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

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

Authors' english version of the paper originally published in Spanish at REVISTA ESPAÑOLA DE DOCUMENTACIÓN CIENTÍFICA (ISSN-L:0210-0614) in July 2023.

How to cite / Citation: Pastor-Sánchez, J. A., Saorín, T., Baños-Moreno, 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

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 pres-ence 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.

* Departamento de información y documentación. Universidad de Murcia, https://orcid.org/0000-0002-1677-1059
** Departamento de información y documentación. Universidad de Murcia, https://orcid.org/0000-0001-9448-0866
*** Departamento de información y documentación. Universidad de Murcia / Odilo Europa, https://orcid.org/0000-0001-9137-1330

1. INTRODUCTION

This work is based on a very simple working hypothesis: Could we use Wikidata and Wikipedia as a source to identify a global literary canon? A literary canon is understood to be a cultural selection strongly affected by the point of view of the dominant group which has established it. Therefore, it is contended from the different positions which have emerged from various geographical, identity-based and cultural peripheries which seek to broaden the vision of the Western literary canon popularized by literary critic Harold Bloom or the essential authors and works contained in school textbooks and also in syllabus of higher education. In addition, any canon which is taken as a benchmark is not immutable and is subject to an endless process of attention, oblivion and recovery over centuries, eras, and decades. As the canon is a changing cultural construction, could the autonomous and unplanned activity of the community of Wikidata and Wikipedia editors be used to obtain another, complementary, point of view? These are communities involved in writing and categorizing articles in all languages and in defining descriptive data of all kinds. Supported by the idea of a neutral point of view and decentralized and multilingual collaboration, the Wikimedia ecosystem could be a candidate source for obtaining results that have not been directly mediated by any authors, academies, nations, or stakeholders of any type.

Studies on thematic coverage in Wikipedia have addressed various fields, such as science, biographies, cultural heritage, mass culture, and current social issues (Hill; Shaw, 2020; Reznik; Shatalov, 2016; Minguillón et al., 2017). However, there is not yet a good, broad overview of Wikipedia's participation in the knowledge of literary works or printed works over time. That is ground traditionally covered by library catalogs, reference works on the history of literature and books, magazines providing literary criticism or suggested reading, or bibliographic repertories. In addition, since Wikidata was launched in 2012, an infrastructure has been available to store factual data on Wikipedia articles in a structured way. There is an active movement interested in establishing the procedures to use Wikidata as a multipurpose bibliographic open database: references in Wikipedia itself, bibliometric analysis, universal repertory, etc. In short, it is possible to perceive growing interest and interrelation between the universe of books and Wikimedia projects.

Considering the above, it is hypothesized that Wikipedia and Wikidata can be used jointly as data sources to build a literary canon. Consequently, this work establishes a series of objectives and a working methodology to determine the necessary data to be extracted, the processes to perform such extraction and the way in which they should be used to define an indicator to identify and weight those works that should be part of such a literary canon.

2. COLLABORATIVE ENCYCLOPAEDIC DATA ABOUT BOOKS AND LITERARY CANON

The omnipresence of Wikipedia as a multi-domain information source is commonplace in studies on collaborative content production (Reagle; Koerner, 2020) and on digital-information use practices. Wikipedia has become enormously well-known and present in our daily lives. In this way. Wikipedia is relevant not only because of the volume of general and local content, but also due to the place it occupies in the daily practices of using the Internet to obtain information, including its invisible role as a component of the answers provided by search engines and voice assistants (Haider; Sundin, 2019).

A significant amount of Wikipedia content is devoted to cultural objects and their context: monuments, paintings, plays, authors, music albums, books, movies, sculptures, etc. A marked local component has been identified in this content, since each cultural community has a different heritage, linked to language or territory (Miquel-Ribé; Laniado; 2018). Authors call it Cultural Context Content and calculate it to represent 25% of the main encyclopedias. The *Wikipedia Diversity Observatory* project indicates, in its *Topical Coverage* section, that 1-2% of the articles on the main Wikipedias correspond to the generic topic of "books"[1]. In this context, Wikipedia provides worthwhile information on literary works, taking into account as well that it is not limited to a single discourse, since each language community writes articles on literary works that incorporate their own cultural differences (Jemielniak; Wilamowski, 2017). Despite the exceptional size of Wikipedia in English, and the fact that it is often seen as a "catch-all encyclopedia", there are considerable content gaps between editions, especially in local content (Miquel-Ribé, 2019). Most of the great works of literature, which are part of the cultural canon and historical traditions, have merited the drafting of detailed encyclopedic articles. Before delving into the treatment of literary works on Wikipedia, it should be noted that the free online encyclopedia is not intended to be a catalog of books. Those appearing on Wikipedia must be "notable" entities with enough encyclopedic relevance.

Wikipedia has a clear tendency to pay more attention to the phenomena of mass culture and its constant production of new trends and releases. This is reflected, in the case of books, in a strong focus on recent popular or well-regarded literary works and not only on classical and established literature, which in this paper we refer to as the "universal literary canon".

Articles about books in Wikipedia show great variability in length and treatment. They usually include a brief summary of the plot, explain the writing and publishing context, talk about the characters, style, literary technique, and reception in its time. They also tend to contain a descriptive template (infobox) that excerpts the essential bibliographic data, links to digital libraries to access the full text of works in the public domain, and a categorization system.

In the context of the literary canon, it can be observed that there is greater coverage of authors compared to that of works. Studies of people are a frequent focus of research about Wikipedia from the perspective of network analysis (Hube, et al., 2017). There is not always an individual article on Wikipedia about each of the works of some canonical authors, however it is common to find basic information (normally an enumerative list with a brief commentary) about their main works. It is also possible to find articles about the fictional universes themselves: fictional characters, objects, and settings.

The articles from the encyclopedia usually conforms to the abstract level of Work (Work) in accordance with the conceptualization of the library reference model LRM-FRBR. The correct modeling of these levels is a area of interest to the library community involved in linked open data (Lemus-Rojas; Pintscher, 2018), so that

Wikipedia and Wikidata become a more precise bibliographic information space. Furthermore, the very definition of what a literary work is remains an open concept. In a very broad and historical sense it is understood as “belles-lettres”, including essays and philosophical texts, and in a more modern sense, as creative fiction (Damrosch, 2009: 6).

Although each article in each encyclopedia is an individual content item, edited and revised by its own community of editors, through the Wikidata knowledge base they are interconnected, so that there is a single entity to represent a work and link it to the articles in the languages in which it exists.

The relationship between Wikipedia and the literary canon has not been specifically studied. It would be framed within the schools of thought on literature in which the focus is placed on the "literary system", or "literary field" in the terminology of Bourdieu (1995), and therefore more on its impact and reception over time and less on its intrinsic literary quality. The study of reviews and critiques published in journals and literary supplements, and the presence of authors and works in monographs, dictionaries, and literary encyclopedias, is one of the methodologies used to study the literary field. Furthermore, the "Distant reading" research movement (Moretti, 2013), approaches the study of the literature by expanding the usual set of sources and data. In this manner, advantage is taken of accessibility to most of the literary production of recent centuries, enabling large volumes of data on literary activity to be processed, including computerized analysis of the full texts themselves. In this sense, Wikipedia and its articles, in each of the languages in which it exists, is a wide, dynamic source of data. The exploration of new, interesting sources is of interest as a starting point to define and understand the dimensions of a canon, as well as the criteria to study it (Algee-Hewitt, 2018). In the case of Wikipedia, we also have a very large and, at the same time, clearly delimited amount of curated content which, above all, is clearly marked and codified, in formats that are easy to process and with APIs and parameterized query systems, in particular due to having the information structured in Wikidata.

The cluster of more than 250 Wikipedias in different languages is aligned with the field of study of "World Literature" (Damrosch, 2009). This allows the focus to be broadened from a Western canon containing strong bias to one that is broader and more global. It also makes it possible to go beyond the "translated canon", where there is a very strong bias towards languages with large publishing markets, such as English, French and Spanish (Venuti, 2008). As stated before, researchers on Wikipedia are conscious of the "culture gap" between editions local and cultural contents (Miquel-Ribé; Laniado, 2021). Therefore, to explore the global canon as well as in each language, in accordance with Wikipedia, it will be necessary to start from the editions in each language to obtain data that reflect its true nature as a diverse source of information.

Each Wikipedia edition for each language works independently, representing the choices of its editors and its context. However, Wikidata is a common database, produced simultaneously by editors from all languages. It is a unique project whose objective is the creation of a collaboratively produced knowledge graph by editors of any language. Wikidata has a universal scope and models different areas of knowledge through the collaborative and supervised creation of properties. It integrates both data about instances (Charles Chaplin; Azteca Stadium; Mount Everest), as well as properties to establish relationships and collect data (Date of birth; Capacity; Coordinates), and also classes, subclasses, and controlled vocabulary to describe them (Actor; Football stadium; Mountain). Regarding the “books” topic, in a broad sense, there is a Wikiproject in which metadata and description guidelines and other aspects of interest for book description are agreed upon[2].

There are several proposals for the automatic evaluation of quality aspects of Wikipedia content using quantitative approaches, that per se represent one of the most frequent subfields of research about Wikipedia (Nielsen, 2019). Some of them use network analysis techniques based on the resulting graph of links between articles. Others use metrics of the content of the articles: number of words, number of references, incoming links, etc., complemented by the study of the activity of editors, their reputation and collaboration networks. Similar research exists on Wikidata to measure data quality and completeness (Shenoy et al., 2022). Automatic metrics serve as an indirect measure of “expected quality”, probability of quality or credibility (Claes; Tramullas, 2021). Among the applied results of this research intense activity, Wiki3DRank website[3] is a very close case that illustrates how to construct rankings of articles segmented by content type, using aggregated indicators called "Popularity", "Authors’ Interest" (AI) and "Citation Index" (Lewoniewski et al., 2019). The most well-known work on ranking is that of Skiena and Ward (2014), in which historical figures are compared, differentiating between celebrity (current popularity) and gravitas (consolidated popularity).

3. OBJETIVES AND METHODS

Wikidata is a knowledge graph that employs its own RDF-compatible data model. Its primary entities are items, each with a unique identifier starting with the letter "Q." For instance, the book " One Hundred Years of Solitude" by Gabriel García Márquez is represented by the item Q178869, linked to 74 articles across various Wikipedias (Spanish, Japanese, Italian, Russian, etc.). Each item is further described through properties, identified by labels starting with the letter "P." Properties define relationships between elements or refer to literal values (strings, numbers, dates). For example, the book above is stated to have as author (P51) the item Q5878 (the writer García Márquez) and its publication date (P577) as 1967. Wikidata doesn't explicitly define distinct classes apart from other elements. Instead, some elements play a class role by fitting into a taxonomy of classes and subclasses connected through the P279 property (subclass of). Item membership in classes is determined by the P31 property (instance of). This setup allows, to some extent, understanding Wikidata as a

"collaborative ontology" that not only contains primary data but also a sort of formalized schema for organizing knowledge (Piscopo and Simperl, 2018). Within each item, there is a section called "Identifiers" that establishes connections with various external records and databases, such as the VIAF international authority control system (Bianchini and Sardo, 2022).

Based on the considerations presented so far, it is proposed to reuse the available data from both the encyclopedic content of Wikipedia and the structured knowledge base of Wikidata to develop a procedure for defining a literary canon. Consequently, in order to demonstrate the hypothesis introduced in this work, the following general objectives are established:

- Identify the set of encyclopedic data related to literary works from all periods and in any language.
- Validate an automated analytical procedure to establish groupings and rankings of literary works with coverage in various Wikipedia editions.
- Identify representative measures of the impact of each literary work within the Wikimedia ecosystem.
- Analyze the temporal distribution of literary canon works from the perspective of their publication or, alternatively, creation.

The research methodology consisted of four steps:

- Step 1: Selection of the item that would define the class from which to retrieve the items of literary works.
- Step 2: Construction of the dataset.
- Step 3: Aggregation of certain data from the dataset.
- Step 4: Analysis of the aggregation results.

Both the obtained dataset, the aggregated data, and the Python and Orange Data Mining scripts are available for public consultation and reuse[4].

In the first step, the item “Literary work” (Q7725634) was used as the starting class for the exploration of Wikidata. Items related by property P31 to this class were retrieved. The taxonomy of classes used for the bibliographic universe is broad and has significant inaccuracies in its hierarchical organization and application. Only elements with direct assignment to this class were retrieved. Taxonomies derived from items Q471 (book) and Q47461344 (written work) were not considered. This decision was made in order to reduce the presence of results that fell outside the scope of the work (noise) and would have required detailed (almost individual) validation procedures for the recovered classes.

The dataset was built in the second step. This study limited the scope of the dataset to literary works items that have at least one article in a Wikipedia in any language. This criterion of relevance or notability made it possible to extract information only about those works in which an editorial effort –and not just the mere existence of data stored in Wikidata– can be identified. Due to the close interrelationship between the

two projects (encyclopedias and knowledge base), most of the Wikidata items also have a Wikipedia article.

SPARQL queries were launched in the Wikidata Query Service (WDQS), enabling the retrieval of:

- All items defined as instances of the item “Literary work” (Q7725634) with one or more equivalences in Wikipedia editions and a list of all the properties and direct claims used to describe them. This study uses the term "literary work" to refer to each of the retrieved items.
- The languages in which those literary works were written.
- The URLs of the correspondent articles in Wikipedias in different languages. The term "sitelink" refers to each of these URLs.
- The date of publication or inception of the works to which the items refer.
- The English and Spanish title of each work, and, failing that, the title in the original language.
- A complete listing of all Wikidata properties, distinguishing those used in the identifiers section (ID properties).

In addition to WDQS, the Wikimedia Xtools[5] service was utilized, accessed through Python scripts to automate queries. Through the respective API of this service, statistical information about the structure of each article corresponding to the retrieved Wikidata items was retrieved. Thus, for each Wikipedia article, data such as the number of words, references, number of edits, creation and modification dates, external links, etc., were obtained.

It was necessary to carry out a data consolidation process. For example, not all retrieved items included explicit statements about the language of the work (P407) or the date of publication (P577). However, in some cases this information could be obtained by extracting the language in which the original title of the work was registered (P1476) and its inception date (P571). In either case, the dataset indicates the properties used to obtain this information.

In the third stage, the dataset was processed to obtain results with more aggregated data. A Python script was developed for aggregation and the extraction of statistical measures.

For each literary work's item, the following data was aggregated from the previously generated dataset:

- The Wikidata "Q" identifier, in the namespace or prefix "wd:".
- Original language of the work.
- Title or label of the work.
- Date of publication or inception of the work.
- Total number of Wikipedias in which the item is present with its corresponding article (N_{wikis}).

- Total number of statements: a distinction is made between ID properties and the rest of the properties used in claims (N_{props}).
- Total number of words used in all the articles corresponding to the item in the different Wikipedias (N_{words}), calculated with data obtained from Xtools.

For each language the following data was aggregated or calculated:

- Standard language identification code. The different regional variations of the same language were grouped together.
- Number of items retrieved from literary works written in that language.
- Arithmetic mean of the number of Wikipedias in which the items of the literary works of the language are present.
- Arithmetic mean of the number of statements with non-ID properties.
- Arithmetic mean of the number of words in the Wikipedia articles corresponding to the item of the literary work.

To complete this stage, a matrix of languages/Wikipedias was generated that represents the number of articles on literary works in a certain language that are present in each of the different editions of Wikipedia. However, this data has not been explored in this work.

In the last step, the data obtained so far were analyzed with the tool Orange Data Mining[6]. For this, the normalized distribution of the items of each work was represented based on the N_{Wikis}, N_{Props}, and N_{Words} values. Items were clustered using the K-means method (Hartigan & Wong, 1979; Arthur & Vassilvitskii, 2007). The number of clusters was determined by the score obtained through the Silhouette method (Rousseeuw, 1987).

After analyzing the results obtained and studying the distribution N_{Wikis}, N_{Props}, and N_{Words}, a metric that combined the three variables was calculated. This indicator, named as Wiki3DRank, enables the items of literary works to be ranked with a normalized distribution that takes into account the three factors established in the research objectives: *presence in Wikipedias, depth of description in Wikidata,* and *length of the articles in Wikipedia*. Once this was accomplished, it was verified that the Wiki3DRank results were consistent with those obtained in the clustering process.

4. RESULTS

Firstly, data is presented regarding the research question of which and how many literary works, based on Wikimedia community activity, could constitute a universal canon, delimiting a subset from the retrieved literary works in our dataset. Secondly, aspects of literatures in each language are analyzed. Thirdly, there is a presentation of the temporal distribution of those canonical works.

4.1. Universal literary canon according to Wikipedia data: diffusion and editorial effort

This work establishes the presence of an article about a literary work in any Wikipedia as an indispensable condition to consider an item relevant. Therefore, the resulting dataset includes a total of 107,434 Wikidata items[7], defined as instances (P31) of the item-class “Literary work” (Q7725634). Without this filter, the total items retrieved would have amounted to 192,236. This implies that over 44% of the items were discarded. Those “only-data” items can be considered mere "catalog records" and not entities with enough relevance or notability to require an explanatory encyclopedic article. This fact indicates a certain tendency to use Wikidata as a general purpose bibliographic database, as seen in WikiCite project.

The distribution of the items of literary works is defined according to the number of Wikipedias in which they appear (N_{Wikis}), the number of statements in Wikidata (N_{Props}) and the total number of words of their articles in Wikipedia (N_{Words}). Table 1 details some statistical indicators for each variable. The greatest dispersion of values (Cv) occurs for N_{Words} and N_{Wikis}. The three variables have a distribution with a strong positive skewness (Fisher's Skewness Coefficient) and a high kurtosis. A large part of the items in the dataset have low values in each of these variables. This implies a low presence in Wikipedias, less depth in their descriptions and shorter articles.

Variable	Mean	Median	C_v	Min	Max	Skewness	Kurtosis
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

Table 1: Statistical analysis of N_{Wikis}, N_{Props}, N_{Words}. Source: own preparation.

The analysis of correlation between the three variables (Table 2) shows that there is a correlation between N_{Words} and N_{Wikis}. This is obvious: a greater number of Wikipedia editions in which a Wikidata item has an equivalent article, a greater total number of words from that group of articles. This analysis also shows that the lowest correlation occurs between N_{Words} and N_{Props}., that is, between description (data) and the article (text).

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	-

Table 2: Pearson and Spearman correlation coefficients between N_{Wikis}, N_{Props}, N_{Words}. Source: own preparation.

Despite this, there are other items whose values for some of the variables (and even all three) are higher than the rest. These data would allow us to verify the hypothesis of this paper, since the items with higher values than the rest would allow us to identify

the works that could be part of the literary canon. In other words, the items of literary works that are part of the literary canon have a greater presence in different editions of Wikipedia, a higher level of description in Wikidata, and a greater degree of develop of the Wikipedia articles compared to the rest of the works.

What number of works would make up that select group of universally outstanding works? K-means++ was used to group the items into clusters to identify the works that could belong to the literary canon. The results of the Silhouette method indicated the possibility of using K-means++ to obtain two or three clusters. The application of K-Means++ with two clusters identified 1,008 items. This number seems excessive for the idea of a literary canon as a list of works that can be easily covered by one person or 'to take to a deserted island,' although perhaps not so much for creating a select inventory of universal written culture spanning over three millennia.". For this reason, the application of K-means was extended, performing the corresponding calculations with up to seven clusters.

Depending on the size of the upper cluster for each iteration of K-means++, the level of agreement with N_{Wikis}, N_{Props}, and N_{Words} was evaluated. The variable with the highest matching ratio is N_{Words}. However, the set of literary works that should form part of the canon was different depending on the variable used. For this reason, the dimensionality was reduced using two methods. The first of them was the PCA method (Dig & He, 2004) calculated from N_{Props} and N_{Words} since they are the variables with the least correlation. An indicator was also defined and calculated, which has been called Wiki3DRank, as the aggregation of the logarithmic transformation of each of these variables (Shatnawi, 2015). For each item Wiki3DRank would be calculated as:

$$Wiki3DRank = \log(1+NWikis) + \log(1+NProps) + \log(1+NWords)$$

This equation, the calculation of which is very simple, integrates N_{Wikis}, N_{Props}, and N_{Words} into a single metric with a relatively normalized distribution (Table 3).

Variable	Mean	Median	C_v	Min	Max	Skewness	Kurtosis
Wiki3DRank	7.556	7.745	0.365	1.386	21.874	-0.014	0.610

Table 3: Statistical analysis of Wiki3DRank. Source: own preparation.

In each iteration n of K-means++ it is possible to calculate the matching between the set of items of the cluster C_n (upper cluster) and the subset delimited between the interval [1,Sn] of each of the rankings established by integrates N_{Wikis}, N_{Props}, and N_{Words}, PCA and Wiki3DRank. Based on the number of matching elements and the size of the upper cluster (S_n), a matching ratio was calculated (see Table 4). Wiki3DRank reaches the highest match ratios in any iteration and the highest value happens in the iteration with three clusters. It can also be seen that of the three variables that define an item, N_{Words} is more representative than N_{Wikis} or N_{Props} with respect to the matching with the results of K-means+++.

n	S_n	Silhouette	Matching ratio				
			N-Wikis	*N-Props*	*N-Words*	PCA	*Wiki3DRank*
2	1,008	0.909	0.869 (876)	0.499 (503)	0.802 (808)	0.882 (889)	0.927 (934)
3	163	0.827	0.822 (134)	0.595 (97)	0.822 (134)	0.822 (134)	0.939 (153)
4	152	0.493	0.822 (125)	0.559 (85)	0.849 (129)	0.822 (125)	0.934 (142)
5	74	0.499	0.676 (50)	0.608 (45)	0.824 (61)	0.676 (50)	0.919 (68)
6	65	0.493	0.615 (40)	0.600 (39)	0.846 (55)	0.615 (40)	0.908 (59)
7	36	0.457	0.472 (17)	0.556 (20)	0.750 (27)	0.472 (17)	0.833 (30)

Table 4: Matching ratios for the different iterations of K-means (in parentheses the number of matching items). Source: own preparation.

Taking into account these results, it was decided to use K-means++ to obtain three clusters. The size of C_1 is 105,100 items, Cluster C_2 (labelled as secondary cluster) contains 2,171 items and Cluster C_3 (main cluster) includes 163 works. This main cluster contains the items of those literary works that are candidates to be considered part of the Universal Literary Canon. Cluster C_1 could be called "bibliographical production", a vast set of books and works achieving greater or lesser success, a more local impact, and little attention. Secondary Cluster C_2 includes a (relatively attainable) set of works that represent to a certain extent the middle class of literature: works that have become well-known in a handful of languages and with varying levels of encyclopedic attention. Figure 1 clearly shows the three clusters and it shows how the works of the cluster C3 have higher values in the analyzed variables (N_{Wikis}, N_{Props}, and N_{Words}).

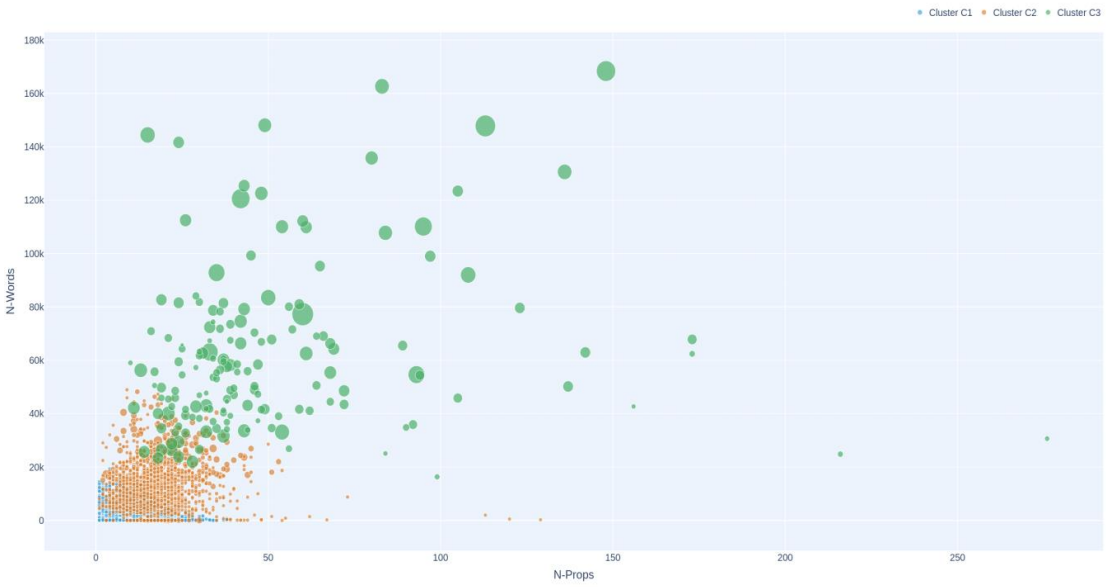

Figure 1: N_{Props}/N_{Words} distribution of the clusters (the size of the bubble represents N_{Wiki}). Source: own preparation.

By way of example, Table 5 shows the data for one literary work from each cluster.

Item	Title	Cluster	*Wiki3DRank*	*N-Wikis*	*N-Props*	*N-Words*
Q8275	Illiad	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

Table 5: Example of literary work from each cluster. Source: own preparation.

In addition to clustering, the Wiki3DRank metric provides a value for selection and ordering operations. This allows the items in the dataset to be sorted independently of the cluster to which they belong and, more useful, within it. Figure 2 represents a selection of the first fifty works from C_3. In the final annex, a complete list of the works in this cluster of "universal classics of all time" can be consulted.

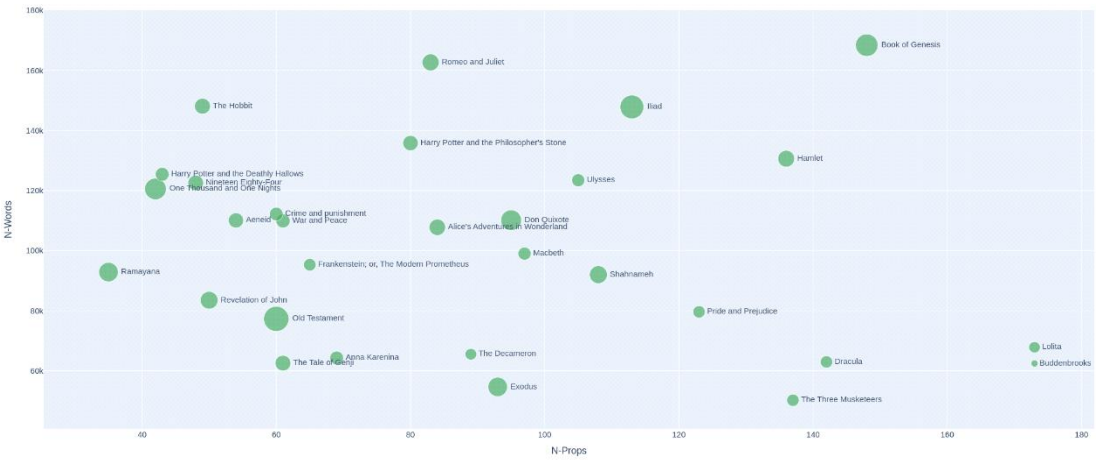

Figure 2: Selection of the 30 works in cluster C_3 with the highest Wiki3DRank. Source: own preparation.

In the top positions of C_3, we find easily recognizable works such as: *Genesis, The Iliad, Hamlet, Romeo and Juliet, Don Quixote, Shahnameh, Ulysses, Harry Potter and the Philosopher's Stone, Alice in Wonderland, Lolita, Macbeth, Pride and Prejudice*, etc. By applying these same metrics, the literary canon written in any language can be extracted. For example, Table 6 includes the top ten works of all time written in Italian.

Tabla 6. Top-ranked literary Works written in italian.

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

4.2. Classical and Contemporary Literature: Local Literary Canons and the Weight of Tradition

The available data allows addressing different facets of studies on cultural phenomena, such as the endurance of certain works over time, the even or uneven representation of different languages, or the permeability of the global canon and canons in each language across different linguistic domains. In this work, we present only the basic data on the language distribution of the globally considered canon, and we analyze with a bit more detail its temporal distribution based on the date of production or publication.

The language of literary works was studied from two perspectives: the language of the works themselves and the Wikipedia editions in which they existed. The language of each work was determined using two mechanisms: explicit description using the P407 property or (in some cases) extracting the language from the original title. For each language, the number of works was counted and the mean values of N_{Props} and N_{Words} were calculated. Figure 3 shows the dispersion of each language based on the means of N_{Props} and N_{Words}. The size of the elements is defined based on the total number of works in each language according to such means. This initial analysis indicates that most of the works are written in English and have a high degree of description. The element labeled as “<none>” represents those works in whose Wikidata items there are no statements with the P407 property and it has not been possible to extract the language of the original title. These literary works are numerous (39,465) but their Wikidata items have a low level of both description and editorial content of the corresponding articles in the different Wikipedia editions. Among the other languages, Spanish, French, Japanese, Russian, and German stand out. It is worth highlighting the case of Latin, Sanskrit, and classical Greek with a low volume of works but with numerous descriptive statements and extensive articles in the different Wikipedias.

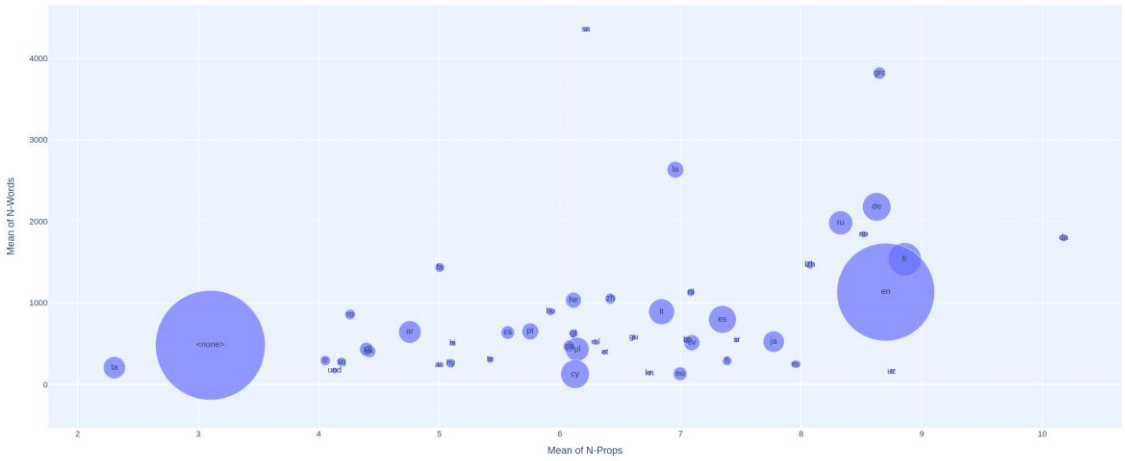

Figure 3: Distribution of languages with a minimum of 100 works based on the means of N_{Props} (x-axis) and N_{Words} (y-axis). The size of the points represents the number of literary works of the language. Source: own preparation.

In relation to the date of the selected literary works, the data obtained allow for outlining an overview of the period to which the works that are part of the universal or local literary canon belong. From these data it is possible to analyze the temporal distribution of the works. It must be pointed out that for a large number of items of literary works there is no information about the date of publication or inception. Only 61,702 items (just over 57%) include a property from which to obtain this information. Most of the data was obtained from property P577 (publication date) and only 2.2% through property P571 (inception date). The results, grouped by centuries, can be seen in Table 7.

Century	Items	C_1	C_2	C_3	Total Wiki3DRank	% items	% with date	% Wiki3DRank	Wiki3DRank Ratio
21	22,319	22,095	217	7	171,355.06	24.87	40.32	35.01	7.68
20	31,836	30,990	802	44	256,594.02	35.47	57.51	52.42	8.06
19	5,174	4,675	454	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

Table 7: Temporal distribution by centuries of the number of items and Wiki3DRank. Source: own preparation.

More than 87% of the items that have some type of date (50% of the total items in the dataset) have a publication or inception date corresponding to the 20th and 21st centuries. The data can be grouped or analyzed in more detail. Figure 4 shows a Wiki3DRank distribution by century (top) and of all works published or created in the 20th century by year (bottom). It also shows the cluster to which each work belongs. As it seems reasonable, only a few works from the distant past are included if they lack a certain relevance (main and secondary clusters). In general, we can see that the data has a varied temporal distribution, with an emphasis on the mid-years of the 20th century."

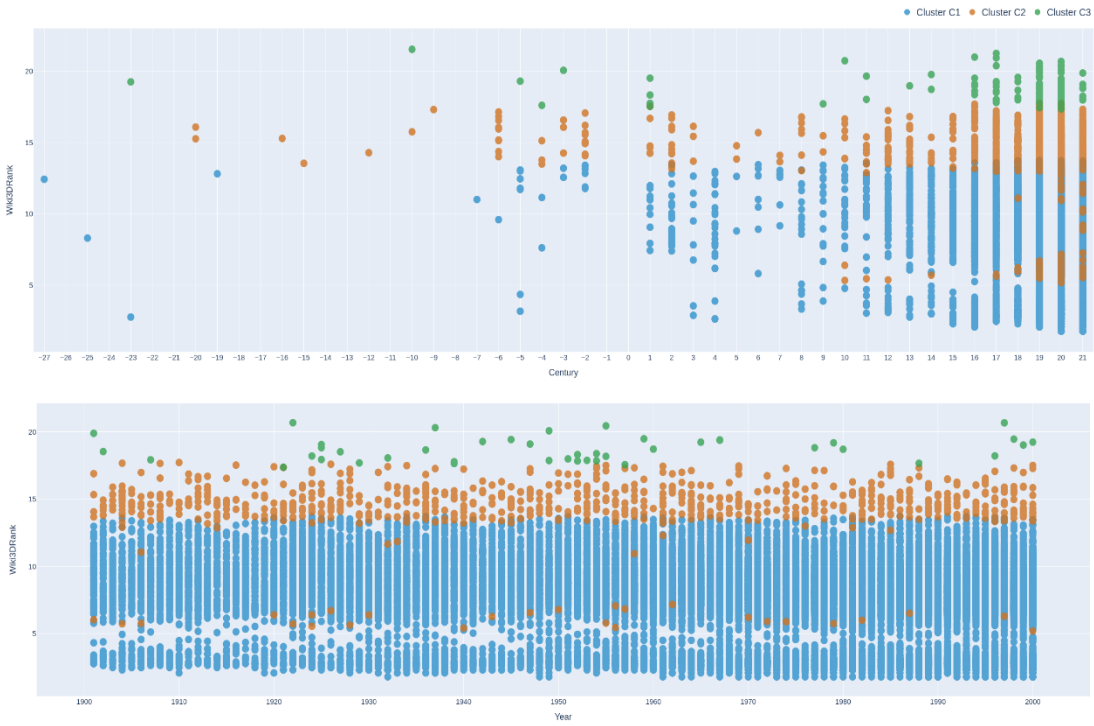

Figure 4: Distribution by centuries (top) and by years of the 20th century (bottom) and Wiki3DRank of the items according to date of publication or inception. Source: own preparation.

Filtering by language is another interesting possibility offered by the study of the temporal data. The evolution and prominence in literary production for each language can be observed and even compared with other languages (Figure 5).

Figure 5: Temporal distribution of works in Spanish (top) and English (bottom). Source: own preparation.

5. DISCUSSION

This research presents several elements that could be valuable: it establishes a quantitative threshold for the number of works that we could identify as exceptionally relevant globally, a way to weigh them individually, and a list of works from the universal literary canon. In this list, titles commonly referred to as 'classics of all time' can be found. These are narratives easily identifiable as part of tradition and can be situated in certain moments and places in history.

The quantification of N_{Wikis}, N_{Props}, and N_{Words} offers three indirect measures of something that we can term as “encyclopedic effort” or perhaps “encyclopedic attention”. First, there is the length of Wikipedia articles in any language, using a cumulative measure and not a weighted measure of centrality (N_{Words}). Secondly, there is the depth of description enrichment in Wikidata, which reflects another type of attention that focuses on data and factual details about a work (N_{Props}). Finally, there is the dissemination of the work through different languages, which acts as an indicator of global presence or reception (N_{Wikis}). Integrating these three variables enables a richer representation than using each of them separately. The position of the works with respect to the axes of the scatter diagrams reflects the balance between text and data (Wikipedia/Wikidata) and neatly shows irregularities and asymmetries. Few works in the secondary cluster (C_2) reach magnitudes that are comparable, in any of the three parameters, to the works in the main cluster (C_3).

If we compare the works of the main cluster C_3 with the results of the WikiRank website for the category “Books” we found a coincidence of 74.3%. However we appreciate a better ordering of works applying Wiki3DRank and a surprising amount of recent and mainstream works in the results shown in WikiRank. The three clusters obtained bear a certain resemblance to other proposals based on other premises from the academic or literary field. The main cluster C_3, containing 163 works, is of a similar size to that proposed by Christiane Zschirnt in her study “Libros, todo lo que hay que leer” [“Books, everything you need to read”] (Zschirnt, 2011). This author picked 141 works, and her selection and those obtained in this study[8] coincide substantially. Addiotionally, a broader study such as “1001 Libros que hay que leer antes de morir” [“1001 books you must read before you die”] (Boxall, 2005), a number chosen due to its catchiness, is loosely closer to the set of C_3 and C_2 (2,334 works), although it almost perfectly aligns with the alternative of calculating only two clusters, which would result in 1008 works.

One noteworthy aspect is the presence of works from religious traditions, especially the Judeo-Christian tradition. Generally, these works are not considered literary works in studies in the academic field of literature studies (“Genesis”, “Leviticus”, “Epistle to the Philippians”, etc). These mythological-spiritual works constitute the basis of religious communities and would merit being differentiated in order to obtain a picture more consistent with what is now, strictly speaking, considered literature. This same problem arises when we find the essay genre, philosophical works and popular science (“The Wealth of Nations” or “The Republic”). The publishing industry and booksellers tends to differentiate between fiction and non-fiction, placing literature in the first group. Strikingly, books such as "Mein Kampf" and "The Guinness Book of Records" appear in the main cluster. This indicates that it would be necessary to study how to better identify the area of interest and prevent results that fall outside the scope of what is commonly understood to be a literary canon.

A thorough review of C_3 makes it possible to detect the absence of certain well-known books (The Divine Comedy, Odissey). One of the reasons is the existence of inconsistencies in the assignment of the appropriate classes to the items. In the Wikidata knowledge base, the class "book" is frequently used for current works, or "written work" for many other cases. It is also due to the level of specificity in typification, assigning, in many cases, subclasses with a deeper level of detail under the taxonomy “literary work”. It would be necessary to investigate further into the precise selection of classes and subclasses within the Wikidata concept schema. In this way, account would be taken of the tendency towards disorganization and inconsistency when involving more elements of the taxonomy, and it thus would require fine-grained validation and denoising processes. A similar situation can be observed for works that are presented in the form of sagas or series, for which the results are distributed between the individual works and the complete series, depending on how they have

been described. This aggregate format hinders precise identification. “The Lord of the Rings” does not appear as such, since it is linked to the “literary trilogy” (Q13593966) and “novel” (Q1667921) classes, although some of its volumes do. Nor is it clear whether in the case of “Don Quixote” both parts of the work are together under the title included in our list. The well-studied work-aggregation duality indicates the suitability of establishing procedures to rank those works that appear individually and grouped, such as “Book of Genesis” and “The Bible” or each of the releases of sagas and series of novels such as Sherlock Holmes. In the same way, the canon seems to harm poetic works, tales and short stories, surely due to the conditions of its edition and publication in numerous and varied compilations or anthologies.

The very consistency of the data recorded in Wikidata makes it difficult to systematically explore other aspects such as authors, genres, topics, etc. The variability in the use of properties can be observed, but the use made of them is also very heterogeneous, as there are no agreed content standards for description. Even so, if the works' score on the Wiki3DRank metric is higher, or if they belong to the main or secondary cluster, the quality of their description is greater.

In relation to recent literary works, it was detected that these included numerous best-sellers, which are not usually considered by classic literary criticism to be literary works with perdurable value. However, this is the case for new trends of study on "canonical best-seller" (Muñoz Rico; García Rodríguez; Cordón García, 2020). Examples of this are the “Harry Potter” or “The Hunger Games” sagas. This suggests there is a certain difficulty in correctly capturing, through the quantitative methodology used in this research, the relevance of works emerging in a context of global campaigns and mainstream releases within mass culture system. However, works from the 19th and 20th centuries seem to fit the model used in the research.

The temporal distribution of the works shows books from all periods. However, works from the 20th and 21st centuries predominate, in keeping with the emergence of a mass market for books and the rise of the mass media. The data for C_2 and C_3 could represent the generic idea of “current and all-time classics”. The greater their temporal distance from the present day, the more usual it is for the works collected to be only those of particular relevance and whose interest has been proven over time.

The language of the C_3 works reflects wide linguistic variety and roughly represents the idea of a global canon. The cluster balances the "Eurocentric" trend of literary canons proposed by other authors. However, it continues to reflect the disparity in the spread of languages associated with colonial empires and economic powers. Despite this, it permits a greater opportunity to highlight books written in dead and non-Western languages. It should be noted that, outside of the dominant Western languages, the

presence of works in other languages is related to remote antiquity, well before the printing age.

6. CONCLUSIONES AND FURTHER RESEARCH

The results obtained show that the combined use of Wikidata and Wikipedia is another source of information to delineate a literary canon, and therefore the hypothesis stated at the beginning of this research would be verified. The observation and measurement of the attention paid by the Wikimedia community to literary works allows us to learn more about the relevance and visibility of the works, and it also serves as a complement to the canon proposals made by the media, academia, and the publishing industry. Thus, here is another valuable source for debating about an open and multiple canon, which contains the result of numerous autonomous voices and individual agents.

Unlike surveys of literary taste, Wikipedia reflects individual decisions to build a canon by putting effort into enriching articles and descriptions. The study presented bears, in a certain way, parallels with studies on translation, print runs, reissues, and sales in publishing and book retail communities, all of them fields for which there are no easily accessible and actionable longitudinal data sources. The data obtained conform, to a certain extent, to the LRM conceptual model. This model differentiates the Work from its Expressions and Manifestations, so it is possible to obtain a collaborative inventory of literature in each language and, by aggregating all Wikipedia editions, a global one.

The visibility of literary works on Wikipedia is still consistent with the educational or academic canon as it is represented in most world literature handbooks. However, a gradual shift towards a greater attention to current works is perceived (presentism), related to the massively successful and transmedia works of the 20th and 21st centuries. Analysis of the editions in each language, combined with the literature produced in each language, makes it possible to quickly sketch out geographical spaces reflecting cultural proximity and influence. This effect could be mitigated by introducing a new variable that takes into account the publication or creation date of the works and increases the Wiki3DRank scoring of older works, or some other domain attributes that may be related to aspects of quality and impact.

In addition, we are aware that the selection of items and articles analyzed (those directly classified as a "literary work" in Wikidata) covers only a portion of the real universe of this type of works. For this reason, in further research, it is necessary to design knowledge-base exploration mechanisms that analyze other classes used to classify literary works. It is essential to take into account, at the same time, the validation of the organizational chaos of the taxonomy of classes caused by the very nature of collaborative description. For this reason, it is essential to understand that

the methodology used is based exclusively on existing data from literary works explicitly identified as such in Wikidata.

The above also implies that works present in other literary canons created subjectively according to author's criteria, but not present in Wikidata, would be excluded. Additionally, as the results show, the coverage of a work in different Wikipedia editions is crucial to establish its position in the ranking calculated by Wiki3DRank. This means that works with limited or language-specific diffusion would be less representative in a universal canon. However, the proposed method would still be valid for establishing a literary canon for a specific language.

The use of metrics related to editorial depth and activity of editors on items in Wikidata/Wikipedia about literary works also seems necessary. In addition, they must be applied in a way that makes it possible to capture in a more detailed fashion the attention and effort put into each article and item, as an indirect measure of its value.

It is necessary to mention that the aggregation of the logarithmic transformation of N_{Wikis}, N_{Props}, and N_{Words} to calculate Wiki3DRank yields coherent results regarding the proposed hypothesis. As a future line of work, an alternative calculation is considered in which works are represented as vectors. The components of these vectors would correspond to the logarithmic transformations of these variables. Through this method, Wiki3DRank could be obtained by calculating the module of the corresponding vector for each work.

This paper also opens up the gates to the use of Wikipedia for the extraction of a proposal of a transmedia cultural canon, which would include the other great formats of narrative fiction, such as movies, comics, video games, and television series. All this would allow us to look deeper into how narrations are intertwined, since they are consumed and published in iterative cycles of versions, adaptations, updates, reboots, and recreations, which is in fact not a completely new phenomenon either, although its rhythm and impact may be.

7. NOTES

[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] It is important to note that an initial version of the dataset for this work, obtained on November 20, 2021, only included 89,744 items.

[8] Only 94 of the works proposed by the author are explicitly categorized as 'Literary Work' in Wikidata, and they are the only ones that could appear in our study. Of these, C3 includes 92 works (97%). Considering all the author's works, the degree of agreement would be 65%.

8. REFERENCES

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.

APPENDIX: LIST OF LITERARY WORKS IN THE CANON

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
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
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
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