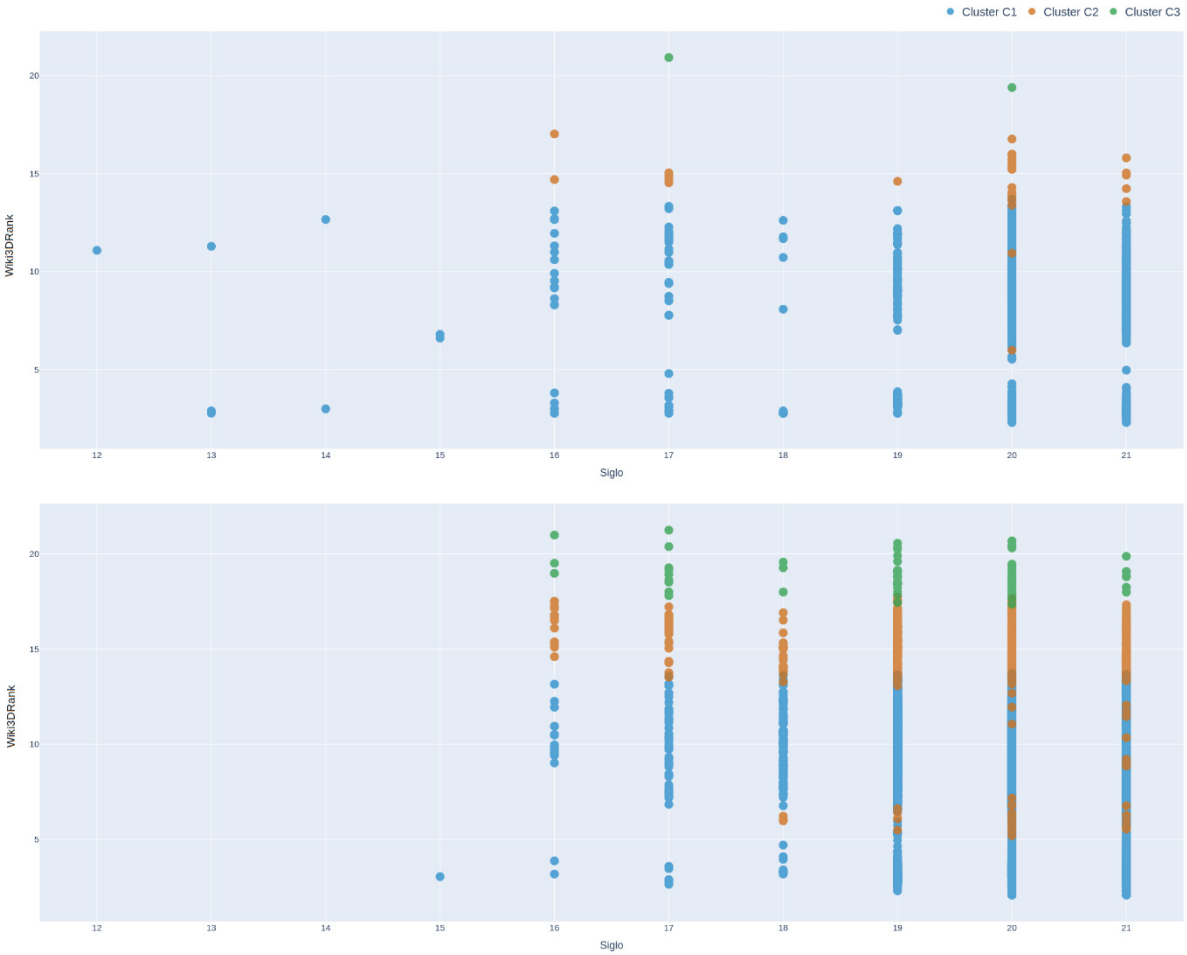
Puede observarse la evolución en la producción literaria y su éxito en el tiempo para cada idioma, e incluso la comparación de esta faceta entre diferentes idiomas (Figura 5).

5. DISCUSIÓN

El trabajo presenta varios elementos que podrían ser valiosos: establece un umbral cuantitativo de la cantidad de obras que podríamos señalar como excepcionalmente relevantes globalmente, una forma de ponderarlas individualmente y una lista de obras del canon literario universal. En esta lista pueden encontrarse títulos generalmente señalados como "clásicos de todos los tiempos". Son relatos fácilmente identificables como parte de la tradición y que pueden situarse en ciertos momentos y lugares de la historia.

La cuantificación de N_{Wikis}, N_{Props} y N_{Words} refleja tres medidas indirectas de algo que podemos llamar "esfuerzo enciclopédico" o también "atención enciclopédica". Por un lado, la extensión de los artículos en Wikipedia en cualquier idioma, usando una medida acumulativa y no una medida ponderada de centralidad (N_{Words}). Por otro lado, la profundidad descriptiva en Wikidata, que refleja otro tipo de atención orientada a datos y detalles factuales sobre la obra ($_{NProps}$). Complementa la visualización la difusión de la obra a lo largo de diferentes idiomas, que actúa como indicador de presencia global (N_{Wikis}). La integración de estas tres variables permite una representación más rica que el uso de cada una de ellas por separado. La posición de las obras con respecto a los ejes de los diagramas de dispersión indica hacia donde se inclina la balanza del equilibrio entre texto-datos (Wikipedia/Wikidata) y permite detectar irregularidades y asimetrías. Pocas obras del clúster secundario C_2 alcanzan magnitudes comparables, en alguno de los tres parámetros, con las obras del clúster principal C_3.

Figura 5: Distribución temporal de las obras en español (arriba) e inglés (abajo).



Si comparamos las obras del clúster principal C_3 con los resultados del sitio web WikiRank para la categoría "Books" encontramos una coincidencia del 74,3%. Sin embargo, apreciamos una mejor ordenación de las obras aplicando Wiki3DRank y una sorprendente cantidad de obras recientes y comerciales en los resultados mostrados en WikiRank. Los tres clústeres obtenidos guardan cierto parecido con otras propuestas realizadas desde otros presupuestos. El clúster principal C_3, de 163 obras, tiene unas dimensiones similares al que propone Christiane Zschirnt en su estudio "Libros, todo lo que hay que saber" (Zschirnt, 2011). Esta autora escoge 141 obras, existiendo una coincidencia sustancial entre su selección y las obtenidas en este trabajo[8]. Por otro lado, un enfoque más abarcador como "1001 libros que hay que leer antes de morir» (Boxall y Mainer, 2016), número escogido por su vistosidad, es más cercano al conjunto de C_3 y C_2 (2336 obras) aunque se ajusta casi a la perfección con la alternativa del cálculo de únicamente dos clústeres, que resultaría en 1008 obras.

Un aspecto que llama la atención es la presencia de obras de las tradiciones religiosas, especialmente la judeocristiana. Generalmente estas obras no tienen consideración de obras literarias en los estudios del ámbito de la crítica e historia de la literatura ("Génesis", "Levítico", "Epístola a los Filipenses", etc.). Estos textos mitológico-espirituales constituyen la base de comunidades religiosas y merecerían diferenciarse para obtener un cuadro más ajustado a lo que hoy se considera literatura en sentido estricto. Este mismo problema surge cuando encontramos obras de ensayo, pensamiento o divulgación ("La Riqueza de las naciones" o "La República"). Desde el ámbito de la distribución de libros, se tiende a diferenciar entre los bloques de ficción y no-ficción, situándose la literatura en el primer grupo. Llama la atención la inclusión de obras como "Mein

Kampf" o "El libro Guiness de los récords" en el clúster principal. En este sentido sería necesario estudiar cómo identificar mejor el objeto de interés y prevenir resultados fuera del concepto de lo que se entiende como canon literario.

Un repaso minucioso de C_3 permite detectar también ausencias notables. La razón se debe a la existencia de inconsistencias en la asignación de las clases adecuadas a los ítems. Se utiliza frecuentemente la clase "libro" para las obras actuales, o bien "obra escrita" para otros muchos casos. También se debe al nivel de especificidad en las tipificaciones, asignándose, en muchos casos, clases de un mayor nivel de detalle dentro de la clase "obra literaria". Se requeriría profundizar en la selección precisa de las clases y subclases dentro del esquema de conocimiento de Wikidata. De este modo se tendría en cuenta la tendencia a la desorganización e inconsistencia cuando se recorren más elementos de la taxonomía de clases. Esto requiere meticulosos procesos de validación y eliminación de ruido. Puede advertirse una situación similar para las obras que se presentan en forma de sagas o series, para las cuales los resultados se distribuyen entre la obra individual y la serie completa, según hayan sido descritas. Este funcionamiento agregado dificulta su identificación precisa. "El señor de los anillos" no aparece como tal, ya que se vincula con las clases "trilogía literaria" (Q13593966) y "novela" (Q1667921), pero sí alguno de sus volúmenes. Tampoco queda claro si en Don Quijote están reunidas ambas partes de la obra. La dualidad obra-agregación señala la conveniencia de establecer procedimientos para asignar ranking a las obras que aparecen individualmente y agrupadas, como "El Génesis" y "La Biblia" o cada uno de los libros de sagas y series de novelas como las de Sherlock Holmes. De la misma forma, el canon parece perjudicar a las obras poéticas y a los cuentos, seguramente por las condiciones de su edición y publicación en numerosas y variadas recopilaciones.

Por otro lado, la propia consistencia de los datos consignados en Wikidata dificulta la exploración sistemática de otros aspectos como autores, géneros, temas, etc. Sin entrar en un análisis detallado, se observa en el conjunto de datos analizado una importante variabilidad en el uso de propiedades descriptivas, y un uso muy heterogéneo de ellas, al no existir pautas de descripción consensuadas, y usar diferentes niveles de detalle en la asignación de categorías para las propiedades que deberían corresponder a vocabularios controlados dentro de la taxonomía de clases disponible. Aun así, conforme las obras tienen mayor Wiki3DRank, o pertenecen al clúster principal o secundario, tienen mayor calidad descriptiva.

En relación con las obras recientes se detecta la presencia de numerosos libros superventas, que no suelen entenderse como obras reconocidas por la crítica literaria más convencional, pero sí por nuevas corrientes de estudio sobre "best sellers canónicos" (Muñoz Rico y otros, 2020). Ejemplos de ello son la saga de Harry Potter o Los juegos del hambre. Esto sugiere cierta dificultad para captar correctamente, mediante el mecanismo de cálculo usado en este estudio, la relevancia de estas obras nacidas en un contexto de grandes fenómenos de difusión en la cultura de masas. Sin embargo, las obras de los siglos XIX y XX parecen encajar con el modelo utilizado en la investigación.

La distribución temporal de las obras muestra obras de todas las épocas. No obstante, predominan las obras de los siglos XX y XXI, en consonancia con el surgimiento de un mercado masivo para el libro y el auge de los medios de comunicación de masas. Los datos de C_2 y C_3 podrían representar la idea genérica de "clásicos actuales y de todos los tiempos". Cuanto mayor es la distancia temporal con las obras recogidas, más habitual es que predominen solo aquellas con cierta relevancia y cuyo interés ha sido decantado por el paso del tiempo.

Los idiomas de las obras de C_3 reflejan una cierta variedad lingüística que se acerca a la idea de canon global. El clúster equilibra la tendencia "eurocéntrica" de cánones literarios propuestos por otros autores. No obstante, sigue reflejando la disparidad de la difusión de las lenguas asociadas a los imperios coloniales y las potencias económicas. Pese a ello, permite una mayor oportunidad de destacar a las lenguas muertas y no occidentales. Conviene señalar que, fuera de los idiomas occidentales dominantes, la presencia de obras en otros idiomas se relaciona con la antigüedad remota.

6. CONCLUSIONES E INVESTIGACIONES EN CURSO

Los resultados obtenidos muestran que el uso combinado de Wikidata y Wikipedia puede utilizarse como fuente de datos para definir un canon literario y por lo tanto la hipótesis planteada al inicio de este trabajo quedaría verificada. La observación y medición de la atención prestada por la comunidad Wikimedia a las obras literarias permite conocer algo más sobre su relevancia y visibilidad, y actuar de complemento a la propuesta de canon que realizan los medios, la academia y la industria editorial. He aquí otra fuente más para debatir so-

bre un canon plural, abierto y múltiple, que recoge el resultado de muchas voces autónomas y actores individuales.

A diferencia de una encuesta de gustos, Wikipedia refleja actos individuales para cuidar la información sobre la literatura poniendo esfuerzo en el enriquecimiento de artículos y descripciones. El estudio presentado guarda, en cierto modo, paralelismos con estudios sobre la traducción, las tiradas, reediciones y ventas en literatura, todos ellos terrenos sobre los que no existen fuentes longitudinales fácilmente accesibles y procesables. Los datos obtenidos muestran hasta cierto punto que se amoldan al modelo conceptual LRM, que diferencia la Obra de sus Expresiones y Manifestaciones, por lo que es posible obtener un inventario colaborativo de literatura en cada idioma y global por agregación de todas las ediciones de Wikipedia.

La visibilidad en Wikipedia de las obras literarias concuerda todavía bastante con el canon escolar y académico de los manuales de literatura universal. Pese a ello, se percibe una tendencia paulatina hacia una mayor "presentización", en donde ganan espacio las obras de éxito masivo y transmedia de los siglos XX y XXI. El análisis de las ediciones en cada idioma, combinado con la literatura producida en cada lengua, permite dibujar de forma ágil espacios geográficos de proximidad cultural e influencia. Este efecto podría reducirse introduciendo una nueva variable que considerara la fecha de publicación o creación de las obras y que incremente el valor de Wiki3DRank de las obras más antiguas, o algunos otros atributos de dominio que puedan relacionarse con aspectos de calidad e impacto.

Por otra parte, somos conscientes de que la selección de ítems y artículos analizados (aquellos clasificados directamente como "obra literaria" en Wikidata) abarca tan solo una porción del universo real de este tipo de obras. Por este motivo, es preciso diseñar mecanismos de exploración que analicen otras clases utilizadas para tipificar las obras literarias. Es imprescindible tener en cuenta la validación del caos organizativo de la taxonomía de clases producido por la descripción colaborativa. Por dicho motivo, es esencial comprender que la metodología utilizada se basa exclusivamente en los datos existentes de aquellas obras literarias identificadas de forma explícita como tales en Wikidata.

Lo anterior también implica que quedarían fuera aquellas obras que estén presentes en otros cánones literarios elaborados de forma subjetiva según el criterio del autor, pero que no estén presentes en Wikidata. Además, como muestran los resultados, la cobertura de una obra en las diferentes ediciones de Wikipedia es determinante para establecer su posición en el ranking calculado mediante Wiki3DRank. Esto significa que las obras con una difusión reducida o limitada al ámbito de un idioma, serían poco representativas en un canon universal. No obstante, el método propuesto seguiría siendo válido para definir un canon literario para un idioma específico.

También parece necesario el uso de métricas relacionadas con la profundidad editorial y la actividad de los editores en los artículos sobre obras literarias en Wikidata/Wikipedia. Además, deben elaborarse de manera que permitan captar de una forma más detallada la atención y el esfuerzo puesto en cada artículo e ítem, como medida indirecta de su valor.

Es necesario mencionar, que la agregación de la transformación logarítmica de N_{Wikis}, N_{Props} y N_{Words} para calcular Wiki3DRank ofrece resultados coherentes respecto a la hipótesis planteada. Como línea de trabajo futura se plantea un cálculo alternativo en el que las obras se representen como vectores. Los componentes de dichos vectores se corresponderían con las transformaciones logarítmicas de dichas variables. Mediante este método Wiki3DRank podría obtenerse a partir del cálculo del módulo del vector correspondiente de cada obra.

Este estudio abre la puerta al uso de Wikipedia para la extracción de una propuesta de canon cultural transmedia. Dicho canon incluiría a los otros grandes formatos de ficción narrativa, como el cine, cómic y televisión. Todo ello permitiría profundizar en sus relaciones, puesto que se consumen y publican en ciclos iterativos de versiones, adaptaciones, actualizaciones y recreaciones, lo cual tampoco es un fenómeno nuevo del todo, aunque sí lo sea su ritmo e impacto.

7. NOTAS

1 https://wdo.wmcloud.org/topical_coverage
2 https://www.wikidata.org/wiki/Wikidata:WikiProject_Books
3 https://wikirank.net
4 https://github.com/j-pastor/wd-literary-canon
5 https://xtools.wmflabs.org
6 https://orangedatamining.com
7 Se hace preciso indicar que una primera versión del dataset para este trabajo, obtenido el 20 de noviembre de 2021, únicamente incluía 89.744 ítems.
8 Únicamente 94 de las obras, propuestas por la autora, están catalogadas expresamente como "Obra literaria" en Wikidata, y son las únicas que podrían aparecer en nuestro estudio. De ellas el C3 recoge 92 obras (el 97%). Teniendo en cuenta todas las obras de la autora, el grado de concordancia sería del 65%.

8. REFERENCIAS

Algee-Heweitt, M., Allison, S., Gemma, M., Heuser, R., y Moretti, F. (2018). Canon/archivo: dinámicas de largo alcance y campo literario. En F. Moretti (Ed.), *Literatura en el laboratorio: canon, archivo y crítica literaria en la era digital*, 131-181. Gedisa.

Arthur, D., y Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027-1035.

Bianchini, C., y Sardo, L. (2022). Wikidata : a new perspective towards universal bibliographic control. *JLIS*, 13(1). DOI: https://doi.org/10.4403/jlis.it-12725

Bourdieu, P. (1995). *The Rules of Art: Genesis and Structure of the Literary Field*. Stanford University Press.

Boxall, P., y Mainer, J. C. (2016). *1001 libros que hay que leer antes de morir: relatos e historias de todos los tiempos* (7ª ed.). Grijalbo.

Claes, F., y Tramullas, J. (2021). Estudios sobre la credibilidad de Wikipedia: una revisión. *Área Abierta*, 21(2), 187-204. DOI: https://doi.org/10.5209/arab.74050

Damrosch, D. (2009). *How to read world literature*. Wiley-Blackwell.

Ding, C., y He, X. (2004). K-means clustering via principal component analysis. *Twenty-First International Conference on Machine Learning - ICML '04*, 29. DOI: https://doi.org/10.1145/1015330.1015408

Haider, J., y Sundin, O. (2019). *Invisible Search and Online Search Engines: The Ubiquity of Search in Everyday Life* (1.ª ed.). Routledge. DOI: https://doi.org/10.4324/9780429448546

Hartigan, J. A., y Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 28(1), 100. DOI: https://doi.org/10.2307/2346830

Hill, B., y Shaw, A. (2020). The Most Important Laboratory for Social Scientific and Computing Research in History. En J. Reagle y J. Koerner (eds.), *Wikipedia @ 20: Stories of an Incomplete Revolution*. The MIT Press. DOI: https://doi.org/10.7551/mitpress/12366.001.0001

Hube, C., Fischer, F., Jäschke, R., Lauer, G., y Thomsen, M. R. (2017). *World Literature According to Wikipedia: Introduction to a DBpedia-Based Framework*. arXiv. Disponible en: http://arxiv.org/abs/1701.00991

Jemielniak, D., y Wilamowski, M. (2017). Cultural diversity of quality of information on Wikipedias. *Journal of the Association for Information Science and Technology*, 68(10), 2460-2470. DOI: https://doi.org/10.1002/asi.23901

Lemus-Rojas, M., y Pintscher, L. (2018). Wikidata and Libraries: Facilitating Open Knowledge. En M. Proffitt (ed.), *Leveraging Wikipedia: Connecting Communities of Knowledge*, 143-158. IL: ALA Editions. Disponible en: https://scholarworks.iupui.edu/handle/1805/16690

Lewoniewski, W., Węcel, K., y Abramowicz, W. (2019). Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics. *Computers*, 8(3), 60. DOI: https://doi.org/10.3390/computers8030060

Minguillón, J., Lerga, M., Aibar, E., Lladós-Masllorens, J., y Meseguer-Artola, A. (2017). Semi-automatic generation of a corpus of Wikipedia articles on science and technology. *El Profesional de la Información*, 26(5), 995-1004. DOI: https://doi.org/10.3145/epi.2017.sep.20

Miquel-Ribé, M. (2019). *The Sum of Human Knowledge? Not in One Wikipedia Language Edition*. Wikipedia@20. Disponible en: https://wikipedia20.mitpress.mit.edu/pub/26ke5md7/release/15

Miquel-Ribé, M., y Laniado, D. (2018). Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. *Frontiers in Physics*, *6*, Article 54. DOI: https://doi.org/10.3389/fphy.2018.00054

Miquel-Ribé, M., y Laniado, D. (2021). The Wikipedia Diversity Observatory: helping communities to bridge content gaps through interactive interfaces. *Journal of Internet Services and Applications*, 12(1), 10. DOI: https://doi.org/10.1186/s13174-021-00141-y

Moretti, F. (2013). *Distant reading*. Verso.

Muñoz Rico, M., García Rodríguez, A., y Cordón García, J. A. (2020). Hacia una teoría del bestseller canónico: la constitución de un modelo estructural. *Revista General de Información y Documentación*, 30(1), 149-165. DOI: https://doi.org/10.5209/rgid.69673

Nielsen, F. Å. (2019). *Wikipedia research and tools: Review and comments*. Disponible en: http://www2.imm.dtu.dk/pubdb/edoc/imm6012.pdf

Piscopo, A., y Simperl, E. (2018). Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata. *Proceedings of the ACM on Human-Computer Interaction*, 2, 1-18. DOI: https://doi.org/10.1145/3274410

Reagle, J., y Koerner, J. (eds.). (2020). *Wikipedia @ 20: Stories of an Incomplete Revolution*. The MIT Press. DOI: https://doi.org/10.7551/mitpress/12366.001.0001

Reznik, I., y Shatalov, V. (2016). Hidden revolution of human priorities: An analysis of biographical data from Wikipedia. *Journal of Informetrics*, 10(1), 124-131. DOI: https://doi.org/10.1016/j.joi.2015.12.002

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7

Shatnawi, R. (2015). Deriving metrics thresholds using log transformation. *Journal of Software: Evolution and Process*, 27(2), 95-113. DOI: https://doi.org/10.1002/smr.1702

Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., y Szekely, P. (2022). A study of the quality of Wikidata. *Journal of Web Semantics*, 72, 100679. DOI: https://doi.org/10.1016/j.websem.2021.100679

Skiena, S. S., y Ward, C. (2014). *Who's bigger? where historical figures really rank*. Cambridge University Press.

Venuti, L. (2008). Translation, interpretation, canon formation. En A. Lianeri y V. Zajko (eds.), *Translation and the Classic: Identity as Change in the History of Culture*, 27-51. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199288076.001.0001

Zschirnt, C. (2011). *Libros: todo lo que hay que saber* (1a ed). Taurus.

ANEXO: LISTADO DEL CANON DE OBRAS LITERARIAS

Ítem	Título	Idioma	Fecha	Clúster	Wiki3DRank	N_{Wikis}	N_{Props}	N_{Words}
Q9184	Génesis	hbo	-	C3	21,8743	148	125	168.391
Q8275	Ilíada	grc	-800	C3	21,5304	113	132	147.831
Q41567	Hamlet	en	1602	C3	21,2433	136	93	130.612
Q83186	Romeo y Julieta	en	1597	C3	20,9841	83	94	162.665
Q480	Don Quijote de la Mancha	es	1614	C3	20,9273	95	115	110.128
Q8279	Shahnameh	fa	1000	C3	20,7261	108	99	92.005
Q6511	Ulises	en	1922	C3	20,6773	105	72	123.428
Q43361	Harry Potter y la piedra filosofal	en	1997	C3	20,6678	80	85	135.803
Q92640	Alicia en el país de las maravillas	en	1862	C3	20,5523	84	91	107.783
Q127149	Lolita	en	1955	C3	20,4429	173	63	67.843
Q130283	Macbeth	en	1623	C3	20,3785	97	72	99.014
Q170583	Orgullo y prejuicio	en	1813	C3	20,3396	123	68	79.634
Q19786	Antiguo Testamento	-	0	C3	20,3151	60	140	77.301
Q74287	El hobbit	en	1937	C3	20,3062	49	88	148.085
Q8258	Las mil y una noches	fa ar	0	C3	20,2569	42	120	120.556
Q41542	Drácula	en	1897	C3	20,2473	142	68	62.963
Q9190	Éxodo	hbo	0	C3	20,1433	93	108	54.648
Q161531	Guerra y paz	ru fr	1869	C3	20,1038	61	78	109.886
Q208460	1984	en	1949	C3	20,074	48	86	122.550
Q165318	Crimen y castigo	ru	1866	C3	20,0696	60	75	112.190
Q60220	Eneida	la	-100	C3	20,0588	54	84	110.065
Q140527	Los tres mosqueteros	fr	1844	C3	19,9853	137	68	50.207
Q150827	Frankenstein o el moderno Prometeo	en	1818	C3	19,903	65	69	95.304
Q326909	Los Buddenbrook	de	1901	C3	19,89	173	39	62.443
Q46758	Harry Potter y las reliquias de la Muerte	en	2007	C3	19,8675	43	76	125.423
Q42040	Apocalipsis	grc	-	C3	19,8595	50	98	83.496
Q16438	Decamerón	it	1348	C3	19,7643	89	64	65.521
Q37293	Ramayana	sa	-	C3	19,7137	35	108	92.856
Q147787	Ana Karenina	ru	1877	C3	19,6634	69	76	64.285
Q8269	El relato de Genji	ja	1010	C3	19,6483	61	87	62.555
Q180736	Los miserables	fr	1862	C3	19,6159	59	67	80.985
Q164974	Oliver Twist	en	1837	C3	19,599	68	70	66.319
Q191838	El conde de Montecristo	fr	1844	C3	19,5688	94	60	54.393
Q483034	Robinson Crusoe	en	1719	C3	19,5634	68	81	55.414
Q183157	Los hermanos Karamazov	ru	1880	C3	19,5389	45	66	99.269
Q184742	Las metamorfosis	la	100	C3	19,5073	105	60	45.844
Q104871	El sueño de una noche de verano	en	1595	C3	19,5068	66	63	69.090
Q899334	El tambor de hojalata	de	1959	C3	19,4809	276	33	30.653
Q47209	Harry Potter y la cámara secreta	en	1998	C3	19,4587	43	80	79.226

Ítem	Título	Idioma	Fecha	Clúster	Wiki3DRank	N_{Wikis}	N_{Props}	N_{Words}
Q1396889	Rebelión en la granja	en	1945	C3	19,4242	42	84	74.635
Q178869	Cien años de soledad	es	1967	C3	19,3977	24	74	141.675
Q4577	Libro de Job	he	-	C3	19,3854	72	73	48.573
Q188538	El maestro y Margarita	ru	1967	C3	19,377	56	56	80.091
Q123397	República	grc	-379	C3	19,296	26	78	112.496
Q25338	El Principito	fr	1942	C3	19,2799	33	109	63.135
Q181488	Los viajes de Gulliver	en	1726	C3	19,2651	51	65	67.793
Q86440	La tempestad	en	1623	C3	19,2644	57	55	71.585
Q8272	Poema de Gilgamesh	akk	-2100	C3	19,2484	15	98	144.456
Q46751	Harry Potter y el cáliz de fuego	en	2000	C3	19,2333	42	78	66.347
Q190192	Dune	en	1965	C3	19,2292	64	49	69.065
Q463108	La historia interminable	de	1979	C3	19,1845	156	31	42.723
Q181598	El rey Lear	en	1606	C3	19,1646	37	67	81.432
Q523076	Mujercitas	en	1869	C3	19,1377	216	37	24.838
Q185118	La isla del tesoro	en	1883	C3	19,1134	72	62	43.466
Q70784	Viaje al Oeste	zh	1592	C3	19,0994	92	58	35.925
Q6911	Diario de Ana Frank	nl	1947	C3	19,0917	34	70	78.719
Q46887	Harry Potter y el misterio del príncipe	en	2005	C3	19,0734	33	77	72.426
Q174596	Moby Dick	en	1851	C3	19,0665	47	67	58.441
Q202975	Cumbres Borrascosas	en	1847	C3	19,0663	64	57	50.587
Q48244	Mi lucha	de	1925	C3	19,0496	37	81	60.189
Q219552	Grandes esperanzas	en	1861	C3	19,0371	46	55	70.372
Q47598	Harry Potter y el prisionero de Azkaban	en	1999	C3	19,0178	39	77	58.233
Q41490	Levítico	hbo	-	C3	18,9899	54	96	33.118
Q206400	El mercader de Venecia	en	1600	C3	18,9734	48	52	66.919
Q131554	Cantar de los nibelungos	gmh	1203	C3	18,9714	39	58	73.495
Q26833	Otelo	en	1604	C3	18,9161	38	72	57.645
Q191380	Nuestra Señora de París	fr	1831	C3	18,907	68	52	44.472
Q2222	La cabaña del tío Tom	en	1852	C3	18,8678	36	53	78.269
Q80817	Harry Potter y la Orden del Fénix	en	2003	C3	18,8295	31	74	62.710
Q183565	Veinte mil leguas de viaje submarino	fr	1869	C3	18,8262	59	59	41.668
Q214371	El gran Gatsby	en	1925	C3	18,8236	46	64	48.975
Q79762	El Silmarillion	en	1977	C3	18,8183	36	55	71.826
Q1219561	La vuelta al mundo en ochenta días	fr	1872	C3	18,8099	62	56	41.097
Q217352	El extraño caso del doctor Jekyll y el señor Hyde	en	1886	C3	18,7985	90	45	34.855
Q81689	El código Da Vinci	en	2003	C3	18,7907	24	70	81.562
Q8065468	Las aventuras de Pinocho	it	1883	C3	18,7697	49	67	41.696
Q134425	Dào Dé Jing	lzh	-	C3	18,769	44	72	43.125
Q82464	El retrato de Dorian Gray	en	1890	C3	18,7638	44	55	55.924
Q191663	Los cuentos de Canterbury	enm	1387	C3	18,7193	46	56	50.317
Q212340	Para matar a un ruiseñor	en	1960	C3	18,7168	30	52	81.839

Ítem	Título	Idioma	Fecha	Clúster	Wiki3DRank	N_{Wikis}	N_{Props}	N_{Words}
Q172850	El nombre de la rosa	it	1980	C3	18,7041	41	53	58.536
Q215894	Cándido o El optimismo	fr	1759	C3	18,6591	39	46	67.511
Q2870	Lo que el viento se llevó	en	1936	C3	18,6532	29	49	84.120
Q81240	Libro de los Jueces	-	-	C3	18,649	43	84	33.595
Q28754	El paraíso perdido	en	1667	C3	18,6123	36	57	56.438
Q131719	El Príncipe	it	1532	C3	18,6091	19	72	82.696
Q48203	Epístola a los Romanos	grc	-	C3	18,6073	32	84	42.962
Q326914	Las aventuras de Tom Sawyer	en	1876	C3	18,5336	40	57	47.078
Q45192	El sabueso de los Baskerville	en	1902	C3	18,532	53	52	39.053
Q182961	Jane Eyre	en	-	C3	18,5278	39	56	48.816
Q464928	En busca del tiempo perdido	fr	1927	C3	18,514	41	46	55.610
Q221211	Noche de reyes	en	1623	C3	18,5064	47	47	47.287
Q148643	Edipo rey	grc	-	C3	18,4991	37	46	60.559
Q130295	El maravilloso mago de Oz	en	1900	C3	18,4962	48	52	41.527
Q274744	Sentido y Sensibilidad	en	1811	C3	18,4813	37	46	59.490
Q241077	Antígona	grc	-	C3	18,4559	40	50	49.536
Q80038	Libro de Rut	he	-	C3	18,4538	37	85	31.629
Q128608	Epístola a los hebreos	grc	-	C3	18,4447	29	79	42.679
Q215410	Las aventuras de Huckleberry Finn	en	1885	C3	18,4284	51	55	34.607
Q308918	Historia de dos ciudades	en	1859	C3	18,4166	30	51	61.781
Q193417	Madame Bovary	fr	1857	C3	18,4141	38	55	45.485
Q210784	El idiota	ru	1869	C3	18,3968	35	48	55.352
Q208002	La Comunidad del Anillo	en	1954	C3	18,3767	34	50	53.610
Q213019	La guerra de los mundos	en	1898	C3	18,3757	35	49	53.112
Q48922	Orlando furioso	it	1532	C3	18,3561	34	35	74.398
Q214132	Diez negritos	en	-	C3	18,3528	34	43	60.670
Q80355	Primera Epístola a los Corintios	-	54	C3	18,3266	32	82	33.232
Q19871	Esperando a Godot	cy fr	1952	C3	18,3064	24	59	59.469
Q11678	Los juegos del hambre	en	2008	C3	18,2417	25	49	64.315
Q469690	Mansfield Park	en	1814	C3	18,2274	33	35	67.343
Q1751870	Juego de Tronos	en	1996	C3	18,2126	21	53	68.362
Q212898	La montaña mágica	de	1924	C3	18,202	30	40	63.224
Q41675	Libro Guinness de los récords	en	1955	C3	18,1776	21	88	40.052
Q151883	Las penas del joven Werther	de	1774	C3	18,1544	37	49	40.328
Q11829	Hansel y Gretel	de	1812	C3	18,1423	35	60	34.471
Q29478	Fausto	de	1832	C3	18,1368	33	52	41.780
Q219457	Viaje al centro de la Tierra	fr	1864	C3	18,0712	56	45	26.890
Q208971	1Q84	ja	2010	C3	18,0693	84	32	25.088
Q191949	Un mundo feliz	en	1932	C3	18,0567	25	48	54.544
Q6113985	Upanishad	sa	-	C3	18,0546	13	87	56.286
Q185427	Cantar de Roldán	fro	1100	C3	18,0203	26	62	39.392
Q205875	Tartufo	fr	1669	C3	18,006	38	45	36.818
Q332387	La fierecilla domada	en	1623	C3	18,003	39	41	39.198
Q223880	Emma	en	1815	C3	17,9946	29	37	57.287

Ítem	Título	Idioma	Fecha	Clúster	Wiki3DRank	N_{Wikis}	N_{Props}	N_{Words}
Q233562	La riqueza de las naciones	en	1776	C3	17,9914	16	53	70.914
Q11834	El Gato con Botas	fr	1695	C3	17,988	34	49	37.070
Q183883	El Guardián entre el Centeno	en	1951	C3	17,9853	19	64	49.771
Q28306	Danza de dragones	en	2011	C3	17,9833	47	35	37.366
Q240617	Papá Goriot	fr	1835	C3	17,9785	37	40	41.245
Q50948	Eugenio Oneguin	ru	1825	C3	17,9759	23	54	48.557
Q471005	La isla misteriosa	fr	1874	C3	17,9516	44	40	33.906
Q36097	El proceso	de	1925	C3	17,9385	23	55	45.937
Q726254	El maravilloso viaje de Nils Holgersson	sv	1907	C3	17,9183	99	36	16.353
Q329989	Los endemoniados	ru	1872	C3	17,9035	30	40	46.905
Q128620	Epístola a los Gálatas	grk	-	C3	17,8915	24	79	29.454
Q192649	Rojo y negro	fr	1830	C3	17,8865	38	43	34.158
Q181937	I Ching	och	-	C3	17,8787	17	57	55.707
Q202009	Fahrenheit 451	en	1953	C3	17,8771	30	48	38.227
Q62407	Madre Coraje y sus hijos	de	1949	C3	17,8654	38	32	44.592
Q333179	Persuasión	en	1818	C3	17,8533	32	35	47.725
Q271764	El señor de las moscas	en	1954	C3	17,8435	26	49	41.591
Q26505	El viejo y el mar	en	1952	C3	17,8329	18	72	40.053
Q155980	Libro de la Sabiduría de Jesús ben Sira	he	-	C3	17,8191	26	61	32.733
Q237572	Como gustéis	en	1623	C3	17,8018	28	47	38.686
Q6507	Finnegans Wake	en	1939	C3	17,7776	32	37	41.919
Q11859	La sirenita	da	1837	C3	17,7496	30	61	26.594
Q123808	Segunda Epístola a los Corintios	-	-	C3	17,7489	28	79	22.017
Q131115	Primera Epístola a los Tesalonicenses	grc	50	C3	17,7452	22	76	28.735
Q212746	Crónica anglosajona	ang	892	C3	17,7074	21	48	45.458
Q408673	Epístola a los Efesios	grc	-	C3	17,6955	22	79	26.315
Q207332	Sin novedad en el frente	de	1929	C3	17,6897	22	48	42.716
Q179021	El alquimista	pt	1988	C3	17,665	19	67	34.534
Q215983	Las uvas de la ira	en	1939	C3	17,6405	24	51	35.255
Q131180	Primera epístola a Timoteo	he	-	C3	17,6268	24	75	23.794
Q233780	Panchatantra	sa	-299	C3	17,6059	23	55	32.942
Q206870	Doctor Zhivago	ru	1957	C3	17,5588	19	45	45.910
Q47228	Kama sutra	sa	-	C3	17,5409	11	81	42.162
Q51613	Epístola a los Filipenses	-	54	C3	17,5347	19	77	26.429
Q565638	La pequeña Dorrit	en	1857	C3	17,4436	25	21	65.807
Q655717	Tractatus logico-philosophicus	en	1921	C3	17,3589	17	37	50.561
Q131107	Segunda Epístola a los Tesalonicenses	grc	-	C3	17,3222	18	74	23.394
Q131104	Epístola a Filemón	-	-	C3	17,2192	14	77	25.704
Q808428	Evangelio de Bernabé	it es	-	C3	16,9396	10	34	59.060