

More than 1001 Books you must read before you die: Measuring change, diversity and centrality in an evolving contemporary World Literary Canon according to Wikipedia and Wikidata metrics

1

Más de 1001 libros que debes leer antes de morir:
medición del cambio, la diversidad y la centralidad en
un canon literario mundial contemporáneo en evolución
según métricas de Wikipedia y Wikidata

Més de 1001 llibres que has de llegir abans de morir:
mesura del canvi, la diversitat i la centralitat en un
cànon literari mundial contemporani en evolució segons
mètriques de Wikipedia i Wikidata

Tomás Saorín-Pérez  i

Universidad de Murcia

tsp@um.es

© Authors

Juan-Antonio Pastor-Sánchez  i

Universidad de Murcia

pastor@um.es

Received: 02-10-2025

Accepted: 10-03-2026

Recommended citation

Saorín-Pérez, Tomás; Pastor-Sánchez, Juan-Antonio (2026). More than 1001 books you must read before you die: Measuring change, diversity and centrality in an evolving contemporary World Literary Canon according to Wikipedia and Wikidata metrics. *BiD*, 56. <https://doi.org/10.1344/bid2026.56.01>

Abstract

The culture of ranking and selection has a long tradition, increasingly shaped by quantification and digital interactions in contemporary contexts. This phenomenon manifests in various forms, such as bestseller lists and curated compilations of creative works across all time periods, serving as frameworks for cultural promotion and access. Within publishing, audiovisual content, and entertainment, inclusion in such rankings holds significant value. This study proposes methodologies for analyzing and comparing lists and rankings within specific domains, building upon the Wiki3DRank methodology developed for encyclopedic knowledge objects. When applied to literature, this approach offers insights into books as cultural artifacts that capture varying levels of attention. Leveraging data from Wikipedia and Wikidata, the methodology synthesizes both informational and structured data about literary works. This approach is particularly relevant to exploring literary canons, a longstanding construct within institutional, educational, and cultural contexts. The study focuses on a subset of the canon linked to the creation of lists. While many lists lack in-depth literary critique, they provide schematic overviews of notable works, often organized by time periods, genres, or linguistic and national boundaries. All international editions of Peter Boxall's work *1001 Books You Must Read Before You Die*, published during the first quarter of the 21st century, are analyzed and measured, proposing metrics for comparison across editions and with other lists, including insights into the authors' genres and the languages of the works.

Keywords

Literary Canon; book lists; data-driven analysis; Wiki3DRank; Wikidata; Wikipedia.

Resumen

La cultura del ranking y la selección tiene una larga tradición, cada vez más marcada por la cuantificación y las interacciones digitales en los contextos contemporáneos. Este fenómeno se manifiesta de diversas formas, como las listas de bestsellers y las compilaciones curadas de obras creativas de todos los periodos históricos, que funcionan como marcos para la promoción cultural y el acceso a los contenidos. En el ámbito editorial, audiovisual y del entretenimiento, la inclusión en este tipo de rankings tiene un valor significativo. Este estudio propone metodologías para analizar y comparar listas y rankings dentro de dominios específicos, basándose en la metodología Wiki3DRank, desarrollada para objetos de conocimiento enciclopédico. Aplicado a la literatura, este enfoque ofrece perspectivas sobre los libros como artefactos culturales que captan distintos niveles de atención. A partir de datos procedentes de Wikipedia y Wikidata, la metodología sintetiza tanto información descriptiva como datos estructurados sobre obras literarias. Este enfoque resulta especialmente pertinente para explorar los cánones literarios, un constructo de larga tradición en contextos institucionales, educativos y culturales. El estudio se centra en un subconjunto del canon vinculado a la creación de listas. Aunque muchas listas carecen de una crítica literaria profunda, proporcionan visiones esquemáticas de obras destacadas, a menudo organizadas por periodos históricos, géneros o fronteras lingüísticas y nacionales. Se analizan y miden todas las ediciones internacionales de la obra de Peter Boxall, *1001 Books You Must Read Before You Die*, publicadas durante el primer cuarto del siglo XXI, proponiendo métricas para la comparación entre ediciones y con otras listas, así como aportando información sobre los géneros de los autores y las lenguas de las obras.

Palabras clave

Canon literario; listas de libros; análisis basado en datos; Wiki3Drank; Wikidata; Wikipedia.

Resum

La cultura del rànquing i la selecció té una llarga tradició, cada vegada més marcada per la quantificació i les interaccions digitals en els contextos contemporanis. Aquest fenomen es manifesta de diverses formes, com ara les llistes de bestsellers i les compilacions curades d'obres creatives de tots els períodes històrics, que funcionen com a marcs per a la promoció cultural i l'accés als continguts.

En l'àmbit editorial, audiovisual i de l'entreteniment, la inclusió en aquest tipus de rànquings té un valor significatiu. Aquest estudi proposa metodologies per analitzar i comparar llistes i rànquings dins de dominis específics, basant-se en la metodologia Wiki3DRank, desenvolupada per a objectes de coneixement enciclopèdic. Aplicat a la literatura, aquest enfocament ofereix perspectives sobre els llibres com a artefactes culturals que capten diferents nivells d'atenció. A partir de dades procedents de Wikipedia i Wikidata, la metodologia sintetitza tant informació descriptiva com dades estructurades sobre obres literàries. Aquest enfocament és especialment rellevant per explorar els cànons literaris, un constructe amb una llarga tradició en contextos institucionals, educatius i culturals. L'estudi se centra en un subconjunt del cànon vinculat a la creació de llistes. Tot i que moltes llistes no ofereixen una crítica literària aprofundida, proporcionen visions esquemàtiques d'obres destacades, sovint organitzades per períodes històrics, gèneres o fronteres lingüístiques i nacionals. S'analitzen i es mesuren totes les edicions internacionals de l'obra de Peter Boxall, *1001 Books You Must Read Before You Die*, publicades durant el primer quart del segle XXI, i es proposen mètriques per comparar-les entre si i amb altres llistes, així com per aportar informació sobre els gèneres dels autors i les llengües de les obres.

Paraules clau

Cànon literari; llistes de llibres; anàlisi basada en dades; Wiki3DRank; Wikidata; Wikipedia.

1. Introduction: From Literary Canon to "Best Books" Lists

The culture of ranking and selection has a long-standing tradition, as reflected in well-known examples such as the Roman *De Viris Illustribus* collections and Johann Joachim Winckelmann's 18th-century efforts to classify and rank classical artworks—among others.

In the literary field, the notion of a defined set of works that constitute the literary canon is a well-established trend across institutional, educational, academic, and mass communication domains (Sullà, 1998). National literatures construct pantheons and commemorations of their consecrated authors and works. Reference works on literary history identify the pinnacle achievements of literature, elaborating on movements, production contexts, influences, impact, and a wealth of specialized knowledge. These works may serve as a contemporary proxy for a global canon shaped by academia, institutions, culturally sanctioned taste, educational tradition, and "market dynamics. These forces are central to Bourdieu's analysis of the modern processes of canon formation (Bourdieu, 1988) and also in a wide range of studies on literary value, cultural capital, attention economies, critical theory, postcolonial literary studies, and the interplay between popular and academic culture (Pozuelo Yvancos; Rodríguez Alonso, 2021; Guillory, 1993; Kermode, 1999; Kolbas, 2001). While such frameworks are useful for understanding the broader mechanisms underlying cultural shortlists, a detailed engagement with these theoretical perspectives lies beyond the scope of the present study.

Within the field of literature and book culture, the most developed form of cultural selection is found in systematic works of literary history and criticism. In this study we focus specifically on compilations of world literature. Histories of World Literature

could serve as a valuable source for a study about list of most notable books. In the Spanish context, a paradigmatic example is the two-volume *Historia de la literatura universal* by Martín de Riquer and José M^a Valverde, a work that has been reprinted for over four decades. The authors describe it as offering “a clear critical overview of the works that constitute the universal literary heritage of civilized nations,” defining “universal” as any literary creation capable of engaging a contemporary reader beyond national, linguistic, or ideological boundaries (De Riquer; Valverde, 2005). These and other types of academic works are discursive and complex, which makes them less suitable for treatment as structured data. In contrast, formats such as lists or enumerations of creative works are more amenable to analysis through metrics and the methodological approaches associated with the so-called “network turn” (Ahnert et al., 2021), within the broader frameworks of distant reading and digital humanities.

Rating, reviewing, and commenting are social phenomena in themselves that are part of public discourse and are manifested through both classic and current media (Blank, 2007). Ratings and scores coexist with critiques, reviews, and cultural analysis, all functioning as journalistic genres that shape cultural promotion and access. As Kristensen and From (2017) note, these practices blur the boundaries between high and popular culture while contributing to the reconfiguration of cultural hierarchies. In its contemporary form, this phenomenon reveals a progressive process of quantification, particularly influenced by digital social interactions. Within publishing, audiovisual media, and entertainment, the importance of appearing in lists, rankings, awards, and curated selections is particularly pronounced, as these formats signal authority, visibility, and symbolic value (Lawson, 2020). They influence consumer behavior, institutional recognition, and cultural memory, as seen in the dynamics of bestsellers, recommendation algorithms, and prize shortlists, which together shape what is often referred to as the “non-academic canon” (Lluch, 2025).

This study focuses on a specific subset of the literary canon that aligns with the impulse to create lists—a phenomenon with anthropological interest, referring to it as the “fascination of the list” (Von Contzen, 2021)— and which has even been described as “reading listopia” by The Library Journal (Verma, 2022). There are numerous book lists that function merely as lists, unaccompanied by extensive works of literary criticism or history. In such cases, their reviews, when included, tend to be schematic and superficial. In the case of literature, and without aiming for exhaustiveness, numerous examples of these “best books lists” can be identified. For many years, such lists have often been associated with media outlets, general-interest magazines with a focus on readers, and occasionally take the form of books or reading guides. These lists frequently exhibit a tendency to delineate significant temporal periods (year, decade, century), genres (science fiction, noir fiction), or languages and nations. Additionally, there is a noticeable preference for rounded numbers, with their range typically spanning between a concise 100-200 entries. Lists with broader coverage are far less common, where the figure of 1,000 entries often serves as an upper limit. The following is a small collection of both broad and concise lists produced since the beginning of the 21st century:

Table 1. Sample List of Selections of Literary Works of Varying Scope and Size (Source: Author’s Own Work)

Title	Dates	Publisher	Characteristics
1000 novels everyone must read: the definitive list	2009	The Guardian	Created by a panel of experts, these lists are published periodically in supplements organized by genre and eventually compiled into a single list. They are primarily focused on fiction.
1000 Books to read before you die: A life-changing list	2018	Workman Publishing Company	A book published in the USA, curated by James Mustich.

Title	Dates	Publisher	Characteristics
Are You Well-Read in World Literature?	2009-2011	GoodRead	Resultado del foro en GoodReads "Around the world in 80 books", orientada a literatura mundial
1001 books you must read before you die	2006-2021	Cassell y and others	The result of the Goodreads forum "Around the World in 80 Books," focused on world literature.
Bücher Alles, was man lesen muss / Libros: todo lo que hay que saber	2002-2006	RM Buch und Medien (Germany) Taurus (Spain)	Curated by Christiane Zschirnt. Featuring 141 works, with adaptations in Spanish and Portuguese.
The 100 greatest novels of all time: The list	2003	The Guardian	100 novels, curated by Robert McCrum.
100 besten Bücher der Weltliteratur	2004	ZDF (German TV)	100 works. Compiled through a collective consultation of experts and stakeholders, followed by a broad public vote.
Verdensbiblioteke / Bokklubben World Library	2002	Norwegian Book Club	A collection of 100 books of all time, proposed by an international group of writers.
The Big Read	2003	BBC	200 works, primarily of modern English literature, selected through a year-long process of mass voting and participatory engagement.
Los 100 mejores libros de la literatura universal	2018	ABC (Spanish Newspaper)	100 works selected by writers, critics, and cultural figures.
This is the canon: decolonize your bookshelf in 50 books	2021	Hachette	50 works selected by postcolonial literary critics Joan Anim-Addo, Deirdre Osborne, and Kadija Sesay.
The Novel 100: A Ranking of the Greatest Novels of All Time	2003-2010	Facts on file	200 novels, curated by Daniel S. Burt

The aforementioned works share common characteristics —fiction, world literature—and are, to some extent, heirs to a long tradition of publishing collections of national and international classics. Examples of this include *La Pléiade* in France, *Clásicos del Siglo XX* by *El País* in Spain, or Mario Vargas Llosa's *Biblioteca de Plata* for *Círculo de Lectores* also in Spain.

This study was initially driven by the question: How could we compare all these book lists using a quantitative approach? Building on this, we extend a method previously tested in earlier research to compare and measure lists or rankings within a specific domain—in this case, literature as a cultural or media field. Although the methodology will be explained in detail later, it is worth noting that it is based on the processing of open data available on Wikipedia and Wikidata in all accessible languages. Several studies have also explored the use of Wikipedia editions as a source for examining the reception and impact of literary works—a recent example can be found in the special issue *Wikipedia, Wikidata, and World Literature* (Fischer et al., 2023) in the *Journal of Cultural Analytics*. In this study, we use highly specific data derived from the globalized Wikimedia projects. Therefore, it aligns with what Borgman (2017) refers to as "little data" and should not be understood as "big data."

Since our main interest lies in how such lists can be compared, we focus on a single use case that has been issued multiple times over a significant span of years. Comparing lists can be challenging, especially when attempting to comment on the presences and absences in each selection. Therefore, our approach seeks to "weigh" the lists based on the rank and relative weight of their components, allowing us to estimate their overall value. From there, we delve into questions of diversity and hierarchy, adopting a data-driven perspective.

To this end, from among the selections mentioned earlier, the analysis focuses on the proposal led by Peter Boxall, *1001 Books You Must Read Before You Die*, which exhibits a handful of highly intriguing and unique characteristics.

2. Peter Boxall's Universal Literary Canon

The book *1001 Books You Must Read Before You Die*, edited by British scholar Peter Boxall from the University of Sussex, presents a unique set of characteristics that exemplify a particular conventional idea of the literary taste or international literary appreciation. Although we occasionally refer to this work as the "Boxall Canon," the author himself explicitly distances himself from such a label, stating that his list "does not attempt to be a new canon and does not aim to define or exhaust the novel." Nevertheless, given its scope, structure, and editorial authority, it invites canon-like readings, particularly when considered in relation to other curated literary lists and their cultural impact.

This book offers a concrete example of what "world literature" can signify in the context of global literary lists curated for mass readership. It reflects a canon of works that aspire to universal cultural value, while being filtered through the tastes and market dynamics of the Anglophone publishing industry, with its commercial priorities and cultural gatekeeping mechanisms. It exhibits a clearly global ambition, resulting in a list that blends critical recognition with popular accessibility. This duality has led to its being labeled a "literary bucket list" (Dirda, 2018) and it is also seen as a "very loose canon" (Hensher, 2006). While *world literature* remains a contested concept within literary studies (Casanova, 2004), the Boxall project exemplifies a classical, Western-oriented approach: a canon that claims representativeness while remaining rooted in Euro-American editorial frameworks that presume the authority to interpret the world at large.

The book *1001 Books You Must Read Before You Die* is not a standalone publication; its original English edition has undergone five revised editions in the same language, published by various British, American, and Australian publishing houses in multiple reprints (Cassell Illustrated, Universe, ABC Books, Pier 9, Octopus, Quintet, Penguin, Hachette, Quintessence, Barrons, Murdoch Books). Over approximately 15 years (2006–2021), the selection of titles has evolved, making the book a reflection of the state of the universal literary canon for all audiences at the dawn of the 21st century.

In addition to the English revisions previously mentioned, the work has been published in adapted versions for the languages of major publishing markets, such as Spanish, French, German, Portuguese, and others. Beyond translating the original work, these adaptations are tailored, to some extent, to each cultural context, reflecting local literature and authors.

Peter Boxall serves as the editorial director of the project, but the work is supported by a substantial team of advisors who contribute to the selection process and the preparation of the reviews. It typically also includes a main editor or coordinator for each language version, such as Jennifer Byrne (AU), Peter Ackroyd (DE), José-Carlos Mainer (ES), Göran Hägg (SV), Jean d'Ormesson (FR), Ed van Eeden (NL), Olaf Jensen (NO), or Sadanori Betsumiya (JP). This team changes across the different English editions and typically includes a lead editor for each language version. Thus, the book is neither the product of individual authorship nor the result of an open, mass collaboration. Instead, it represents a coordinated effort by a group of recognized academics or literary critics.

Although the figure of 1001 books is merely a nod to the Middle Eastern collection of tales *One Thousand and One Nights*, it is worth emphasizing that it represents a relatively extensive selection of titles, allowing for a broad range of tastes and sensibilities. Our preliminary explorations of 'best books' lists suggest that compilations featuring 100–200 items are much more common, typically focusing on a specific and more recent period. In this context, the *1001 Books* series stands out not only for its scope, but also for its enduring cultural impact. The 1001 concept has been adapted to films, monuments,

comics, music albums, and other creative domains, establishing a cultural blueprint for similar lists. The book itself, as an encyclopedic object, is noteworthy enough to have its own Wikipedia entry in several languages.

The work is better understood as a project of cultural dissemination rather than as a contribution to literary history or criticism, even though its contributors often hail from these domains. While literary history and criticism aim to produce analytical, interpretive, or theoretical frameworks about texts, this project is oriented toward the selection, circulation, and promotion of literary works for a general audience. It has not been the subject of specific studies in the MLA International Bibliography, the primary reference database for literary studies. A recent study, applying other metrics and purposes, has analyzed the selected works in relation to Wikipedia, considered as a ‘hall-raum’ (resonance space) of literary canonicity (Rohe et al., 2025). Nevertheless, it could be seen—at least to some extent—as a representation of mainstream world literature. It reflects certain aspects of the international reception and dissemination of literary works, an area where clear and comprehensive data is often lacking. The book is frequently mentioned on reading-focused social networks like *Goodreads* or *LibraryThing*, as well as in the abundant subgenre of lists, such as *Listology*. It is widely recognized among book enthusiasts and frequently appears in online discussions. Readers often use it to track how many titles they have read, update their progress, or diversify their reading with works from different geographic regions while delving into the “classics”, as seen in these suggested links.

Example 1. References to the 1001 Books You Must Read Before You Die list in social reading networks.

Boxall's 1001 Books You Must Read Before You Die
https://www.goodreads.com/group/show/970-boxall-s-1001-books-you-must-read-before-you-die
Library Thing list 1001 Books You Must Read Before You Die
https://www.librarything.com/list/508/all/1001-Books-You-Must-Read-Before-You-Die
Listopia > 1001 Book Lists
https://www.goodreads.com/list/tag?id=1001&ref=ls_ts
Reading Challenges for 2025 > The Greatest Reading Challenge of Your Life
https://www.goodreads.com/challenges/20983-el-gran-reto-de-lectura-de-tu-vida

The longevity of Boxall's work and its evolution across numerous international editions allow us to observe how certain selection criteria are manifested within a consistent editorial project that is clearly positioned within the present cultural moment. Across successive editions and translations, the structure of the collection has been maintained, while incorporating local variations and periodic revisions. In total, all the books included in the English editions amount to 1,316 works, of which 705 appear in all English editions (referred to at times as the “BYMRBYD Consensus” and the “core list,” respectively). When compiling the works included in all international editions, the number rises to 1,718 titles. However, the most distinctive feature of this work is its international reach, with editions of the same characteristics available in the major languages and publishing markets: French, German, Spanish, Italian, Portuguese, Polish, Danish, Dutch, Norwegian, Swedish, Czech, Hungarian, Romanian, Russian, Chinese, Korean, and Japanese. A thorough reconstruction of the editorial histories and criteria underlying each national edition would constitute a substantial and separate undertaking, well beyond the scope of this study.

Table 2. International Editions of 1001 Books You Must Read Before You Die by Peter Boxall (Source: Author's Own Work)

Language	Title	Publishing period
English	1001 Books You Must Read Before You Die	2006-2018
German	1001 Bücher: die Sie lesen sollten, bevor das Leben vorbei ist	2007-2021
French	Les 1001 livres qu'il faut avoir lus dans sa vie	2007-2013
Spanish	Mil un libros que hay que leer antes de morir: relatos e historias de todos los tiempos	2006-2016

Language	Title	Publishing period
Dutch	1001 boeken die je gelezen moet hebben!	2007-2017
Italian	1001 libri da leggere nella vita. I grandi capolavori della narrativa mondiale	2007-2016
Danish	1001 bøger du skal læse før du dør	2010
Romanian	1001 de cărți de citit într-o viață	2007-2010
Polish	1001 książek, które musisz przeczytać	2008
Czech	1001 knih, které musíte přečíst, než zemřete	2011
Norwegian	1001 bøker du må lese før du dør	2007
Korean	죽기 전에 꼭 읽어야 할 책 1001권	2008
Hungarian	1001 könyv, amit el kell olvasnod, mielőtt meghalsz [a regény történetének teljes körű krónikája]	2007-2008
Japanese	世界の小説大百科：死ぬまでに読むべき1001冊の本	2013
Chinese	有生之年一定要读的1001本书 / You sheng zhi nian yi ding yao du de 1001 ben shu	2021
Portuguese	1001 livros para ler antes de morrer	2010
Swedish	1001 böcker du måste läsa innan du dör	2008-2010
Russian	1001 КНИГА, КОТОРУЮ нужно ПРОЧИТАТЬ. (под редакцией Питера Бокселла)	2007

This breadth of international distribution, paired with consistent structure and branding, situates the series as a rare case of transnational literary packaging aimed at shaping global taste. The range of languages and cultural markets in which it is published—although broad—is far from what could be considered universal. Significant cultural and linguistic spheres remain unrepresented in its editions, particularly those associated with Arabic-speaking countries, the Middle East, South Asia, and Southeast Asia. Rather than offering comprehensive global coverage, the series reflects the most integrated and interconnected segment of the international publishing system (Phillips; Clark, 2019)—one largely shaped by Euro-American and, to a lesser extent, East Asian literary circuits.

Regarding the selection criteria, while declaring its focus on the novel as a genre—a concept with broad boundaries—it endeavors not to exclude other types of fiction and narratives. This list exists in a contradiction between the global and the partial. As articulated by the editor-in-chief in the preface to the second international edition, the stated criterion acknowledges the impossibility of creating a canon valid for all literary cultures. In his own words, it is “a list animated by the spirit of the novel, by a love for what the novel is and does, but which, in any case, neither expects nor attempts to capture it, summarize it, or finalize it” (Boxall; Mainer, 2016, p. 12).

We conclude this overview of the 1001 Books series by highlighting that, although the title of the work refers to “books,” the selection consists almost exclusively of novels and short stories, deliberately excluding—except in some very specific editions—plays, poetry, essays, or works of scientific thought. This limitation is explicitly stated in the book’s introduction and aligns with other well-known publications by its main author (Boxall, 2013; 2019; 2024). It should be noted that this kind of “narrative bias” becomes especially salient when the results are compared with other lists of notable books that apply a broader inclusion criterion.

3. Objectives and Methodology

As noted in the introduction, previous research has recognized that Wikipedia can serve as a valuable source for understanding the reception and dissemination of literary works. In this study, we focus on the presence of literary works across different language editions of Wikipedia and their corresponding entries in Wikidata. The existence of

articles about a work across multiple Wikipedia editions—within a massively used and widely consumed online platform—signals the presence of an observable discourse about that work. This discourse has both an informational dimension, reflected in the multilingual written content across different language editions, and a factual dimension, represented by the structured data linked through Wikidata. Thus, our objective is to explore how a metric based on this multilingual and multi-format presence can be used not only to rank individual books, but also to evaluate sets of works.

This work builds on our own generalized Wiki3DRank methodology, originally developed for encyclopedic knowledge objects and explored in greater detail for literary works (Pastor-Sánchez, Saorín, & Baños-Moreno, 2023; 2024). As explained below, these studies validate the calculation of rankings based on encyclopedic data for elements of the same type—such as books, films, or paintings. When applied specifically to the domain of literature, this approach provides a framework for understanding books through varying levels of engagement, as reflected in the content and structured data voluntarily enriched by a large and globally distributed community of editors in Wikipedia and Wikidata. This type of user engagement differs from mere audience metrics, as it involves active editorial effort and the contribution of time and knowledge to create or enhance articles and their associated data.

We now take a step forward by operationalizing the Wiki3DRank metric—previously developed for evaluating the relevance of individual knowledge objects—in order to derive an aggregate centrality measure suitable for the comparative analysis of literary lists. By aligning each title from Boxall's selections with its corresponding Wikidata entity and associated Wikipedia articles, the ranking algorithm can be systematically applied to the dataset, yielding a unified metric representative of the entire collection.

Simultaneously, the study seeks to validate the performance of these metrics with respect to key aspects such as author's gender, language, and countries.

This study is based on processing the works included in all international editions of *1001 Books You Must Read Before You Die* to address the following research questions through systematized and enriched data:

- Can we assign a quantitative value to each of the title selections studied by applying the Wiki3DRank methodology?
- Is there a correlation between this measure and other metrics related to cultural diversity, gender, or language?

To understand the calculation method used, it is important to outline the fundamentals of the Wiki3DRank methodology. Although the complete details have been presented in previous publications by the authors, the essential components are summarized here to support the reader's understanding of the present analysis.

Wiki3DRank is computed as the magnitude of the vector formed by values from three dimensions related to editorial effort within the Wikimedia ecosystem: the number of articles across various Wikipedia editions (N_{Wikis}), the total number of words in those articles (N_{Words}), and the number of property statements in Wikidata (N_{Props}).

For this study, an additional centrality measure was developed to generate aggregate results for sets of items by calculating the centroid of the vectors that constitute a collection or selection.

Wiki3DRank is a per-item metric with a statistically normalized distribution. Moreover, previous research demonstrates that it is consistent with results obtained through other techniques, such as clustering. The initial proposal computed Wiki3DRank as the direct sum of the logarithmic transformations of the three dimensions. It was ultimately refined to a three-dimensional vector, where each component represents the logarithmic transformation of one of the magnitudes.

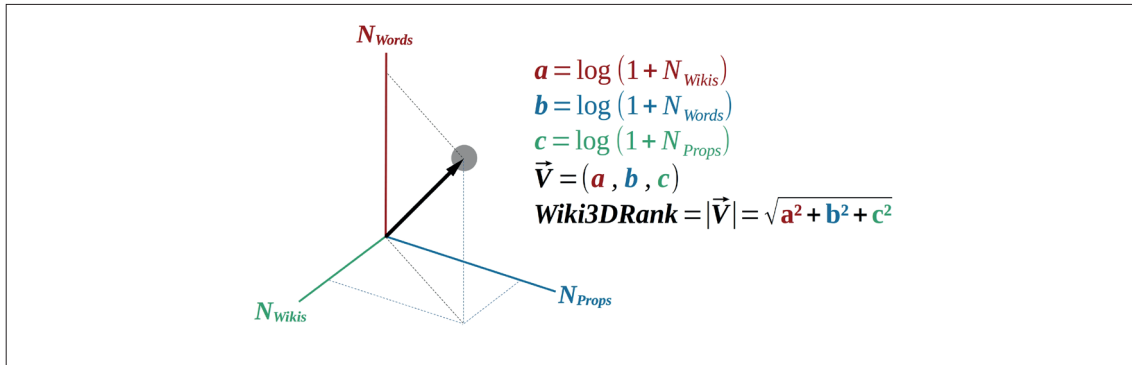


Figure 1: Calculation of Wiki3DRank as the Magnitude of a Vector (Source: Author's Own Work)

With a set of vectors, it is possible to perform various operations based on the three components (N_{Wikis} , N_{Words} , and N_{Props}), such as centroid, aggregation, or maximum calculation. For a given set of elements, the centroid—which we will use as a measure of centrality—allows for the comparison of entire sets without the need to compare each of their numerous individual items. It is calculated as the vector whose components are derived from the mean of the corresponding components of the individual vectors in the set.

$$\text{Wiki3DRank}_{Cent} = |\vec{V}_{Cent}| = \sqrt{\left(\frac{\sum a}{n}\right)^2 + \left(\frac{\sum b}{n}\right)^2 + \left(\frac{\sum c}{n}\right)^2}$$

To gather the data used in this study, the following steps were carried out:

Step 1: Bibliographic exploration in WorldCat and other collective catalogs to locate all editions of Peter Boxall's book in all languages. Subsequently, the list of titles included in each edition was compiled. This list was constructed using various methods, depending on access to the mentioned editions: a) Consultation of the physical copy or digital version, b) Requesting scanned copies of the index pages from Instituto Cervantes libraries for processing, and c) Reviewing digital platforms and social reading networks where the titles of each edition were collected.

Step 2: Once the lists of works from each individual edition were obtained, they were consolidated into a single data file used to map the corresponding items in Wikidata. For this task, automation mechanisms were explored using the title and author of the works. One approach involved using OpenRefine to reconcile the records of the works with Wikidata. Another method utilized SPARQL queries to perform free-text searches based on the title of the work in the Wikidata Query Service (WDQS). However, preliminary results from both methods revealed the need for review and correction, often due to variability in titles and translations of the works, as well as a degree of heterogeneity in Wikidata's labeling data. Since the dataset size was not so vast, the matching process was performed manually. For cases where no results were initially obtained, the search procedures were expanded by including variations of the titles in different languages and conducting broader searches. This ensured that all works from the editions of Boxall's book were unified in a file that included their mapping to Wikidata items.

Step 3: The unified file was processed to obtain additional information. For this, a Python script was used to retrieve data by querying WDQS. The collected data and properties utilized are detailed in Table 3. Assigning Wikidata identifiers to authors was straightforward using property P50. Additionally, cases of anonymous authorship and a few instances of collective authorship were accounted for.

Table 3: Collected Data from Work Items and Properties Used (Source: Author's Own Work)

Gathered information	Wikidata Property
Item class	P31
Author	P50
Artistic form	P7937
Literary genre	P136
Publication or creation date	P577 / P571
Original work's language	P407 / P1476
Country of origin of the work	P495

Step 4: Using the Wikidata identifier for the authorship of the works, a similar process to that executed in Step 3 was carried out, but this time for the authors. In this case, data on gender (property P21) and nationality (P27) were obtained.

11

Step 5: A normalization process was applied to historical variants, both for the languages used in property P407 and for the original language of the title in P1476. Additionally, normalization was performed for the countries of origin of the works (P495) and the nationality of the authors (P27). This normalization process involved mapping both types of elements to their more contemporary counterparts. For example, the "Middle Spanish" language used during the Spanish Golden Age (Q5364419) was mapped to "Spanish"; for the "Weimar Republic" (Q41304), equivalence was established with "Germany" (Q183).

Step 6: The values for $N_{Wikis'}$, $N_{Props'}$ and N_{Words} were retrieved for each of the works compiled from the various editions of Boxall's canon. Using these values, the Wiki3DRank metric was calculated for each work. It is important to note that the date for all data retrieved from Wikidata and the calculated Wiki3DRank results is October 29, 2024.

Step 7: Since the list was coded using unique identifiers, it is possible to construct virtual aggregations—distinct from the printed editions—that are particularly useful for generating panoramic views of the dataset, as outlined in Figure 2. The purpose of these groupings is to enable a differentiated analysis of the original English edition versus the international versions derived from it. The individual study of each edition does not allow for a precise observation of global aspects or for identifying absences, presences, and overlaps. This was achieved by constructing defined sets through the application of intersection, union, and exclusion logics.

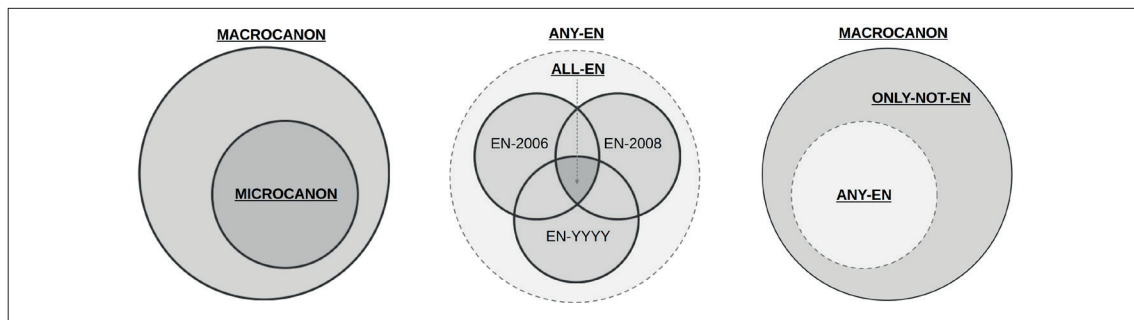


Figure 2: Definition of Groupings of Works from Boxall's Editions (Source: Author's Own Work)

Considering all editions of Boxall's book as individual sets of literary works ($B_1, B_2, \dots, B_n$) including the sets of English editions ($E_1, E_2, \dots, E_n$), the following groupings are defined:

- **MACROCANON:** Works present in any edition of Boxall's book. Defined as the union of all sets $\{B_1 \cup \dots \cup B_n\}$
- **MICROCANON:** Works present in all editions. Defined as the intersection of all sets $\{B_1 \cap \dots \cap B_n\}$

- **ALL-EN:** Works present in all English editions. Defined as the intersection of all English sets $\{E_1 \cap \dots \cap E_n\}$
- **ANY-EN:** Works present in any of the English editions. Defined as the union of all English sets $\{E_1 \cup \dots \cup E_n\}$
- **ONLY-NOT-EN:** Works present in any non-English edition but absent from all English editions. Defined as the difference between MACROCANON and ANY-EN $\{B_1 \cup \dots \cup B_n\} \setminus \{E_1 \cup \dots \cup E_n\}$

Step 8: The data was processed to generate various statistics:

- The number of works from each edition present in Wikidata.
- The number of works by language.
- The number of works by the author's country of origin.
- The number and percentage of co-occurring works across different editions.
- The number of works by the gender of the author in each edition of the canon.
- Calculation of the Wiki3DRank centroid for the groups of works in each edition of the canon.

12

4. Results

The results first present a systematization of the data obtained, including basic information about the different editions of Boxall's book and the aggregations outlined in Table 4. This section also includes the distribution of data related to the language of the works, as well as the nationality and gender of the authors.

Subsequently, a similarity analysis between the different editions of Boxall's work was conducted. This involved examining the overlapping works across the various editions used in this study.

4.1. Descriptive analysis of coverage in Wikidata and languages of the Works

After processing the dataset containing all editions of Boxall's work, a total of 1,718 works were identified, written in 79 languages (57 normalized), by 1,098 non-anonymous authors of 157 different nationalities (80 normalized). The following table provides a breakdown of the various editions, showing the total number of works in each one along with the number of corresponding items present and absent in Wikidata.

Table 4. Basic Data from the Different Editions of Boxall's Work (Source: Author's Own Work)

Editial	Total titles	Wikidata items	% Wikidata items
NL	1000	949	94.9
NO	1038	989	95.28
HU	939	902	96.06
RU	996	947	95.08
RO	1001	950	94.91
DK	1062	1004	94.54
PL	997	947	94.98
KO	993	948	95.47
JP	997	949	95.19
ES	998	936	93.79
IT	1000	949	94.9

Editial	Total titles	Wikidata items	% Wikidata items
FR	1001	952	95.1
DE	989	945	95.55
CZ	999	951	95.2
PT	998	951	95.29
ZH	995	950	95.48
EN-2006	998	955	95.69
EN-2008	1002	952	95.01
EN-2010	1001	953	95.2
EN-2012	1001	954	95.3
EN-2018	1003	957	95.41

The eye-catching title *1001 Books* used by the publisher is, in most cases, merely an approximation. It can also be observed that nearly all the works in each edition are represented in Wikidata, with the Spanish edition having the lowest correspondence ratio (93.79%).

13

In addition to the basic data obtained for each edition, several aggregations were generated, as outlined in the methodology section. These aggregations enable observations that go beyond individual editions. They may support a more comprehensive view of Boxall's work as a whole, facilitate the collective analysis of all English-language editions, and allow for the examination of international editions as a distinct group. Whether such aggregations contribute to refining center–periphery measures and revealing structural patterns that might remain hidden when editions are analyzed in isolation will be assessed in the discussion of the results.

Table 5. Basic Data of the Aggregated Lists (Source: Author's Own Work)

Aggregation	Total titles	Wikidata items	% Wikidata items
ALL-EN	705	679	96.31
ANY-EN	1316	1250	94.98
ONLY-NOT-EN	402	342	85.07
MACROCANON	1718	1592	92.67
MICROCANON	522	511	97.89

We observe a very high correspondence between the titles present across all editions of Boxall's book and items registered in Wikidata. However, as we move further away from the Anglosphere (ONLY-NOT-EN), the coverage rate decreases to 85.07%. Additionally, there is a smaller set of 522 works consistently included in all editions (MICROCANON), which shows the highest coverage in Wikidata. The total set of analyzed works reaches a 92.67% correspondence with Wikidata, confirming that it is a significant source for analyzing the studied objects.

With regard to languages, Table 6 presents data from the different editions and the percentage of works in each one that have a corresponding item in Wikidata. A color-coding system is used to highlight percentages over 50% (green), 10% (blue), 5% (yellow), and 2% (red). For space considerations, the table includes only languages that contribute more than 0.5% of the total works across all editions. The complete data is available in the dataset referenced in the methodology.

Table 6. Percentage of Languages of Works Present in Wikidata and Mentioned in the Various Editions of Boxall's Work (Source: Author's Own Work)

	en	fr	es	de	pt	it	nl	da	ru	no	jp	el	ca	sv	cs	pl
NL	53.0	10.5	5.9	8.3	1.5	3.8	4.2	0.2	3.0	0.7	1.8	0.6	0.3	1.2	0.8	0.7
NO	52.5	10.7	6.6	8.0	1.6	3.6	1.8	0.4	3.0	2.3	1.7	0.8	0.2	1.1	0.8	0.9
HU	71.1	10.5	1.6	6.4	0.8	2.9	0.6	0.2	3.3	0.2	0.8	0.4	0	0.7	0.7	0.2
RU	53.1	10.8	7.0	8.2	1.7	3.9	1.8	0.2	3.3	0.6	1.8	0.8	0.2	1.2	0.8	1.0
RO	52.7	11.1	6.9	8.2	1.7	3.9	1.9	0.2	3.2	0.6	1.8	0.8	0.2	1.2	0.8	0.9
DK	51.2	10.8	5.9	8.0	1.8	3.9	1.6	3.8	2.9	0.6	1.7	1.5	0.2	1.1	0.8	0.9
PL	53.6	11.1	6.4	8.3	1.7	3.8	1.7	0.2	3.1	0.6	1.7	0.7	0.2	1.2	0.8	1.0
KO	54.3	11.1	6.1	8.2	1.8	3.8	1.7	0.2	3.1	0.6	1.8	0.7	0.2	1.2	0.8	0.9
JP	54.1	11.1	6.3	8.4	1.8	3.7	1.7	0.2	3.1	0.6	1.8	0.7	0.2	1.2	0.8	0.8
ES	48.7	10.1	14.9	7.5	2.0	3.6	1.3	0.2	2.7	0.3	1.5	0.5	1.7	0.9	0.9	0.7
IT	52.7	11.5	6.1	8.3	1.6	4.6	1.8	0.2	3.1	0.6	1.8	0.8	0.5	1.2	0.8	0.9
FR	49.8	16.2	6.3	8.2	1.7	4.2	1.5	0.2	2.9	0.6	1.8	0.7	0.5	1.2	0.8	0.8
DE	72.0	9.8	1.6	6.1	0.8	2.8	0.6	0.2	3.2	0.2	0.8	0.4	0	0.6	0.6	0.2
CZ	53.9	11.1	6.3	8.4	1.7	3.7	1.7	0.2	3.0	0.6	1.8	0.7	0.2	1.2	0.8	0.9
PT	51.9	10.7	6.2	8.1	5.9	3.5	1.6	0.2	2.9	0.6	1.7	0.6	0.2	1.1	0.8	0.8
ZH	53.9	11.1	6.5	8.3	1.8	3.8	1.6	0.2	3.1	0.6	1.8	0.7	0.2	1.2	0.8	0.9
EN-2006	71.8	10.1	1.6	6.2	0.8	2.8	0.6	0.2	3.1	0.2	0.8	0.4	0	0.6	0.6	0.2
EN-2008	53.9	11.1	6.3	8.3	1.7	3.8	1.7	0.2	3.0	0.6	1.8	0.7	0.2	1.2	0.8	0.9
EN-2010	53.8	11.1	6.4	8.4	1.7	3.8	1.7	0.2	3.0	0.6	1.8	0.7	0.2	1.2	0.8	0.9
EN-2012	54.0	11.1	6.3	8.4	1.8	3.8	1.7	0.2	3.0	0.6	1.8	0.7	0.2	1.2	0.8	0.9
EN-2018	54.4	11.1	6.1	8.4	1.8	3.9	1.7	0.2	2.9	0.6	1.8	0.7	0.2	1.1	0.8	0.9

From the above data, it can be observed that literature in English consistently contributes approximately 50% of the items, starting from the first revision of the original 2006 edition (71.8%) and in international editions that almost entirely translate it, such as the German and Hungarian editions. French literature ranks second, followed by German literature in third place. An exception is the Spanish edition, which exhibits a "local bias" favoring works in Spanish, a trend also seen (though to a lesser extent) in the Portuguese edition. The first non-Western language represented is Japanese.

The same data for the previously described aggregations are shown below.

Table 7. Percentage of Languages of Works Present in Wikidata and Mentioned in the Various Aggregations of Boxall's Lists (Source: Author's Own Work)

	en	fr	es	de	pt	it	nl	da	ru	no	jp	el	ca	sv	cs	pl
ALL-EN	66.4	12.8	1.3	7.4	1.0	3.7	0.9	0.3	3.7	0.3	0.7	0.1	0	0.7	0.6	0.3
ANY-EN	61.3	9.3	5.4	7.2	1.4	3.1	1.3	0.2	2.6	0.5	1.6	0.8	0.2	1.0	0.8	0.7
ONLY-NOT-EN	7.6	15.5	24.3	1.5	12.9	3.8	7.9	10.5	1.2	5.6	0	2.6	4.4	0	0	0
MACROCANON	49.7	10.6	9.4	6.0	3.9	3.3	2.7	2.4	2.3	1.6	1.3	1.2	1.1	0.8	0.6	0.6
MICROCANON	65.9	12.7	1.2	8.2	0.8	4.3	0.4	0.4	4.3	0.2	0.8	0.2	0	0.6	0.8	0.4

The Anglo-Saxon focus is clearly evident in the table above. When examining only the English editions, the percentage surpasses 60%. A similar trend is observed in the works that consistently appear in all international editions (MICROCANON), where 65.9% are in English. Additionally, two particularities are worth noting:

- For the aggregation that considers all works mentioned in all editions (MACROCANON), the percentages are more diverse, and the prominence of Anglo-Saxon works decreases significantly to 49.7%, though they remain dominant.
- When analyzing the contribution of works not included in the English editions, Spanish has a much more pronounced presence than other widely spoken languages such as French, German, or Portuguese.

4.2. Descriptive analysis of authors' nationality and gender

When considering the percentages of nationalities (normalized to current countries) of authors contributing at least 1% relative to the total works across the different editions, Table 9 is obtained. In this case, the color-coding represents minimum percentages of 10% (green), 5% (blue), 3% (yellow), and 2% (red), respectively.

Table 8. Percentage of Nationalities of Works Present in Wikidata and Mentioned in Any of the Editions of Boxall's Work (Source: Author's Own Work)

	UK	US	FR	ES	IT	DE	RU	BR	NL	DK	IE	AT	NO	IN	AR	CA
NL	17.0	14.1	6.5	3.2	3.1	3.7	3.2	0.4	3.6	0.2	2.1	2.0	0.6	1.6	0.5	1.6
NO	17.2	13.5	6.8	3.4	3.0	3.3	3.1	0.5	1.4	0.4	2.2	2.3	2.3	1.4	0.8	1.5
HU	17.6	13.6	5.7	0.6	2.4	2.2	2.2	0.3	0.6	0.2	2.5	1.1	0.1	0.7	0.1	0.7
RU	17.1	13.5	6.7	3.6	3.3	3.4	3.2	0.5	1.4	0.2	2.3	2.4	0.5	1.6	0.8	1.5
RO	17.1	13.4	6.8	3.6	3.3	3.4	3.2	0.5	1.5	0.2	2.3	2.4	0.5	1.6	0.8	1.5
DK	16.1	13.3	6.8	3.0	3.5	3.5	3.0	0.5	1.2	3.0	2.0	2.1	0.7	1.5	0.6	1.4
PL	17.7	13.5	7.0	3.4	3.2	3.6	3.2	0.5	1.3	0.2	2.1	2.2	0.5	1.6	0.6	1.6
KO	17.0	14.7	7.0	3.2	3.1	3.5	3.3	0.5	1.3	0.2	2.1	2.2	0.5	1.6	0.5	1.6
JP	17.1	14.3	7.0	3.3	3.1	3.7	3.1	0.5	1.3	0.2	2.1	2.2	0.5	1.6	0.5	1.6
ES	15.9	13.1	6.9	8.9	2.9	3.1	2.9	0.6	0.9	0.2	2.0	2.0	0.2	1.2	1.6	1.1
IT	17.4	13.5	7.3	3.7	3.7	3.5	3.2	0.5	1.4	0.2	2.1	2.4	0.5	1.6	0.5	1.6
FR	15.8	13.7	10.8	3.3	3.4	3.6	3.4	0.5	1.2	0.2	1.9	2.4	0.5	1.5	0.4	1.6
DE	17.5	13.7	5.3	0.5	2.3	2.1	2.1	0.3	0.5	0.2	2.5	1.0	0.1	0.6	0.1	0.7
CZ	17.1	13.8	6.9	3.2	3.0	3.7	3.2	0.5	1.3	0.2	2.2	2.2	0.5	1.6	0.6	1.6
PT	16.6	13.7	6.7	3.0	2.8	3.6	2.8	3.6	1.3	0.2	2.2	2.1	0.5	1.6	0.5	1.4
ZH	17.2	13.9	6.8	3.3	3.2	3.6	3.2	0.5	1.3	0.2	2.3	2.2	0.5	1.6	0.6	1.6
EN-2006	17.3	13.6	5.3	0.5	2.4	2.1	2.1	0.3	0.5	0.2	2.5	1.0	0.1	0.6	0.1	0.7
EN-2008	17.6	13.7	6.9	3.3	3.2	3.6	3.2	0.5	1.3	0.2	2.2	2.2	0.5	1.6	0.6	1.6
EN-2010	17.1	13.7	6.9	3.3	3.1	3.7	3.1	0.5	1.3	0.2	2.3	2.2	0.5	1.6	0.6	1.6
EN-2012	17.0	14.4	6.9	3.2	3.1	3.7	3.1	0.5	1.3	0.2	2.1	2.2	0.5	1.6	0.5	1.6
EN-2018	17.0	14.6	6.9	3.1	3.1	3.6	3.1	0.5	1.3	0.2	2.3	2.2	0.5	1.6	0.5	1.6

It is evident that authorship from the European and Anglo-Saxon cultural spheres is clearly dominant. In the case of continental Europe, colonial powers such as Spain, France, Italy, Germany, and Russia hold prominence, followed by others politically or linguistically linked to them, such as Ireland or Austria. Authors from Anglo-Saxon countries (United Kingdom, United States) have the largest representation. They are followed by French authors in second place, closely followed by Spanish authors. For Portuguese-language works, it is notable that Brazil is the predominant nationality.

Table 10 presents the equivalent results for the aggregated lists.

Table 9. Percentage of Nationalities of Works Present in Wikidata and Mentioned in the Various Aggregations of Boxall's Lists (Source: Author's Own Work)

	UK	US	FR	ES	IT	DE	RU	BR	NL	DK	IE	AT	NO	IN	AR	CA
ALL-EN	22.7	17.1	7.4	0.4	2.9	2.8	2.8	0.4	0.7	0.3	2.8	1.2	0.1	0.6	0.1	1.0
ANY-EN	14.3	12.3	5.4	2.5	2.6	2.9	2.5	0.4	1.0	0.2	2.2	1.8	0.4	1.4	0.5	1.2
ONLY-NOT-EN	2.6	2.3	14.3	17.8	3.5	0.6	2.3	9.6	6.7	8.8	0	1.2	5.8	0	3.5	0.6
MACROCANON	11.5	10.1	7.0	5.6	2.8	2.4	2.3	2.2	2.1	1.9	1.7	1.6	1.5	1.1	1.0	1.0
MICROCANON	23.1	19.2	8.4	0.4	2.9	3.3	3.3	0.4	0.6	0.4	2.5	1.2	0.2	0.6	0	1.2

An indicator of the Anglo-Saxon overrepresentation can be observed in the following table, which shows that there are 44 authors with at least five works, 32 of whom belong to the Anglo-Saxon sphere. This means that 4% of the authors contribute nearly 16% of the total works in the Macrocanon.

Table 10. Contribution of Principal Authors to the Global Canon (Source: Author's Own Work)

Author	N. of Works	Nationality	Author	N. of Works	Nationality
Charles Dickens	10	UK	Margaret Atwood	6	CA
J. M. Coetzee	10	ZA	Thomas Bernhard	6	AT
Virginia Woolf	9	UK	George Eliot	5	UK
Graham Greene	8	UK	Anthony Trollope	5	UK
Samuel Beckett	8	EI	Fyodor Dostoyevsky	5	RU
Philip Roth	8	US	Émile Zola	5	FR
Don DeLillo	8	US	H. G. Wells	5	UK
Ian McEwan	8	UK	Joseph Conrad	5	PL
Thomas Hardy	7	UK	Thomas Mann	5	DE
D. H. Lawrence	7	UK	Wyndham Lewis	5	UK
Saul Bellow	7	US	George Orwell	5	UK
J. G. Ballard	7	UK	Ernest Hemingway	5	US
Salman Rushdie	7	US	Italo Calvino	5	IT
José Saramago	7	PT	Vladimir Nabokov	5	RU
Paul Auster	7	US	Georges Perec	5	FR
Jane Austen	6	UK	Toni Morrison	5	US
Ivan Turgenev	6	RU	Martin Amis	5	UK
Henry James	6	US	John Banville	5	EI
Edith Wharton	6	US	Kazuo Ishiguro	5	JP
Henry Green	6	UK	Iain Banks	5	UK
Elizabeth Bowen	6	EI	Haruki Murakami	5	JP
Iris Murdoch	6	EI	TOTAL	273 works	
Thomas Pynchon	6	US			

However, as shown in Figure 2, the number of authors relative to the number of works generally follows a logarithmic series distribution, indicating the probability of an author appearing in the canon with K works.

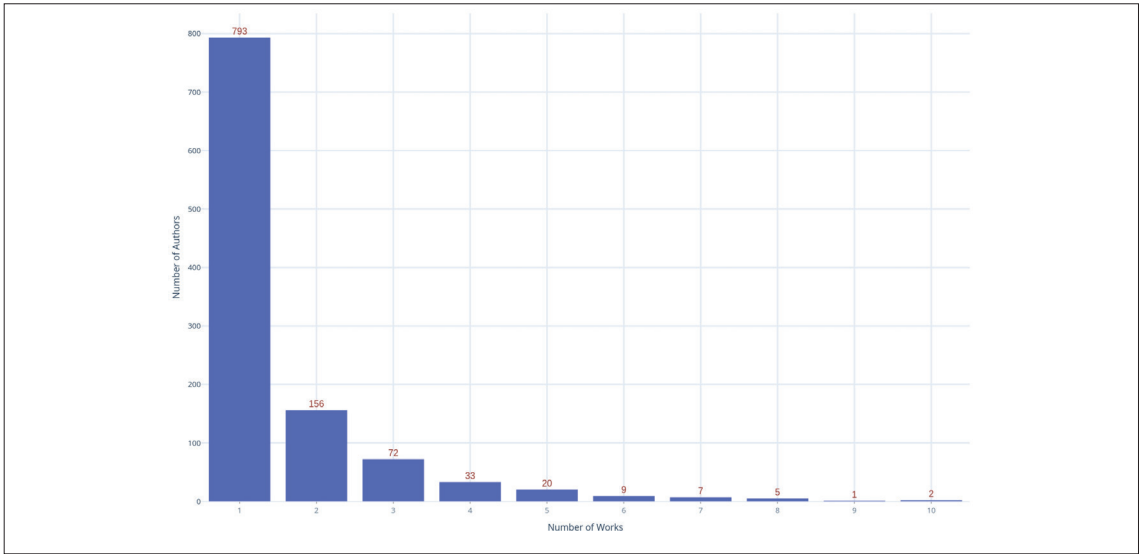


Figure 3. Visualization of the Distribution of Number of Works per Author (Source: Author's Own Work)

This probability is estimated as:

$$f(K)=\frac{-1}{\ln(1-p)}\times\frac{p^K}{K}$$

Where K is the number of works by an author, and the parameter p takes an approximate value of 0.5.

For the gender of authors, the following data has been obtained, including the number of anonymous works. Overall, the gender distribution remains fairly consistent across editions (around 80% male). Therefore, we present only the data for the aggregations. It is noteworthy that for works contributed to the global canon by international editions but not included in the English editions, this percentage rises by nearly ten points (90.35%) compared to the global canon (82.16%).

Table 12. Gender Distribution in Authorship (Source: Author's Own Work)

Aggregation	Total	Men		Women		Others		Anonymous	
		Total	%	Total	%	Total	%	Total	%
ALL-EN	679	557	82.03	129	19	0	0	1	0.15
ANY-EN	1250	999	79.92	258	20.64	1	0.08	3	0.24
ONLY-NOT-EN	342	309	90.35	32	9.36	0	0	3	0.88
MACROCANON	1592	1308	82.16	290	18.22	1	0.06	6	0.38
MICROCANON	511	420	82.19	96	18.79	0	0	1	0.2

Female authorship remains stable across all the versions analyzed, ranging approximately between 18–20%. The presence of other genders and anonymous authorship is minimal.

4.3. Analysis of similarity and overlap in Boxall's editions

Another aspect of interest was the nature of the international editions, aiming to identify local contributions and variability compared to editions for the Anglo-Saxon book market. To this end, the degree of concordance between the various editions was calculated to measure the extent to which one edition is similar to another. The average similarity percentage across all editions is 87.68%.

To represent this information, all concordance values were distributed into quartiles, yielding the following values: Q25=74.47%, Q50=92.58%, and Q75=95.78%. Based on these figures, the following table shows the editions of Boxall's work with the most similar selections of works (above 94%, in green) and those with the greatest differences (below 75%, in red).

It is important to note that the table is not symmetrical because the number of works included in Wikidata across different editions varies. Consequently, the similarity percentages between pairs of editions differ depending on the total number of works in the edition used to calculate the concordance percentage.

Table 13. Overlap Between International Editions (Source: Author's Own Work)

	NL	NO	HU	RU	RO	DK	PL	KO	JP	ES	IT	FR	DE	CZ	PT	ZH
NL	100	92.94	67.97	91.36	91.78	94.84	93.57	95.68	95.26	83.25	91.15	88.09	68.39	94.1	89.99	92.73
NO	89.18	100	67.14	93.73	94.24	92.01	94.03	91.3	92.01	80.69	91.51	85.74	68.86	93.02	88.88	91.51
HU	71.51	73.61	100	74.61	74.83	73.28	74.17	73.39	73.73	68.51	73.28	69.96	98.56	74.39	72.28	73.39
RU	91.55	97.89	71.07	100	99.47	94.19	95.99	93.77	94.51	84.16	93.56	89.33	73.07	95.25	91.24	93.66
RO	91.68	98.11	71.05	99.16	100	94.42	96.11	93.89	94.63	84.11	93.89	89.37	72.95	95.37	91.37	93.79

	NL	NO	HU	RU	RO	DK	PL	KO	JP	ES	IT	FR	DE	CZ	PT	ZH
DK	89.64	90.64	65.84	88.84	89.34	100	91.53	92.43	93.43	80.48	88.55	85.96	66.43	92.53	88.25	90.94
PL	93.77	98.2	70.64	95.99	96.41	97.04	100	96.62	97.36	84.27	96.3	90.18	71.7	98.31	93.77	96.62
KO	95.78	95.25	69.83	93.67	94.09	97.89	96.52	100	98.42	84.7	93.46	90.72	70.36	97.26	92.72	95.78
JP	95.26	95.89	70.07	94.31	94.73	98.84	97.15	98.31	100	85.56	93.99	91.36	70.71	98.1	93.57	96.42
ES	84.4	85.26	66.03	85.15	85.36	86.32	85.26	85.79	86.75	100	84.08	83.65	66.56	86	82.8	84.83
IT	91.15	95.36	69.65	93.36	93.99	93.68	96.1	93.36	93.99	82.93	100	89.67	70.6	95.47	91.36	93.99
FR	87.82	89.08	66.28	88.87	89.18	90.65	89.71	90.34	91.07	82.25	89.39	100	66.91	90.13	86.97	89.08
DE	68.68	72.06	94.07	73.23	73.33	70.58	71.85	70.58	71.01	65.93	70.9	67.41	100	71.75	69.52	70.79
CZ	93.9	96.74	70.56	94.85	95.27	97.69	97.9	96.95	97.9	84.65	95.27	90.22	71.29	100	95.16	98
PT	89.8	92.43	68.56	90.85	91.27	93.17	93.38	92.43	93.38	81.49	91.17	87.07	69.09	95.16	100	93.69
ZH	92.63	95.26	69.68	93.37	93.79	96.11	96.32	95.58	96.32	83.58	93.89	89.26	70.42	98.11	93.79	100

It is clear that two editions stand out as being substantially different from the rest while being similar to each other: the Hungarian and German editions. At the same time, it can be observed that the Spanish and French editions never exceed the similarity threshold of 94% (yellow) with any other edition. A similar pattern, though less pronounced, can be seen with the Portuguese, Italian, and Danish editions.

When the same calculations are performed using the same criteria for the five English editions of Boxall's book in relation to the non-English editions, the following results are observed.

Table 14. Overlap Between English Editions and International Editions (Source: Author's Own Work)

	NL	NO	HU	RU	RO	DK	PL	KO	JP	ES	IT	FR	DE	CZ	PT	ZH
EN-2006	68.9	72.25	94.24	73.3	73.51	70.79	71.94	70.79	71.2	66.18	71.1	67.54	98.85	71.94	69.74	70.99
EN-2008	93.49	97.9	70.59	95.38	95.8	96.85	99.05	96.43	97.16	84.03	96.43	89.71	71.53	98.74	94.22	97.06
EN-2010	93.91	96.85	70.41	94.96	95.38	97.69	98.01	96.96	97.9	84.68	95.38	90.35	71.14	99.69	95.17	98.01
EN-2012	94.97	95.6	69.81	94.03	94.44	98.64	96.86	98.11	99.06	85.74	94.23	91.19	70.44	98.43	93.82	96.86
EN-2018	95.51	94.88	69.49	93.31	93.73	97.49	96.13	98.54	98.01	84.85	93.52	90.28	70.11	97.39	92.89	95.82

In general, there is a strong similarity between all non-English editions and the works included in the English editions. However, it is important to note two aspects. First, the 2006 English edition shows a low percentage of similarity with the other editions, except for the Hungarian and German editions. Second, the Spanish and French editions have the lowest overall degree of similarity with the English editions.

To conclude the analysis of concordance between editions, calculations were also performed among the different English editions. As shown in the following table, all English editions exhibit a high degree of similarity with each other, except for the 2006 edition.

Table 15. Overlap Among English Editions Only (Source: Author's Own Work)

	EN-2006	EN-2008	EN-2010	EN-2012	EN-2018
EN-2006	100	72.25	71.94	71.31	71.2
EN-2008	72.48	100	99.05	97.9	97.48
EN-2010	72.09	98.95	100	98.74	98.01
EN-2012	71.38	97.69	98.64	100	99.16
EN-2018	71.06	96.97	97.6	98.85	100

4.4. Application of Wiki3DRank for calculating encyclopedic relevance metric

After these descriptive analyses of coverage, languages, nationality, and gender of the authors, we now complement this data with the results obtained using the Wiki3DRank metric. This metric provides a quantitative assessment of sets of items to represent aspects of their global encyclopedic relevance.

The Wiki3DRank calculation was performed for the centroids of the works included in each edition and aggregation.

Table 16. Calculation of Encyclopedic Relevance (Wiki3DRank) for Each Selection of Works in Boxall's Canon (Source: Author's Own Work)

Edition / Aggregation	Wiki3DRank _{Cent}	Edition / Aggregation	Wiki3DRank _{Cent}
NL	9.160235	CZ	9.188282
NO	9.131698	PT	9.154538
HU	9.200438	ZH	9.200397
RU	9.189686	EN-2006	9.134628
RO	9.185409	EN-2008	9.194845
DK	9.174244	EN-2010	9.189159
PL	9.203002	EN-2012	9.200532
KO	9.206763	EN-2018	9.195978
JP	9.198508	ALL-EN	9.538547
ES	9.116488	ANY-EN	8.946492
IT	9.193111	ONLY-NOT-EN	7.583044
FR	9.173427	MACROCANON	8.652339
DE	9.132877	MICROCANON	9.805333

The “weight” measure for each selection is fairly consistent across all editions. The standard deviation is 0.027, indicating a minimal range of dispersion in the Wiki3DRank values of the centroids for the editions. The difference between the Wiki3DRank value of the highest-ranked edition (Korean edition) and the lowest-ranked edition (Spanish edition) is less than 1%.

Regarding the aggregations of works, it can be observed that the Wiki3DRank for the aggregated set of all works (MACROCANON) is significantly lower than that of the works appearing in all editions of Boxall (MICROCANON), with a difference exceeding 11%. A similar trend is evident for works appearing only in non-English editions (ONLY-NOT-EN) compared to other aggregations, with a difference of more than 20% compared to the works appearing in all English editions (ALL-EN).

5. Discussion

Based on the data obtained and presented in the previous results section, it is possible to analyze the impact of certain factors on the relevance or centrality outcome of Wiki3DRank. To structure this discussion, the section is organized into four thematic blocks: (1) gender imbalance, (2) linguistic representation, (3) the distribution of work rankings according to gender and language, and (4) comparisons with other selected-book lists.

It is worth debating whether there is any bias in gender representation in the calculation of Wiki3DRank. Although, as previously shown, the initial selection is already markedly unbalanced (with 80% of the works authored by men), it is still possible to further examine how the ranking values assigned by Wiki3DRank to each book vary according to the author's gender.

To address this, the centroid of Wiki3DRank for works written by female authors was calculated. Additionally, the percentage of female authors and the total and percentage differences were computed. In this sense, taking the Dutch edition (NL) as an example, with a centroid of 9.16, the corresponding value restricted to works authored by women is lower, at 8.67. There is a difference of 5% between the two measures. It is important to stress that we are not examining disparities in presence but rather in evaluation.

In this section presents several visualizations designed to facilitate the reading and interpretation of the data, while the complete data tables are provided in Appendix I.

The chart below clearly shows how the editions similar to the first English edition of 2006 exhibit not only lower global centroid rankings but also lower rankings for works authored by women. It is also noteworthy that, in the international editions, the average centroid value for works by female authors is higher than in the English editions, with Spanish standing out prominently, followed by French to a lesser extent. It is difficult to determine to what extent variations in the criteria used to select works and authors influence the rankings obtained, or whether these rankings instead reflect biases inherent in Wikipedia and Wikidata themselves.

20

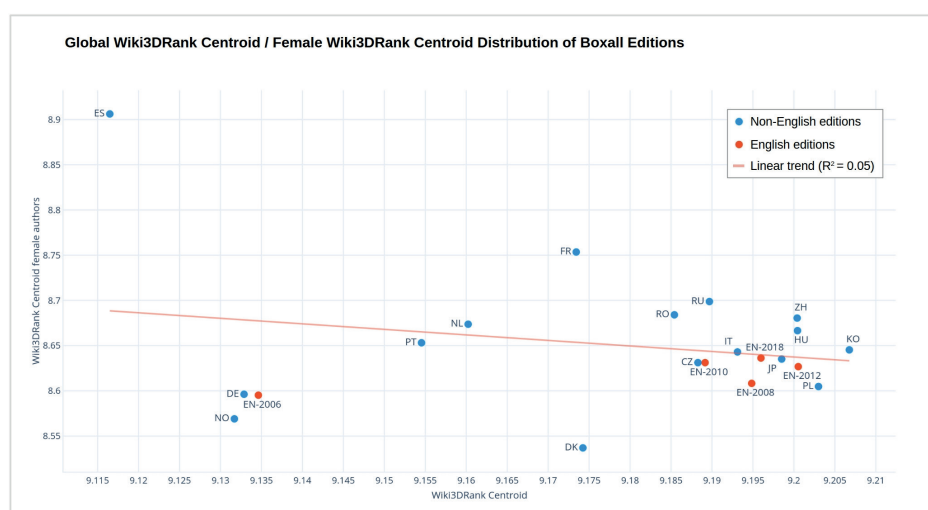


Figure 4. Scatter plot of the different Boxall editions, combining centroid values for works authored by women (Y-axis) and total centroid values (Source: Author's Own Work)

It can be observed that in all editions of Boxall's work, the percentage of female authors compared to their male counterparts ranges from 17.09% (Spanish edition) to 20.36% (Korean edition). Additionally, there is a noticeable difference between the Wiki3DRank centroid for works by female authors and the centroid for the entire edition. This indicates that, beyond the well-documented lower representation of women, their works also receive lower scores in our relevance metric (See table A-1 in Appendix I).

We acknowledge the interest in visualizing the chronological evolution of these metrics. However, as shown by the overlap between editions, a strictly chronological view is problematic: some later editions are simply translations of earlier versions. Between 2006 and 2018, both the global centroid values and those for works authored by women showed slight increases, with the most notable growth occurring with the 2008 revision. After 2008, values remained relatively stable, indicating only minor adjustments over time. This upward trend is observed both in the overall dataset and in the ranking of works by women, even though the proportion of female-authored titles remains relatively stable.

It would be worth exploring further whether this trend applies broadly to literature in general, particularly pre-20th-century literature, or if it is specific to Boxall's selection. Such an analysis would require a more detailed study, which falls outside the scope of this work.

It is notable that the smallest percentage difference between the Wiki3DRank centroids for male and female authors is in the Spanish edition (only 2.307%). This contrasts with the lower representation of women in the selection of works for this edition, while at the same time, the selection of works by these authors holds nearly the same weight as those by male authors.

However, in general, the attention given to female authors on Wikidata and Wikipedia (according to Wiki3DRank data) is generally lower. This explains the reduced centroid values for female authors.

The Pearson correlation coefficient (ρ) between the general Wiki3DRank value and the percentage of female authors in each edition is $\rho=0.366$. Therefore, there is no significant correlation between these two factors.

When conducting a similar analysis for the centroids of the aggregations (see Table A-2, Appendix I), some interesting patterns can be observed. The most notable is the significant difference in Wiki3DRank for works by female authors that appear only in non-English editions (over 11%). On the other hand, the centroid values of Wiki3DRank for works by female authors included in MICROCANON and those appearing in all English editions (ALL-EN) show smaller differences compared to the global Wiki3DRank value for the aggregation. This may indicate greater attention to English-language female authors. However, a detailed and combined analysis of gender and nationality data would be required to confirm this.

21

The subsequent chart presents an integrated visualization of both individual editions and constructed aggregations. Boxall's editions coalesce into a relatively dense cluster, reflecting their shared centroid characteristics. Among the datasets, only the Spanish edition and the Macrocanon exhibit significantly higher centroid values for works authored by women. Importantly, the data suggest that the size of the corpus does not directly correlate with its ranking: *Only-no-EN* (342 works) records the lowest centroid value, whereas the *Microcanon* (511 works) attains the highest. Larger aggregations, such as the *Macrocanon* and *Any-EN*, occupy intermediate positions, with the aggregation encompassing all English editions achieving a comparatively superior ranking. Notably, the selective subset of works across all editions surpasses the English-only aggregation in centroid value, indicating a broader representation of works by women within the former.

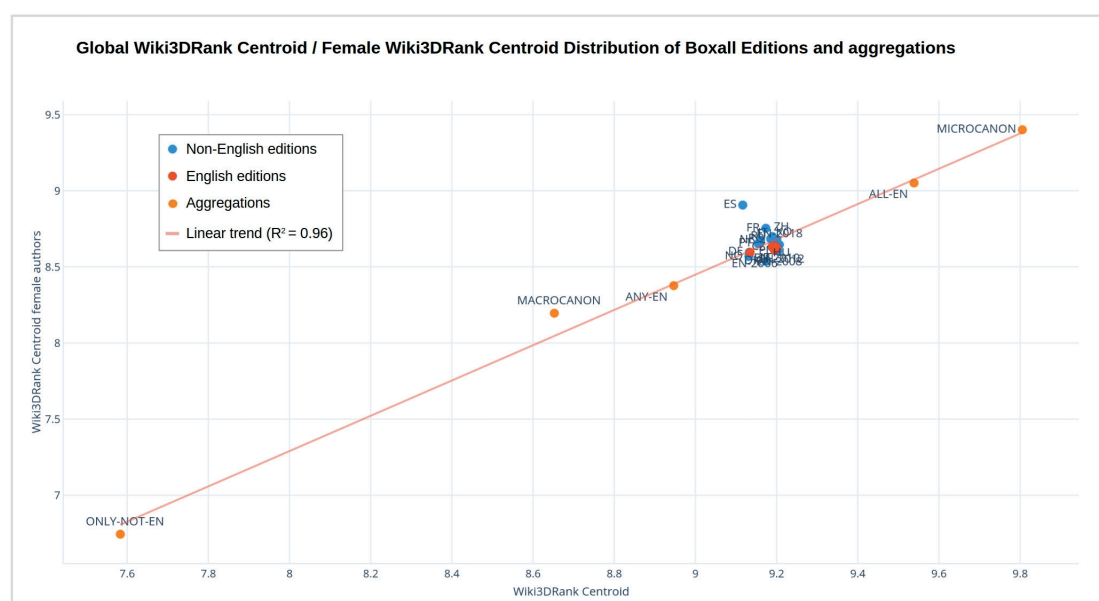


Figure 5. Scatter plot of the different Boxall editions, including constructed aggregations (Macrocanon, Microcanon, etc.), combining centroid values for works authored by women (Y-axis) and total centroid values (Source: Author's Own Work)

- To study the results in relation to the languages of the works, two comparisons were defined:
- Works written in English, as a representative language of economic and cultural globalization, versus works written in any other language.

Works written in one of the main Western languages (English, Spanish, French, Italian, German, and Russian) versus works written in any other language.

Regarding the first comparison (Table A-3, Appendix I), something already reflected in the data from Section 4.1 for the Hungarian and German editions becomes evident: a greater presence of works in English. Additionally, their centroids are the lowest among non-English editions. It can also be observed that the centroid for works in the 2006 edition is the lowest of all the English editions. This reinforces the observation that these three editions are almost identical.

In the following scatter plot, considering the centroid values of works written in English versus those written in other languages, we can observe notable differences between the Microcanon and Macrocanon. In the Microcanon, the centroid value for non-English works is remarkably high (10.122), indicating that, despite its smaller and highly selective nature, this aggregation places significant emphasis on highly central non-English works. Conversely, the Macrocanon, which represents a much larger and more inclusive corpus, records a substantially lower non-English centroid value (8.570), suggesting a broader and less central selection of non-English works. This contrast highlights that while the Microcanon prioritizes highly influential non-English literature, the Macrocanon incorporates a wider diversity of non-English texts, including those located further from the central core of the global literary system.

22

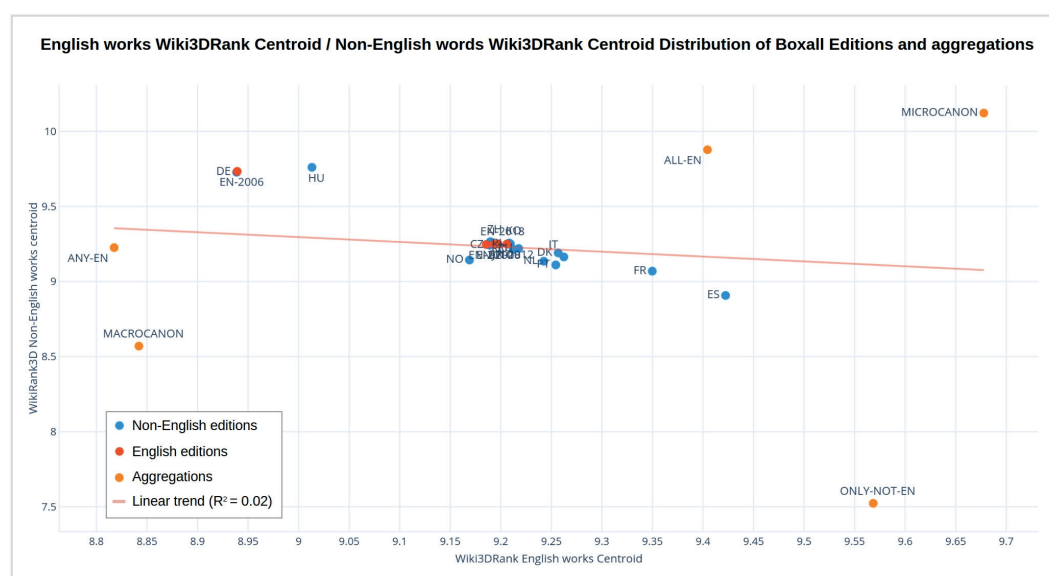


Figure 6. Scatter plot of the different Boxall editions, including constructed aggregations (Macrocanon, Microcanon, etc.), comparing centroid values for works authored in English (X-axis) and in other languages (Y-axis) (Source: Author's Own Work).

No correlation is observed between the general centroid value of each edition and the percentage of works in English ($\rho=-0.19$) or the percentage of works in other languages ($\rho=0.195$).

On the other hand, there is a strong inverse correlation between the Wiki3DRank value of each centroid for works in English or other languages and their corresponding percentages. For works in English, $\rho=-0.931$, and for works in other languages, $\rho=-0.971$. It could be stated that the lower the percentage of works in either category, the more selective the inclusion process, as it favors works of greater relevance. This can be cor-

Using the same format, we have represented the distribution of Wiki3DRank based on the gender of the authors (Figure 8) and to illustrate the different aggregations constructed in this study from the primary data of the various international editions of the book *1001 Books You Must Read Before You Die*, coordinated by Peter Boxall. In all cases, we observe the same organic pattern, with very few works deviating from the central mass. However, it may be worth exploring the existence of particularly outstanding elements, sometimes referred to in the literature as “hypercanonical” (Damrosch, 2006).

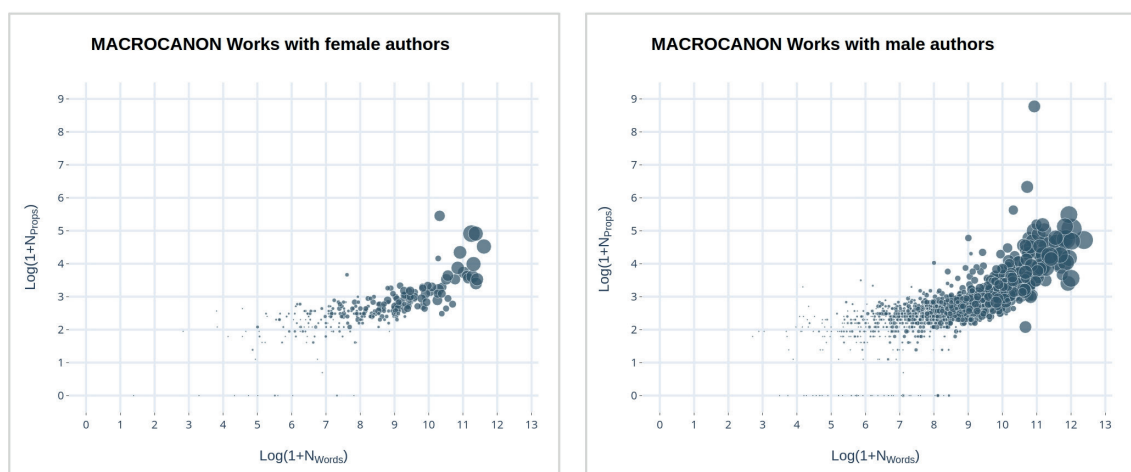


Figure 8. Distribution Graphs: Female Authors vs. Male Authors (Source: Author's Own Work)

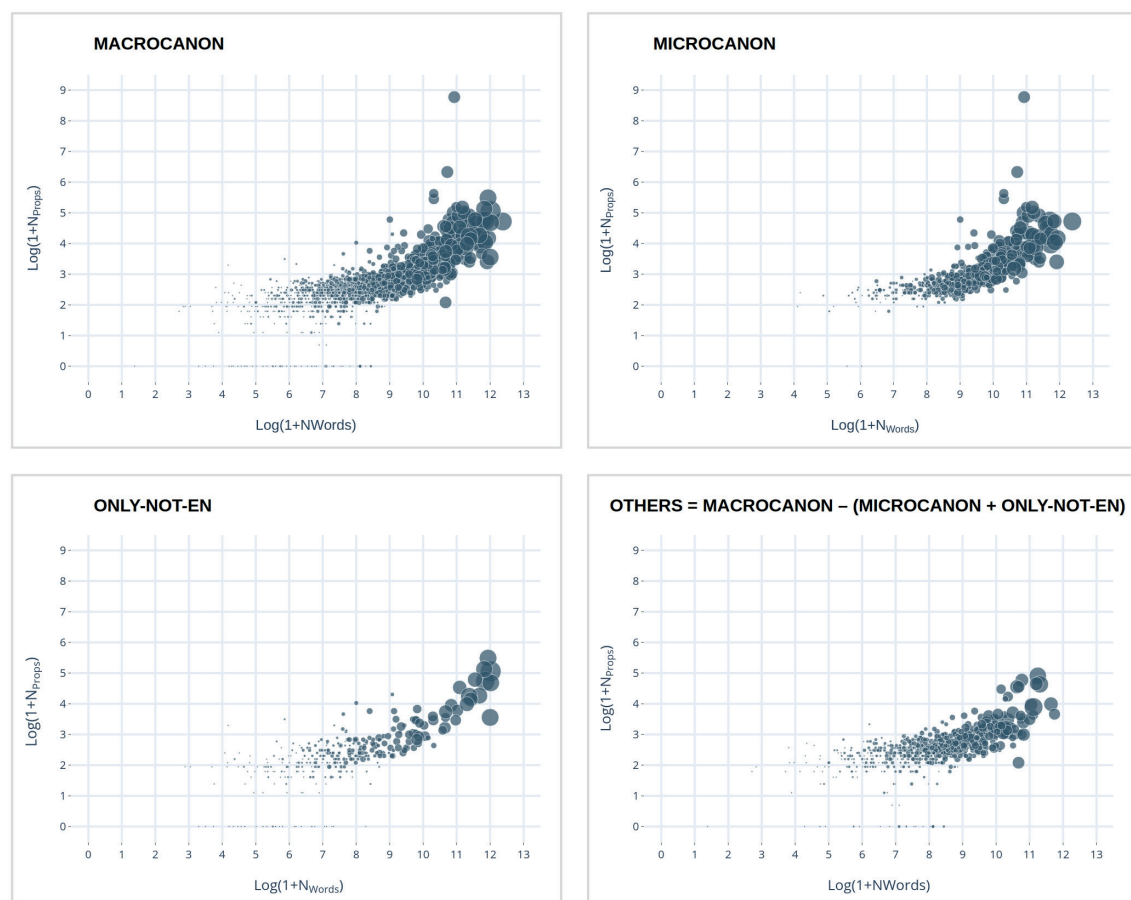


Figure 9. Distribution Graphs of MACROCANON, MICROCANON, ONLY-NOT-EN, and Remaining Works in MACROCANON (Source: Author's Own Work)

To conclude this discussion section, we propose a comparison of the results obtained with two lists of works designed to establish a literary canon.

A brief analysis was conducted on the book by German scholar Christiane Zschirnt, *Books: Everything You Need to Know* (Spanish edition), which includes a list of 141 works. Of the works proposed by the author, 139 have a corresponding item in Wikidata. However, only 30.9% of the works Zschirnt proposes are present in Boxall's MICROCANON, while this percentage increases to 45.3% in the MACROCANON. The Wiki3DRank centroid of the works included in Zschirnt's book is 11.4741.

From the data above, the low percentage of overlap between Zschirnt's and Boxall's selections is striking. One might expect a higher percentage given the small size of Zschirnt's selection (139 works). Particularly notable is the low percentage of overlap with the MICROCANON, which contains 522 works and could be presumed to represent a core of common and indisputable literary works. Even the overlap percentage with the MACROCANON, while higher, remains below 50%.

On the other hand, the difference in the centroid between Zschirnt's canon and Boxall's—1.67 points with the MICROCANON and 2.82 points with the MACROCANON—suggests that Boxall's selection includes works that, as a whole, are less relevant than those offered by Zschirnt.

The second list used for comparison is derived from previous research (Pastor-Sánchez; Saorín; Baños-Moreno, 2023) and is considerably larger than Zschirnt's or Boxall's, containing 107,434 works (WIKICANON). It is important to note that this dataset does not include some significant works due to issues with the assignment of Wikidata items, as it only retrieved books classified under the class *literary work* (Q7725634). For example, *The Divine Comedy* (Q40185) is classified as *written work* (Q47461344) rather than *literary work*, and thus was not included in the aforementioned study. Nevertheless, the dataset contains an extensive number of works, enabling meaningful comparisons with Boxall's selection.

The comparison was conducted by ordering the works according to their Wiki3DRank. The overlap between the 1,592 works with the highest Wiki3DRank and Boxall's MACROCANON is 17.02%, while the overlap with the MICROCANON is 20.74%. The Wiki3DRank centroid value is 10.4816 for the subset of works matching the size of the MACROCANON and 11.36793 for the equivalent subset of the MICROCANON. As in the case of the comparison with Zschirnt, the differences with the centroids of Boxall's MACROCANON and MICROCANON are significant.

An interesting complementary observation appears when examining what happens when the subset size is expanded and the degree of overlap with the MACROCANON and MICROCANON is analyzed, as shown in Figure 10.

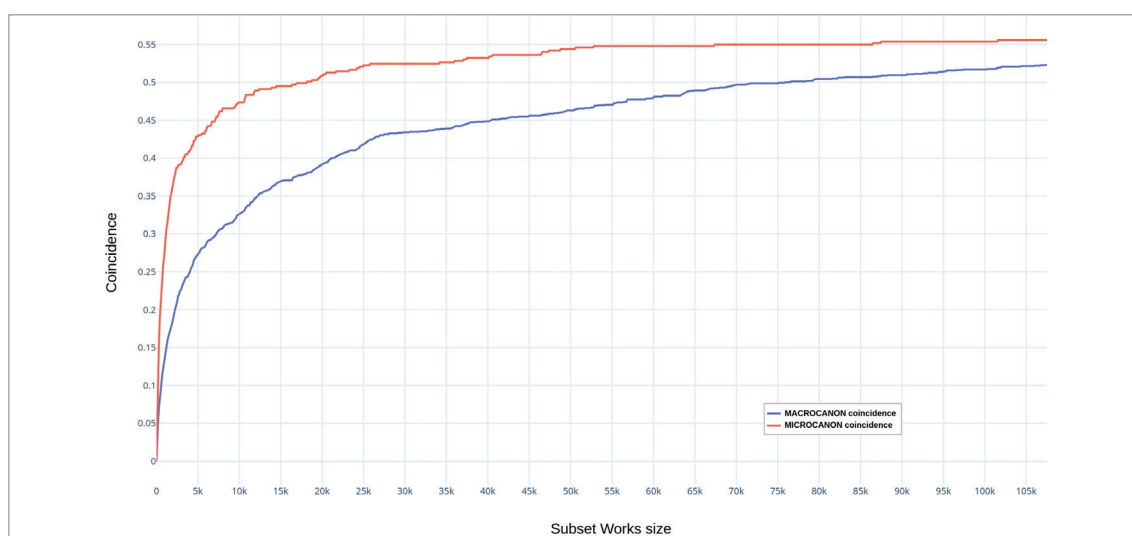


Figure 10. Coincidence Curve of Boxall's MICROCANON and MACROCANON with WIKICANON (Source: Author's Own Work)

It can be observed that the degree of coincidence with the MICROCANON increases rapidly, reaching 38% around a subset of 2,600 works, growing moderately to 50% within the subset of the 18,000 works with the highest Wiki3DRank, and reaching a maximum value of 55% overlap with the entire dataset (107,434 works).

In contrast, the MACROCANON does not exhibit the same rapid growth in overlap. Instead, it shows a moderate overlap of 45% within the first 41,000 works, eventually growing to 52% when considering the entire dataset.

This suggests that WIKICANON excludes a significant number of works from Boxall's selection. This discrepancy could be attributed to the previously mentioned issue of assigning Wikidata items to an appropriate class. Nonetheless, the rapid initial growth of the MICROCANON curve might also indicate that, despite quantitative differences, the selection approach of Boxall's MICROCANON aligns substantially with the works selected in WIKICANON.

26

To summarise the issue of gender in a few words: works by women writers are not only selected far less frequently but also tend to receive lower average scores of cultural importance. Nevertheless, a very slight positive trend can be observed across successive English and international re-editions.

Regarding language, in small sets of selected works, the distribution of scores between works in English and in other languages remains relatively balanced. However, as the number of books increases, a noticeable shift emerges in favour of works written in English.

Finally, when using Wiki3DRank to compare different curated lists, we can obtain a synthetic measure of their value or representativeness, complementing a simple overlap analysis. Preliminary results suggest that the metric becomes more meaningful when the lists being compared are of similar size, allowing for a more balanced assessment of their relative cultural prominence.

6. Conclusions

The first research question asks whether it is possible to assign a quantitative value to each of the selected titles by applying the Wiki3DRank methodology. We observe that the coverage in Wikidata–Wikipedia of works representative of literary culture is remarkably high —above 90% in all cases. Furthermore, by applying centrality calculations, a secondary value is obtained that scores each set of works as a whole. This metric, called Wiki3DRankCen (the centroid), is consistent with the vector-based calculation mechanism on which the measure is built. The second question focuses on whether there is a correlation between this measure and other metrics related to cultural diversity, gender, or language. The approach taken in this study offers a richer perspective by jointly examining both the coverage and the ranking values of the works according to these classic dimensions of cultural diversity.

By applying these metrics to a publishing project with multiple editions, we can observe that, despite the changes in the titles included in each, the overall value remains stable across editions. Therefore, to a certain extent, a degree of equivalence can be established between them—or put more playfully, all of Boxall's editions are worth the same.

At first glance, and as an explicit policy of Boxall's editorial project, it is evident that each new version of the canon incorporates greater diversity in terms of gender, cultures, geographies, and languages, without this affecting the overall value of the collec-

tion as measured by Wiki3DRank. We could conclude that the selected works, regardless of their origin (language, nationality, or gender of the author), are undoubtedly globally relevant.

To provide a better understanding of the metric's behavior, we examined correlations with aspects that characterize socio-cultural biases (author gender and original language) and analyzed the variability of the results. The values obtained do not show pronounced deviations caused by gender or language, which may indicate that the proposed metric can be used in various contexts to analyze selections of literary works.

That said, it is still significant to note the lower average values obtained when evaluating works written by women or in languages other than English. Despite the existence of dedicated Wikipedia editions for each language—and the active efforts of their communities to reduce gender gaps and highlight underrepresented cultures—the data still reflect a lower level of recognition for works from the peripheries.

27

Conversely, despite including language as an element of analysis, our study cannot fully capture the dynamics of circulation, commercialization, translations, and the drivers of the global literary system. We cannot determine whether a work written in Japanese achieves relevance because it has been translated and disseminated in the more powerful language markets. However, it is clear that this must have a considerable impact in a globalized market with very dominant Western actors at its core.

It is worth noting the value of this approach for comparing selections of works, not only through the overlap ratio but also by analyzing the Wiki3DRank values obtained for each selection and, in particular, through a centrality measure that facilitates comparisons—as in the comparative analysis involving Zschirnt's case, whose overall value exceeds that of any individual Boxall edition, including the constructed aggregations composed of the presumably more prominent books.

Although the comparison could be further refined to homogenize criteria for including books, this is not the focus of this study. Suffice it to say that, in both cases, the distribution by language and gender is similar, with lower representation of women and non-colonial cultures and languages.

Additionally, having a specific value for each work allows for the identification of certain works that might be considered lesser significant, potentially explained as editorial team compromises based on representational criteria or quotas (e.g., Irish authors, Galician authors, national treasures). It has also been observed that Wiki3DRank values for authors generally exceed those of any of their individual works. For instance, the authors Jane Austen and Virginia Woolf have scores of 13.66674 and 13.25398, respectively, compared to the highest scores of their works, *Pride and Prejudice* (13.11761) and *Mrs. Dalloway* (11.21329). This could suggest that the author encyclopedic entity also aggregates the relevance of their literary production, making it a more frequent subject in Wikimedia and other reference works or specialized dictionaries.

This study also suggests complementary approaches to better understand the accumulation of value in literary works, particularly for more recent works. These could include analyzing awards and recognitions for works and/or authors, the existence of film or television adaptations, official commemorations, derivative works, and other aspects of cultural lifecycle dynamics. Such an approach could potentially be implemented using data from Wikidata, although it would be necessary to assess whether it offers sufficient coverage for this purpose.

In general terms, the study shows the potential of using open and collaborative knowledge infrastructures, as long as appropriate and validated metric methods are applied. These infrastructures make it possible to observe both global and local dynamics in the visibility of creative works and in the formation of cultural canons. Each social or

collaborative platform contributes data that offer complementary perspectives. In particular, the Wikipedia–Wikidata ecosystem can reveal aspects that do not depend solely on large audiences or mass-attention phenomena. This helps to enrich the theoretical debate on global canonicity and on contemporary frameworks of legitimisation.

Future lines of research and application include comparing literary lists that share similar size and selection criteria. It is also possible to create metalists (lists of lists) or derived lists that combine elements from different sources. These computational methods could also be used to design services that support the Wikimedia community in addressing cultural gaps. One option is to develop dashboards that track literary works across languages. Such tools could include work scores to guide editor participation and to monitor progress. They would combine coverage and valuation metrics for literary works in each language across different Wikipedia editions.

Regarding the scalability of the approach, it is important to note that applying Wiki3DRank to other domains requires careful validation in each field. This is especially true for areas involving creative works or cultural objects. The calculation method may need adjustments for fast-moving and relatively young industries, such as audiovisual production or video games, where release cycles strongly affect visibility. In contrast, other domains—such as cultural or monumental heritage—operate on much slower temporal scales, which would also require specific adaptations. Even so, a measure like Wiki3DRank could still complement other metrics or inclusion criteria that come from different sources or perspectives.

Finally, it is worth emphasizing the utility of being able to “measure and weigh” sets of works from the perspective of their relevance. This simplifies the process of comparing different proposals or even observing the evolution within a single list of cultural objects. These calculations are easy to explain, transparent in their execution, based on open data, and achievable with minimal resources.

7. References

- Ahnert, Ruth, Sebastian E. Ahnert, Catherine Nicole Coleman & Scott B. Weingart. 2021. *The network turn: Changing perspectives in the humanities* (Elements in Publishing and Book Culture). Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108866804>
- Blank, Grant. 2007. *Critics, ratings, and society*. Lanham: Rowman and Littlefield.
- Borgman, Christine L. 2017. *Big data, little data, no data*. MIT.
- Bourdieu, Pierre. 1988. *La distinción. Criterios y bases sociales del gusto*. Madrid, Taurus.
- Boxall, Peter; José Carlos Mainer. 2016. *1001 libros que hay que leer antes de morir. Relatos e historias de todos los tiempos*. Grijalbo.
- Boxall, Peter. 2013. *The value of the novel and twenty-first-century fiction: A critical introduction*. Cambridge University Press.
- Boxall, Peter (coord.). 2019. *The Cambridge companion to British fiction*. Cambridge University Press
- Boxall, Peter. 2024. *The possibility of literature: The Novel and the politics of form*. Cambridge University Press. <https://doi.org/10.1017/9781009314305>
- Casanova, Pascale. 2004. *The world republic of letters*. Cambridge: Harvard University Press.
- Damrosch, David. 2006. World Literature in a postcanonical, hypercanonical age. In: Haun Saussy, ed., *Comparative Literature in an Age of Globalization*. Johns Hopkins University Press, pp. 43–53.
- De Riquer, Martín, José María Valverde 2005–2006. *Historia de la literatura universal*. 2 vols.

- Dirda, Michael. 2018. The ultimate literary bucket list: '1,000 books to read before you die'. *The Washington Post*, 26 September. https://www.washingtonpost.com/entertainment/books/the-ultimate-literary-bucket-list-1000-books-to-read-before-you-die/2018/09/25/ac7906a4-c020-11e8-be77-516336a26305_story.html
- Fischer, Frank, Jacob Blakesley, Paula Wojcik, Robert Jäschke (eds.). 2023. Wikipedia, Wikidata, and world literature. *Journal of Cultural Analytics*, 8(2). <https://culturalanalytics.org/issue/7259>
- Guillory, John 1993. *Cultural capital. The problem of literary canon formation*. Chicago: The University of Chicago Press.
- Hensher, Philip. 2006. A very loose canon. *The Spectator Magazine*, 25 February. <https://archive.spectator.co.uk/article/25th-february-2006/35/a-very-loose-canon>
- Kermode, Frank. 1999. *Formas de atención*. Gedisa.
- Kolbas, E. Dean. 2001. *Critical theory and the literary canon*. Westview Press.
- Kristensen, Nete Nørgaard, Unni From. 2017. *Cultural journalism and cultural critique in a changing media landscape*. Routledge.
- Lawson, Alison. 2020. The UK's reading culture and consumers' emotional response to books. In: Baverstock, Alison, Richaard Bradford, & Madelena Gonzalez (Eds.). (2020). *Contemporary publishing and the culture of books* (1st ed.). Routledge, pp. 58-76, <https://doi.org/10.4324/9781315778389>
- Lluch, Gemma (Dir). 2025. *Los libros que leemos ¿Quién, dónde y cómo se decide? La construcción del canon no académico*. Tirant lo Blanch.
- Pastor-Sánchez, Juan-Antonio, Tomás Saorín, & María-José Baños-Moreno. 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>
- Pastor-Sánchez, Juan-Antonio, Tomás Saorín, & María-José Baños-Moreno. 2024. Wiki3DRank: A model for measuring the relevance of knowledge objects using quantitative data from Wikidata and Wikipedia. *Ibersid: Revista de Sistemas de información y documentación*, 18(1), 55–70, <https://www.ibernid.eu/ojs/index.php/ibernid/article/view/4967/4377>
- Phillips, Angus, Giles Clark. 2019. *Inside book publishing* (6th Edition). Routledge.
- Pozuelo Yvancos, José María; Mariángeles Rodríguez Alonso. 2021. Veinte años de teoría: canon, pensamiento literario e historiografía. *Theory Now, Journal of literature, critique and thought*, 4(1). <http://dx.doi.org/10.30827/TNJ.v4i1.15928>
- Rohe, Jonas, Viktor J. Illmer, Lisa Poggel, Frank Fischer. 2025. Wikipedia as an echo chamber of canonicity: '1001 books you must read before you die'. In: *DH2025: "Accessibility and Citizenship"*, pp. 14–18, July. Universidade NOVA de Lisboa.
- Sullà, Enri. 1998. *El canon literario*. Madrid: Arco libros.
- Verma, Henrietta. 2022. Reading listopia | Books of lists and lists of books. *Library Journal*, 29 November. <https://www.libraryjournal.com/story/Reading-Listopia>
- Von Contzen, Eva (2021). Theorising lists in literature: Towards a listology. In: *Lists and catalogues in ancient literature and beyond*. De Gruyter, pp. 35-55. <https://doi.org/10.1515/9783110712230-003>
- Zschirnt, C. 2011. *Libros: todo lo que hay que saber*. (1ª ed). Taurus.

Appendix I

Appendix I provides detailed comparative data supporting the analyses presented in the main text. It includes tables on the calculation of global Wiki3DRank centroids for each edition of Boxall's work and for works authored by women, data on aggregations, percentages of works written in English, and a summary of correlations related to authorship gender and language.

Table A-1. Comparative Data on the Calculation of Global Wiki3DRank Centroids for Each Edition of Boxall's Work and Centroids for Works by Female Authors (Source: Author's Own Work)

Edition	Wiki3DRank Centroid	Female authors (%)	Wiki3DRank Centroid female authors	Difference	Difference (%)
NL	9.160235	19.71	8.673688	-0.486547	-5.312
NO	9.131698	18.71	8.569059	-0.562639	-6.161
HU	9.200438	19.96	8.666568	-0.53387	-5.803
RU	9.189686	17.95	8.698711	-0.490975	-5.343
RO	9.185409	17.79	8.684057	-0.501352	-5.458
DK	9.174244	19.12	8.536932	-0.637312	-6.947
PL	9.203002	19.11	8.604798	-0.598204	-6.500
KO	9.206763	20.36	8.645325	-0.561438	-6.098
JP	9.198508	19.49	8.635175	-0.563333	-6.124
ES	9.116488	17.09	8.906164	-0.210324	-2.307
IT	9.193111	18.967	8.643003	-0.550108	-5.984
FR	9.173427	18.382	8.753493	-0.419934	-4.578
DE	9.132877	19.788	8.596228	-0.536649	-5.876
CZ	9.188282	19.453	8.631197	-0.557085	-6.063
PT	9.154538	18.612	8.653138	-0.5014	-5.477
ZH	9.200397	19.474	8.68046	-0.519937	-5.651
EN-2006	9.134628	19.895	8.595214	-0.539414	-5.905
EN-2008	9.194845	19.223	8.608263	-0.586582	-6.379
EN-2010	9.189159	19.412	8.631197	-0.557962	-6.072
EN-2012	9.200532	19.706	8.62678	-0.573752	-6.236
EN-2018	9.195978	19.958	8.636265	-0.559713	-6.086

Table A-2. Comparative Data on the Calculation of Global Wiki3DRank Centroids and Centroids for Works by Female Authors in Aggregations (Source: Author's Own Work)

Aggregation	Wiki3DRank Centroid Aggregation	Female authors (%)	Wiki3DRank Centroid female authors	Difference	Difference (%)
ALL-EN	9.538547	18.999	9.051039	-0.487508	-5.111
ANY-EN	8.946492	20.64	8.37676	-0.569732	-6.368
ONLY-NOT-EN	7.583044	9.357	6.74408	-0.838964	-11.064
MICROCANON	9.805333	18.787	9.400959	-0.404374	-4.124
MACROCANON	8.652339	18.216	8.19593	-0.456409	-5.275

Table A-3. Percentage of Works Written in English (Source: Author's Own Work)

Edition or Aggregation	Works in English		Works in other languages	
	Wiki3DRank Centroid	%	Wiki3DRank Centroid	%
NL	9.242499	53.003	9.13526	50.474
NO	9.168975	52.477	9.143527	50.859
HU	9.013188	71.064	9.760985	32.373
RU	9.217745	53.115	9.220148	50.264
RO	9.212276	52.737	9.214416	50.632
DK	9.256744	51.195	9.19108	52.789
PL	9.209472	53.643	9.251075	49.63
KO	9.208277	54.325	9.256977	49.051
JP	9.207017	54.057	9.249222	49.42
ES	9.422324	48.718	8.907519	54.38
IT	9.262358	52.687	9.164234	50.474
FR	9.349805	49.79	9.06945	53.466
DE	8.938745	71.958	9.728161	31.429
CZ	9.188863	53.943	9.241856	49.422
PT	9.2544	51.945	9.110943	51.42
ZH	9.189598	53.895	9.265631	49.474
EN-2006	8.939253	71.832	9.733878	31.414
EN-2008	9.193324	53.887	9.248584	49.37
EN-2010	9.185597	53.83	9.24696	49.528
EN-2012	9.206093	53.983	9.250433	49.476
EN-2018	9.197258	54.441	9.25114	49.007
ALL-EN	9.40445	66.421	9.877913	36.672
ANY-EN	8.817528	61.28	9.225983	42.08
ONLY-NOT-EN	9.568394	7.602	7.523179	95.614
MICROCANON	9.677764	65.949	10.122206	37.965
MACROCANON	8.842007	49.749	8.570496	53.58

Table A-4. Summary of Correlations Related to Gender of Authorship and Language of Works (Source: Author's Own Work)

	% female authors	% English	% not English	% occidental	% not occidental
Wiki3DRank Centroid general	0.366	-0.19	0.195	-0.26	0.266
Wiki3DRank Centroid Female authors	-0.619	-0.305	0.292	0.044	-0.07
Wiki3DRank centroid English	-0.608	-0.931	0.928	-0.738	0.721
Wiki3DRank centroid Not English	0.595	0.974	-0.971	0.835	-0.816
Wiki3DRank centroid Occidental	-0.005	-0.751	0.76	-0.894	0.9
Wiki3DRank centroid Not occidental	0.342	0.95	-0.951	0.963	-0.951