Juan Antonio Pastor Sánchez und Tomás Saorín,
Library and Information Sciences Department. Universidad de Murcia (Spain)
tsp@um.es / pastor@um.es

Measuring the literary field and creative works: the case of world literary canon according to Wikipedia and Wikidata

v.3 Repository

Freie Universität Berlin

International Workshop | Berlin, 10-11th october 2023
Wikipedia, Wikidata and Wikibase: Usage Scenarios for Literary Studies

Literary canon according to Wiki_data_pedia

UNIVERSIDAD DE **MURCIA**

This ongoing research project aims to **verify** the use of Wikidata and Wikipedia as a **source to identify a universal literary canon** in pursuit of a good enough answer to the question of **how many** books could be considered as an all-times global literary canon and, given this quantity, **which titles** should be included.

Concerns:

- Why the expected results are useful?
- Why Wikipedia/Wikidata?
- What is a literary work? What is a work? What is a book?
- How does these selected titles fit with a global point of view in World Literature field?

First public results

UNIVERSIDAD DE **MURCIA**

- There's more to the picture than meets the eye.
- Research datasets and scripts are openly released.

A universal literary canon based on multilingual encyclopedic data: Proposal of a method for the ranking of literary works using quantitative data obtained from Wikidata and Wikipedia.
REVISTA ESPAÑOLA DE DOCUMENTACIÓN CIENTÍFICA, , vol. 46, 3 (2023)
https://redc.revistas.csic.es/index.php/redc/article/view/1519

How to end lists of the best books once and for all.
THE CONVERSATION, April 2023

Wiki3DRank, a quantitative method for measuring relevance of knowledge objects using data and contents from Wikidata and Wikipedia.
IBERSID, XXVII International Meeting on Information and Documentation Systems, Zaragoza, october, 2022.

Please, choose your position

UNIVERSIDAD DE **MURCIA**

It's not informetry

It's not SEO

Digital Humanities

It's not audicence measurement

It's not bibliographic studies

Digital discourses & data about literaty works

Distant reading

Digital Cultural studies

Digital Media studies

Between mass media and humanities

UNIVERSIDAD DE **MURCIA**

- **It's cool to talk about literary canon**; it sounds **amusing**, polemic, colourful… but also trivial, pointless or **outmoded**.
- To influence over what is considered literary canon in a certain time span is the wet dream of many academics in the literary field.
- The very idea of **one canon** is a contended concept: Diversity in canon(s)
- **Who really needs a canon?**
- How concepts as **popularity / audience / sales** have to be included when studying literary canon?

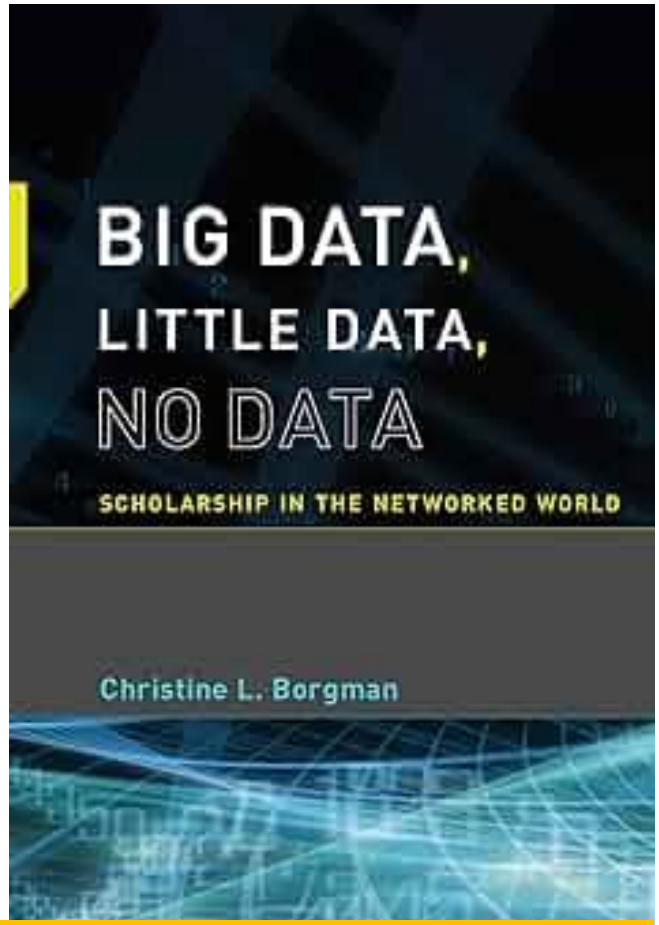
Canon is a kind or ranking

- Numerous **academics or institutions** elaborate lists of best authors, works, based in their own institutional and regional reputation.
- **Media and Publishing Industries** impact broadly in the global attention regarding books and creative works and also promote new proposals to the “big bag” of books to read.
- **Literary studies** foster authorized **systematic reference works** that remarks the most relevant works and authors in every language, culture, period or even in the global arena.
- **Canon is a kind of ranking;** ranks permeate all social fields; ranks are at the same time dumb and important.
- A global and diverse point of view is required.

Where are the suitable literary data?

UNIVERSIDAD DE **MURCIA**

- **Long-term** data about editions, printings, and sales are sparse and obscure.
- Global data about **publications is still disparate**, very incomplete and disperse.
- Data about sales, editions, readers, translation is not good data; there is a **data-hole** that makes measuring the kind of books that last over time and place unworkable.
- Media audience and digital tracking is a very recent phenomena; literature is a long-term cultural pratice.



In plain english: Does structured data sources about the number of editions of any of the books featuring Sherlock Holmes and their print run in whatever language and in whatever time exist?

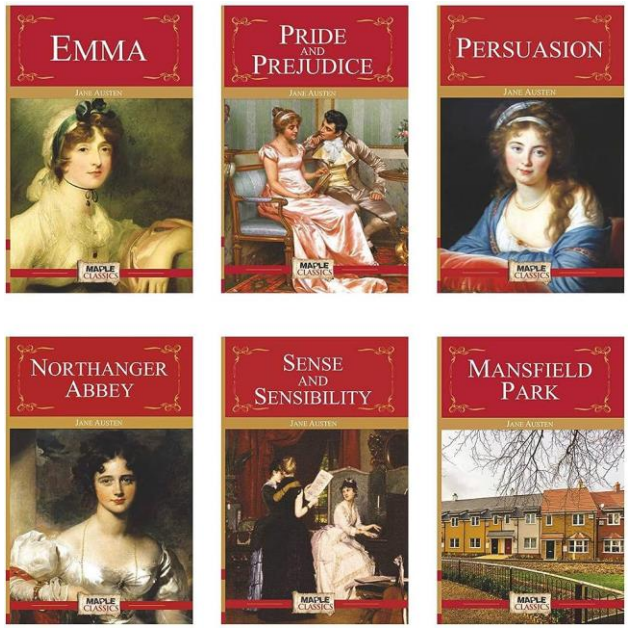
Indirect measuring from the Wikisphere

UNIVERSIDAD DE MURCIA

- **Why don’t we use the knowledge derived from Wikimedia Communities?**
- There are editions in any almost language, representing different cultural spaces.
- The extent and quality of Wikipedia could be considered as an indirect indicator of social and cultural aspects, for instance:
    - Civic compromise and participation
    - Access to education and information resources
    - Curation of own culture and interest in good public information sources.
- NPV (Neutral Point of View) is a well-known value in Wikimedia projects.
- Coverage in Wikipedia(s) is vast and diverse, both for current and past knowledge objects.
- “Ready-made” multilingual communities and content.
- Data-friendly ecosystem, powered by Knowledge Base Wikidata.

Works, not authors

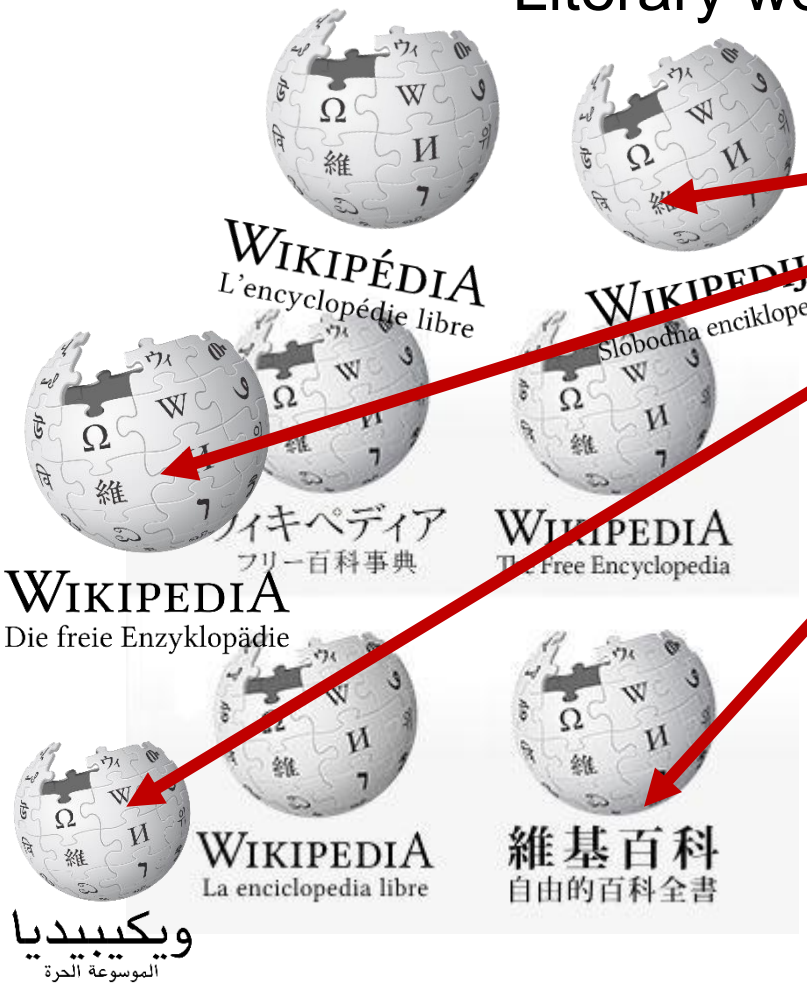
- There is **more consensus in a list of authors whose fingerprint in the literary history** is unquestionable.
- The author entity **gathers** the relevance of its many work, so it's an object more spread at Wikimedia and other specialized reference works and dictionaries.
- Identify which of their works is more valuated than the others in competition with the whole universal literature is a **challenge**.



HYPOTHESIS:
Canon candidates should also be relevant in Wikipedia(s)

UNIVERSIDAD DE **MURCIA**

A Wikidata item has to be **typed** as "Literary work" (Q7725634)



WIKIDATA

This literary work must have **its own Wikipedia** article in at least one language or edition.

Notability criteria for being included in Wikipedia, combined with enough **multicultural communities interest** to curate the encyclopaedic article, are the **two drivers** that suggests Wikipedia could be an alternative source of insights at a certain level of analysis.

How to measure: Low-Data approach

UNIVERSIDAD DE **MURCIA**

- Big data is not the unique valid approach when understanding things.
- Even the use of the editions of each and all Wikipedias and every single statement in Wikidata doesn't reach the level of Big Data *stricto sensu*.
- Several studies about Wikipedia content and communities use the social network metrics, analyzing the whole graph. **We don't**.
- How many data is enough to get insights with affordable resources (Data, Time, Code, Platforms)

Our focus

Selecting enough parameters to trade-off between sense and calculation processes.

Low-Cost from the early stages of the research

Wikipedia stands-in for a Global Catalog

UNIVERSIDAD DE MURCIA

- Wikipedia is not a catalogue of all the bibliographic production, but it replaces it in absence of world-scale library catalogues that accomplish with the Work-Expression-Manifestation-Item conceptual model of IFLA.
- Librarian holding databases is still far for resolve this cultural question about works.: Not Library of Congress, Not OCLC WorldCat, Not any catalogue
- More Linked Open Data from a big number of catalogs would be valuable.
- Online bookstores catalogs cover poorly this field.

Total Wikidata ítems	Wikidata items with link(s) to Wikipedia articles	Subset after first cut-off criteria
192.236	**107.434**	**40 % deprecated**
Books in Wikidata as a bibliographic source for Wikimedia Projects	Books that deserved the creation of an enciclopaedic article in one or many Wikipedias.	60% of Literary Works Items are candidates to play the “universal all-times literary canon game”

Which data from Wikipedia-Wikidata?

UNIVERSIDAD DE **MURCIA**

Only capturing **3 magnitudes** to characterizing every item selected.

Not considered other data as: numbers of views, structure of the texts, references, wiki links or data outside Wikimedia platforms.

NWikis	Number of Wikipedias with an article about the literary work	SPARQL in WDQS retrieving sitelinks
NWords	Total number of words used in all the articles in any language	XTools API to obtain numbers for each Wikipedia
NProp	Number of descriptive properties in Wikidata. Excluding identifiers	SPARQL in WDQS retrieving statements

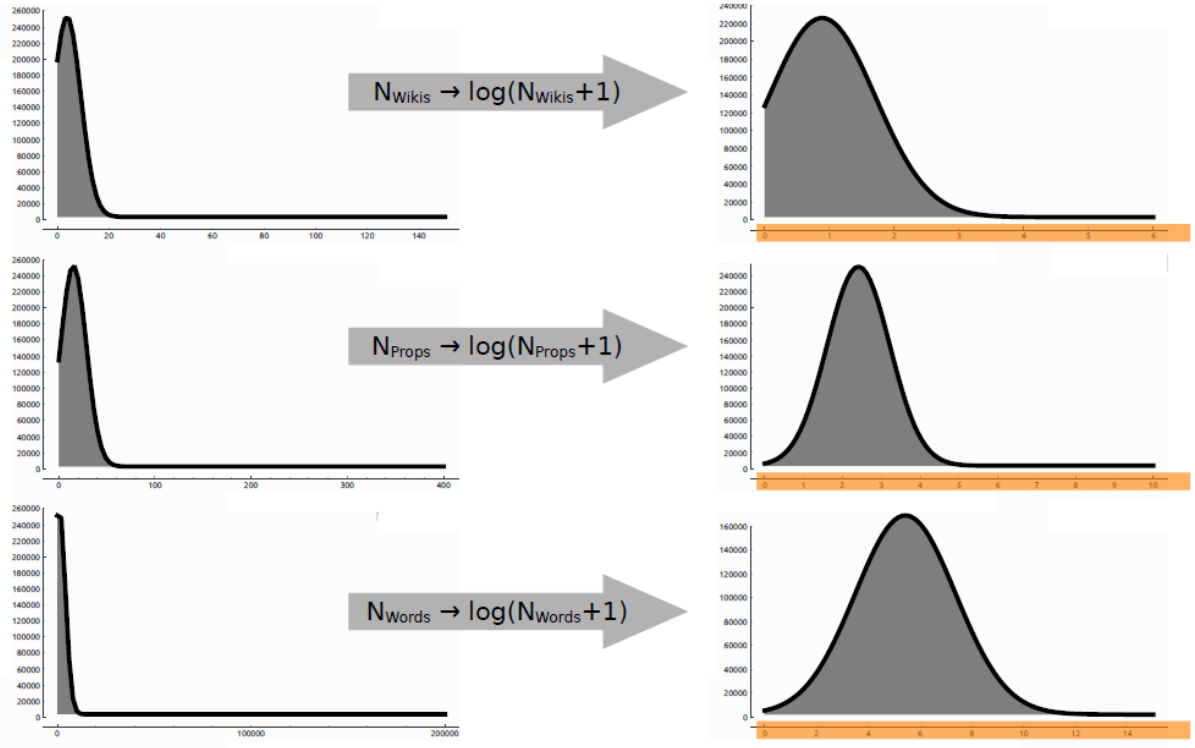
Statistic validation of results obtained with: Skewness, Kurtosis, Pearson and Spearman' correlation coefficients.

Also were obtained, for secondary uses publication date and original language.

Normalizing distributions

UNIVERSIDAD DE **MURCIA**

Logarithmic transformation

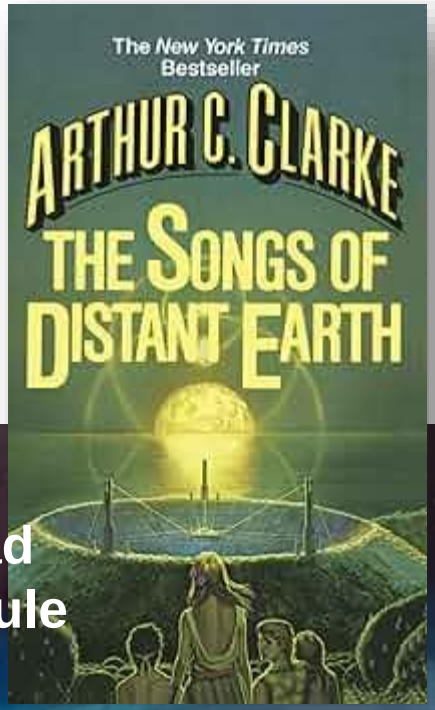


N_{Wikis}, N_{Props} and N_{Words} have a **pronounced positive skew**: *low number of sitelinks, poor redaction of articles and only a careles descriptions*. Logarithmic transformation allow operations with value intervals very different and to obtain a more normalized distribution when they are combined into the Wiki3DRank indicator.

How many books must be preserved?

UNIVERSIDAD DE **MURCIA**

- Once the dataset is constructed, the **Clustering K-means++** method is applied.
- **Silhoutte** test suggests the possibility of dividing ítems in 2 or 3 clusters.
    - 2 clusters seems too “Long-Tail”. Main cluster contains 1.008 books, and the secondary the vast rest.
    - 3 clusters seems like a more fine-grained approach, and it was the one we took.



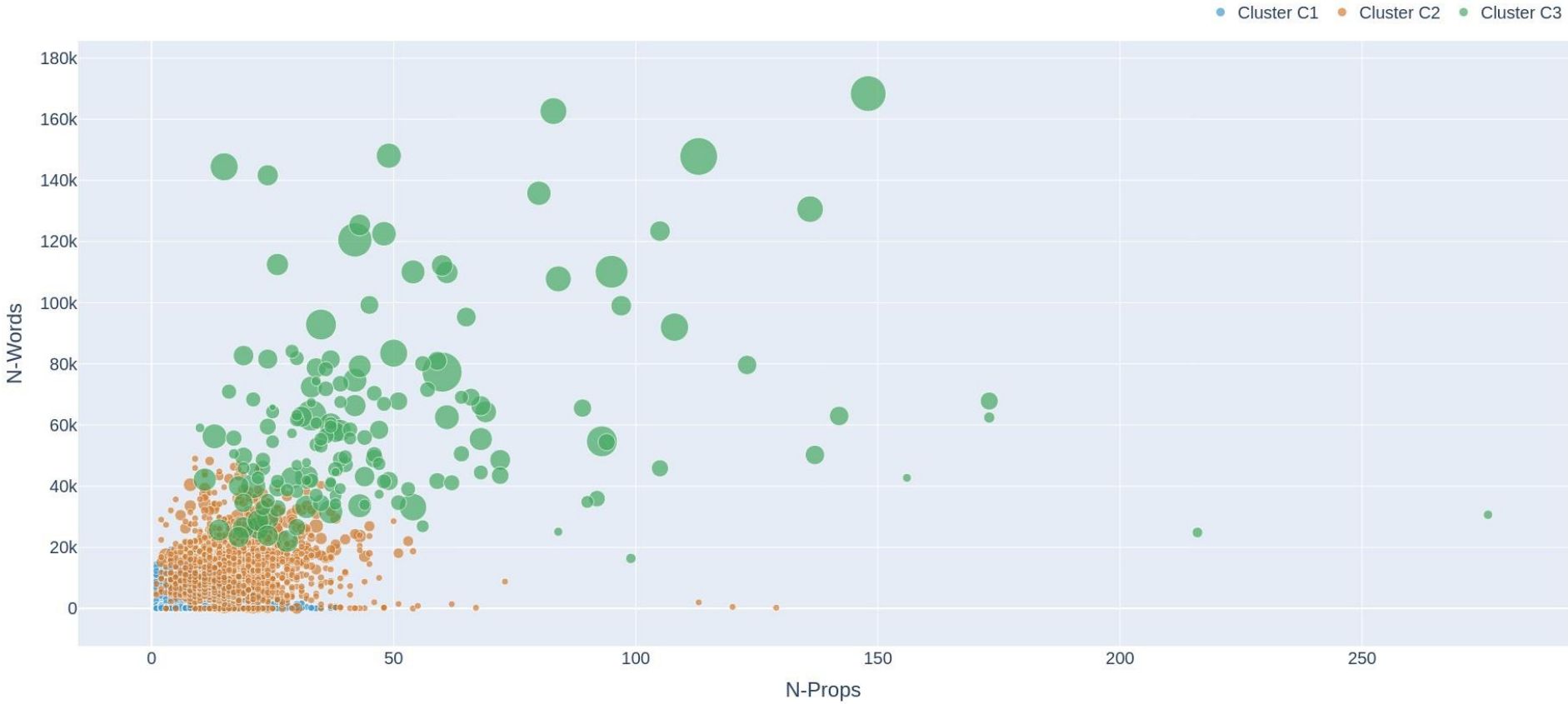
Now we might figure out the volume of books to load in the *preservation ark* in an hypothetic space capsule surviving in case planetarian terminal dissaster.

Three Clusters to group them all

UNIVERSIDAD DE **MURCIA**

Clusters show three well-differenced groups

- **C3 – Literary canon** = 163 items
- **C2 – Essential books** = 2.711 items
- **C1 - Literary or bibliographic production** = 105.000 items



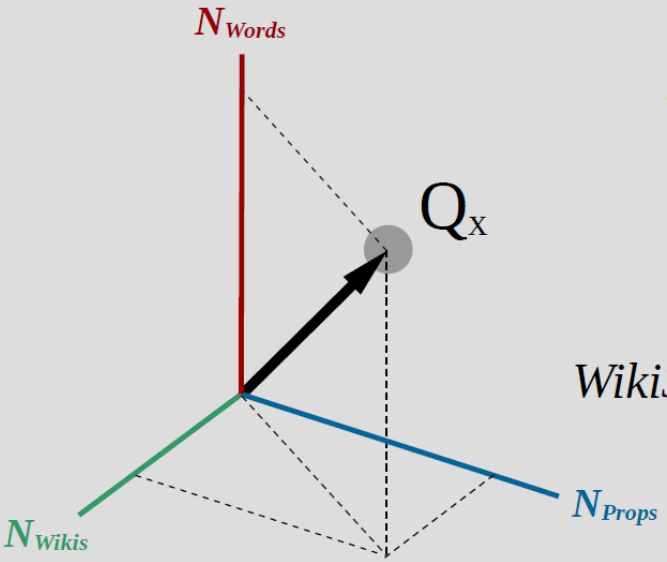
Synthetic Metric: Wiki3DRank

Inside each clusters each item has equal value.

Wiki3DRank is a single metric with a consistent stadistic normalized distribution.

First calcuted as the direct addition of the **logarithmic transformation** of each of the three metrics.

Finally, we used it as a **vector**, so the rank is **the module of a vector** with three dimensions.



$$a = \log\left(1 + N_{Wikis}\right)$$
$$b = \log\left(1 + N_{Words}\right)$$
$$c = \log\left(1 + N_{Props}\right)$$
$$\vec{V} = (a, b, c)$$

$$Wiki3DRank = |\vec{V}| = \sqrt{\left(a^2 + b^2 + c^2\right)}$$

Wiki3DRank as ranking metric

- Items inside the clusters are not ordered, so we propose a Rank metric.
- This metric works fine at the same level that the two clusters case, with a 93'9 – 92,7% of shared ítems (1008 or 163 outstanding ítems).
- Now we have established a quantitative frontier for the very best items (163)

Tabla III: Análisis estadístico de Wiki3DRank.

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

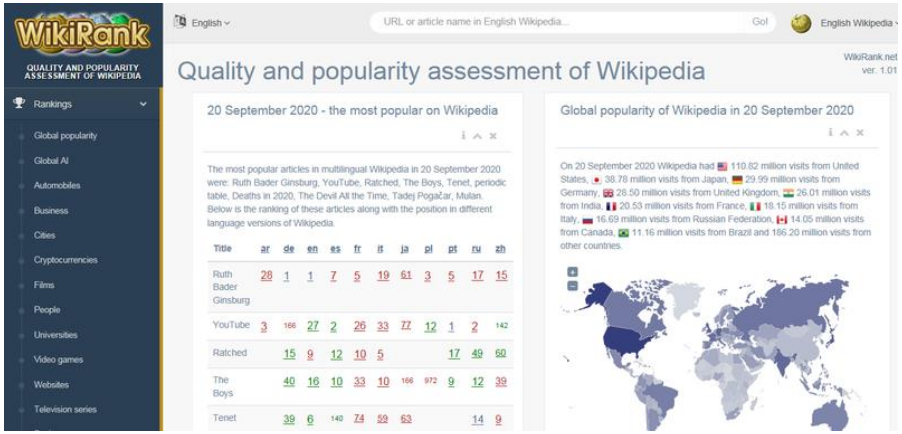
Tabla IV: Ratios de coincidencia para las diferentes iteraciones de K-means++ (entre paréntesis el número de ítems coincidentes). Fuente: elaboración propia.

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

What are we measuring?

UNIVERSIDAD DE **MURCIA**

- Our metric (Wiki3DRank) tries to capture **the value of the object described** in Wikipedia and Wikidata, not the articles theirselves.
- There are a huge **subfield of research** about quality aspects of articles, popularity, linking patterns or global quality of different editions of Wikipedia.
- **WikiRank** is another interesting project, that **measures and ranks the articles** by themselves, not the cultural object they represent. It also serves to compare quality of articles in various languages.

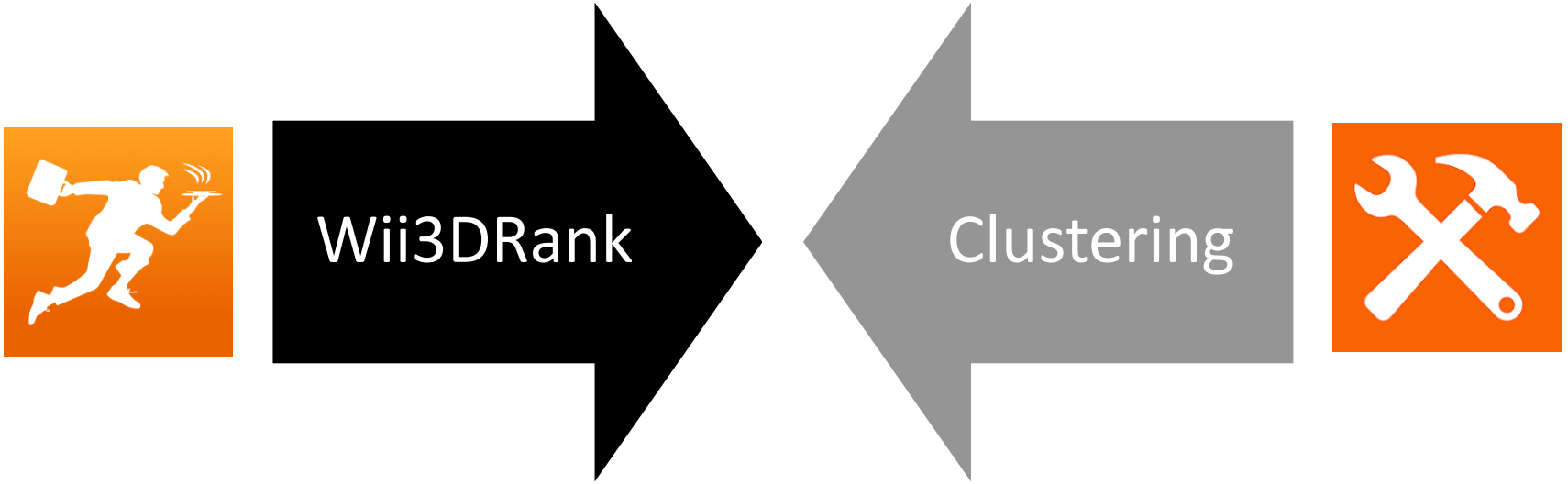


https://wikirank.net

- By the Wikipedia-Wikidata analysis we derive a measure of social global relevance of a precise type of cultural object.
- Focus is on the real object.

Wiki3DRank as ranking metric

The synthetic metric Wiki3DRank reproduces the results of the Clustering technique. Once tested, Clusters are no longer needed.



Obtained by querying a single item.

One at a time.

On-The-fly

Must be calculated retrieving the whole subset.

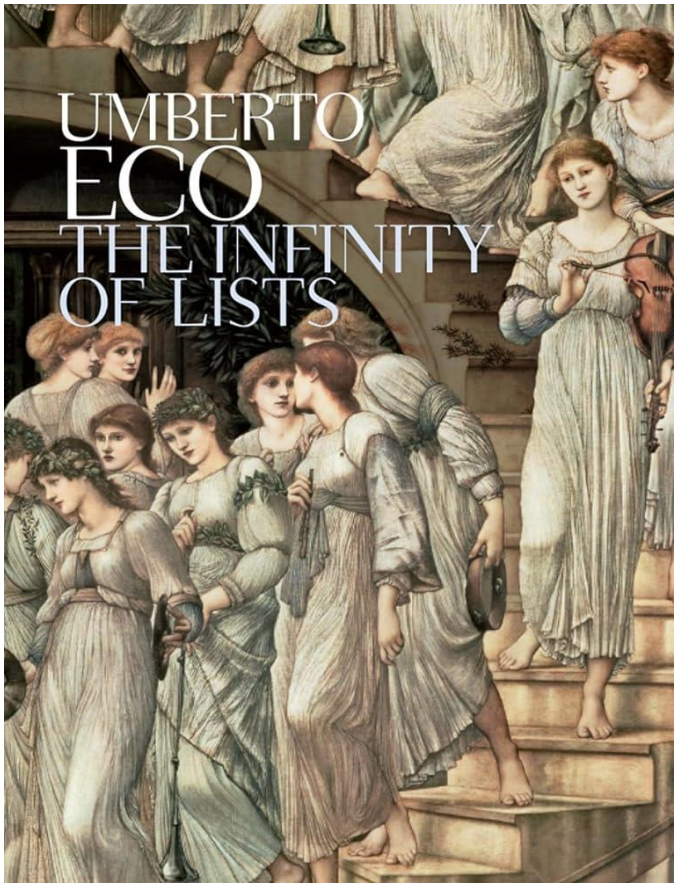
All items each time

Batch-processing

The list, here it is!

UNIVERSIDAD DE **MURCIA**

1. **Génesis**, story from The Bible
2. **Ilíada,** by Homer
3. **Hamlet**, by William Shakespeare
4. **Romeo and Juliet**, by William Shakespeare
5. **Don Quijote de la Mancha**, by Miguel de Cervantes
6. **Shahnameh**, persian epic poem
7. **Ulises**, by James Joyce
8. **Harry Potter and the philosopher's stone**, by J.K. Rowling
9. **Alice in Wonderland,** by Lewis Carroll
10. **Lolita**, by Vladimir Nabokov
11. **Macbeth**, by William Shakespare
12. **Pride and prejudice**, by Jane Austen
13. **Old Testament**, first division of the Christian biblical canon
14. **The hobbit**, by J. R. R. Tolkien
15. **One Thousand and One Nights**, collection of arabian folktales
16. **Dracula**, by Bram Stoker
17. **Exodus**, story from The Bible
18. **War and peace**, by Leon Tolstói
19. **1984**, by George Orwell
20. **Crime and punishment**, by Fiódor Dostoyevski



… and so on up to 163 ítems of the Main Cluster (C3).

Italian canon from a global point of view

UNIVERSIDAD DE MURCIA

- Once each selected ítem has its own rank, it's easy to filter by original language or date, to obtain an ordered list of N items.
- The table bellow presents the literary Works written in Italian.

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

Comparing ítems on-the-fly

UNIVERSIDAD DE **MURCIA**

- It's feasible to build a script to obtain the rank of an ítem (query by Q) on-the-fly, because it doesn't depend on the analysis and processing of the whole network of ítems (dump).
- https://gicd.inf.um.es/wiki3drank/index.php?id=Q29478,Q386431,Q82464

Faust (Q29478) by Johann Wolfgang von Goethe

Doctor Faustus (Q386431) by Thomas Mann

The Picture of Dorian Gray (Q82464) by Oscar Wilde

Wiki3DRank results

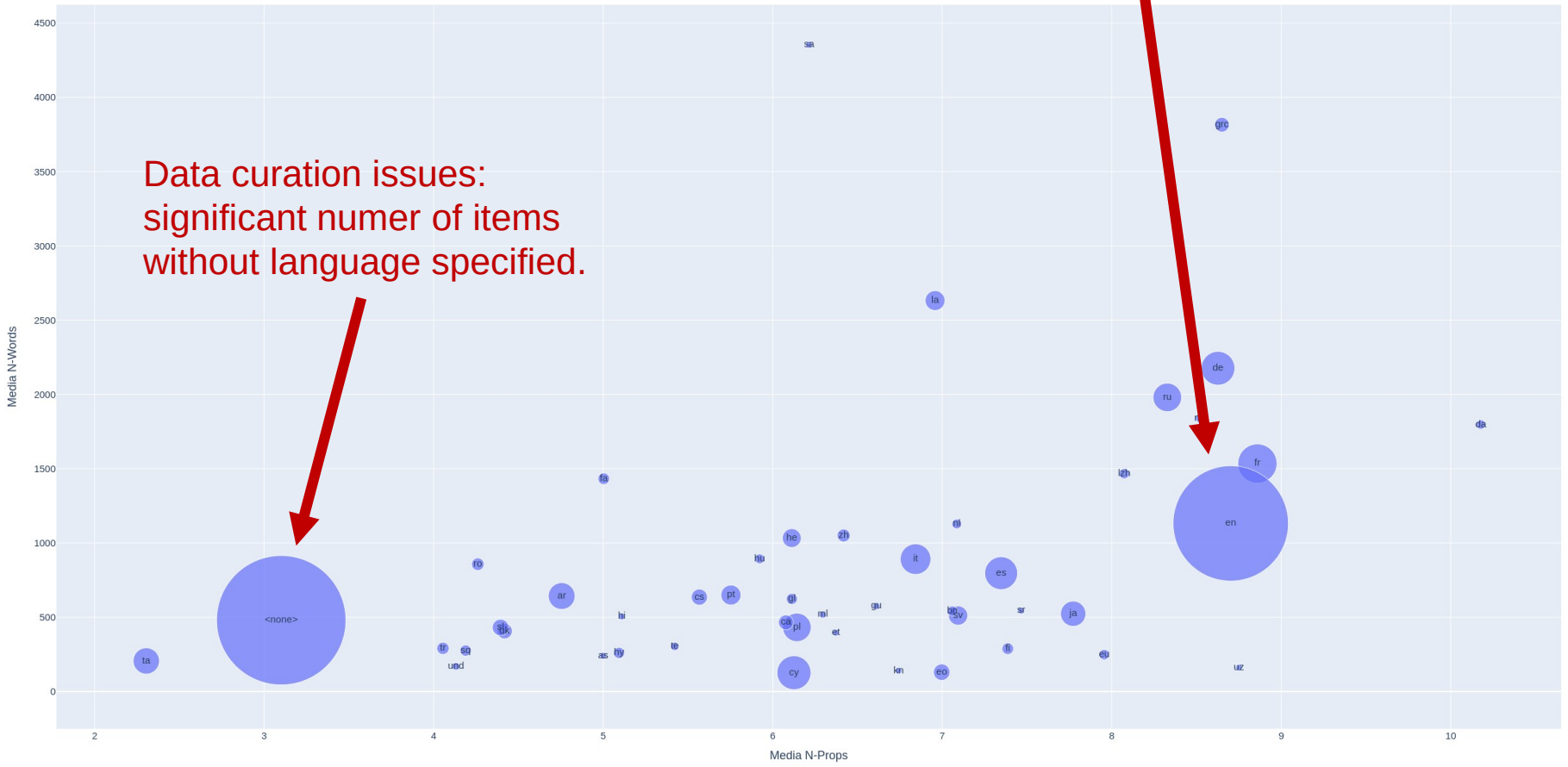
Ítem	Label	N_{Wikis}	N_{Props}	N_{Words}	Wiki3DRank
Q29478	Fausto	56	42	44169	12.037152949663502
Q386431	Doktor Faustus	27	29	30551	11.37200463461269
Q82464	El retrato de Dorian Gray	60	50	55750	12.320468385852626

Who do you think taht would win the contest of 80s best-seller?
"Das Parfum" by Patrick Süskind vs "The name of the Rose" by Umberto Eco

Linguistic and geographic diversity

UNIVERSIDAD DE **MURCIA**

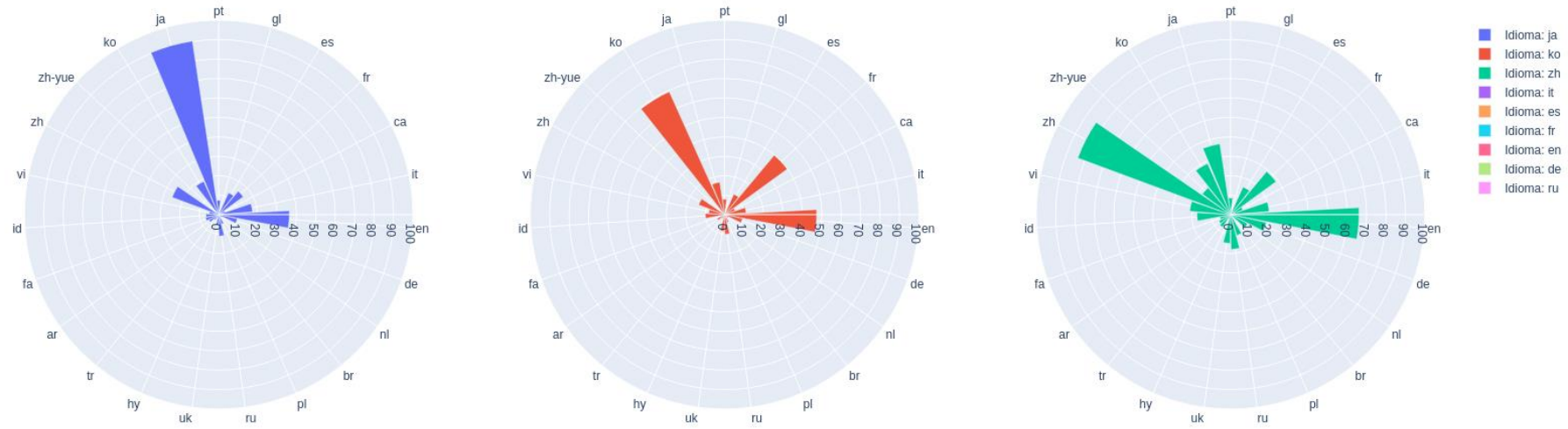
- English literature is enormous in number of articles, extent and enrichment of descriptive data.
- Presence of classical and also dead languages.



Language communities and foreing canon

UNIVERSIDAD DE **MURCIA**

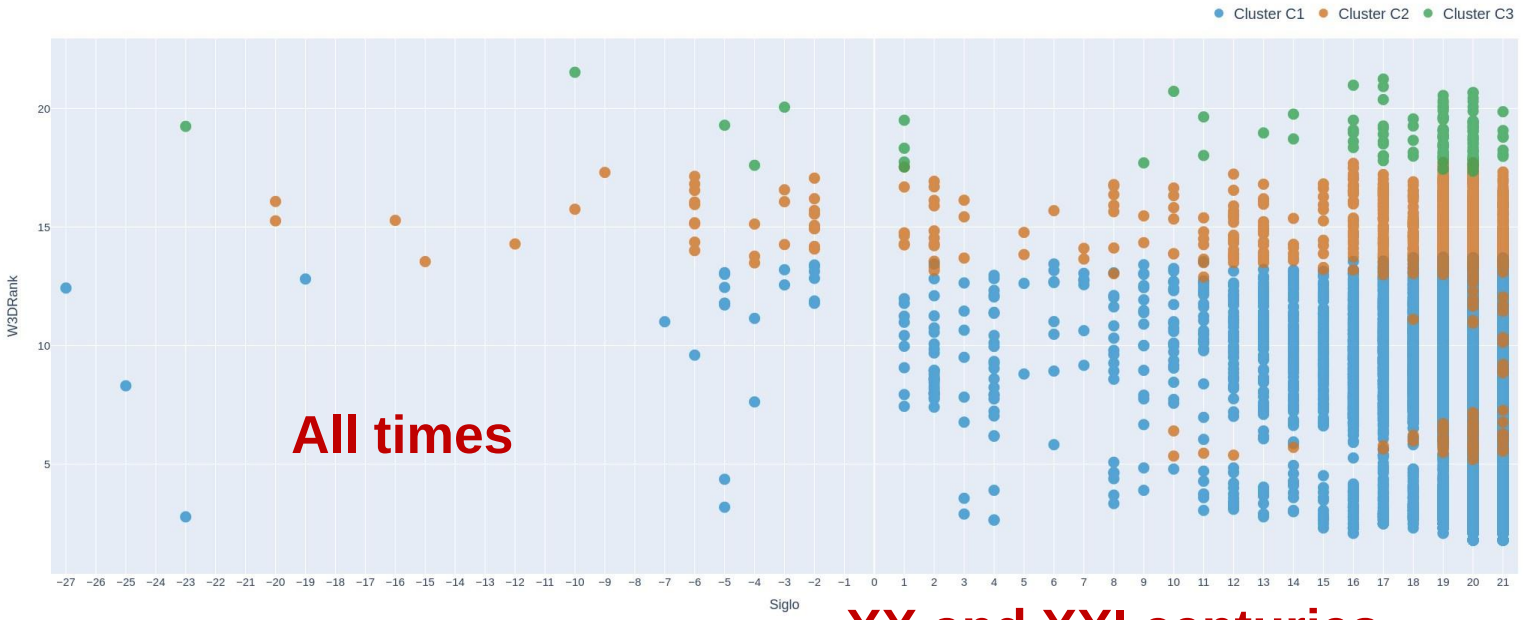
- Preliminary results for the research question: How much are the literary works – canon or not – in an specific language covered by other linguistic communities? **Literarure closeness**
- Coverage varies on the basis of the cluster (Prominence).
- Measures that take into account the rank of each works: higher rank books are gived more score.
- Is it possible to measure cultural literary proximities?

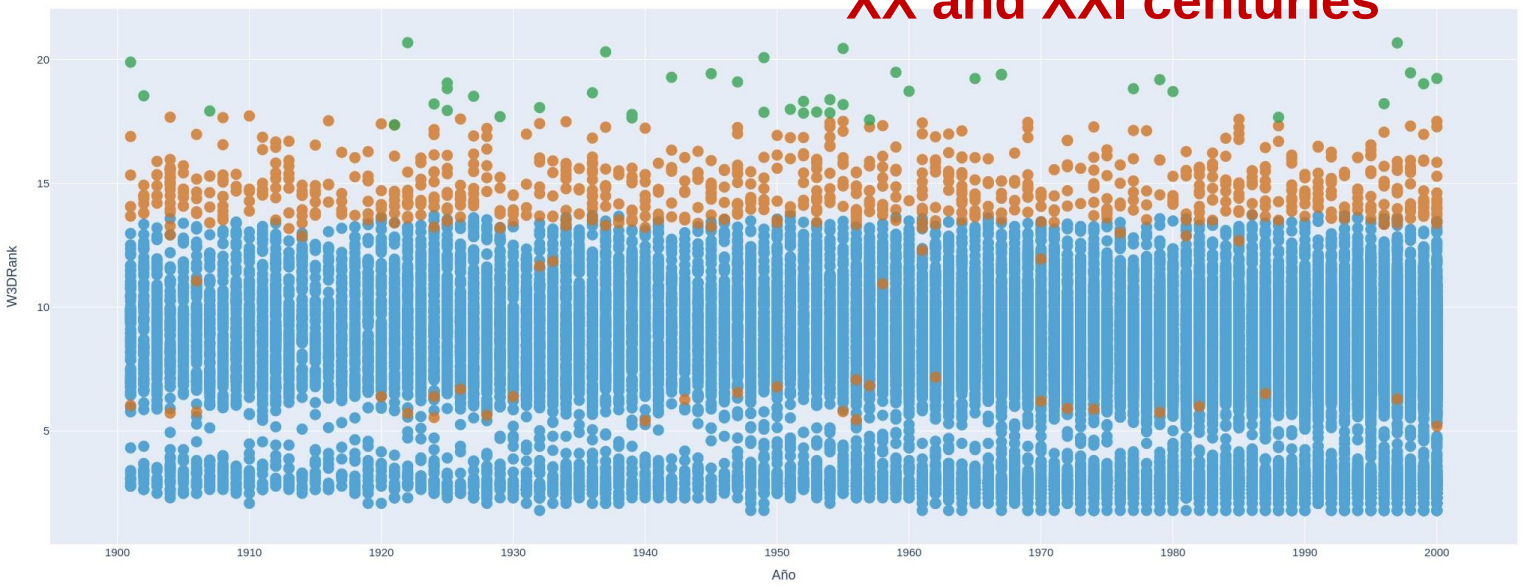


Japanese literature of JA-WIKI present in other Wikipedias: Besides EN-WIKI (a quasi-catch-all-wiki) Chinese and Korean Wikipedias are the most related, but also Spanish an Italian ones.

Times of rise and fall of languages

UNIVERSIDAD DE **MURCIA**

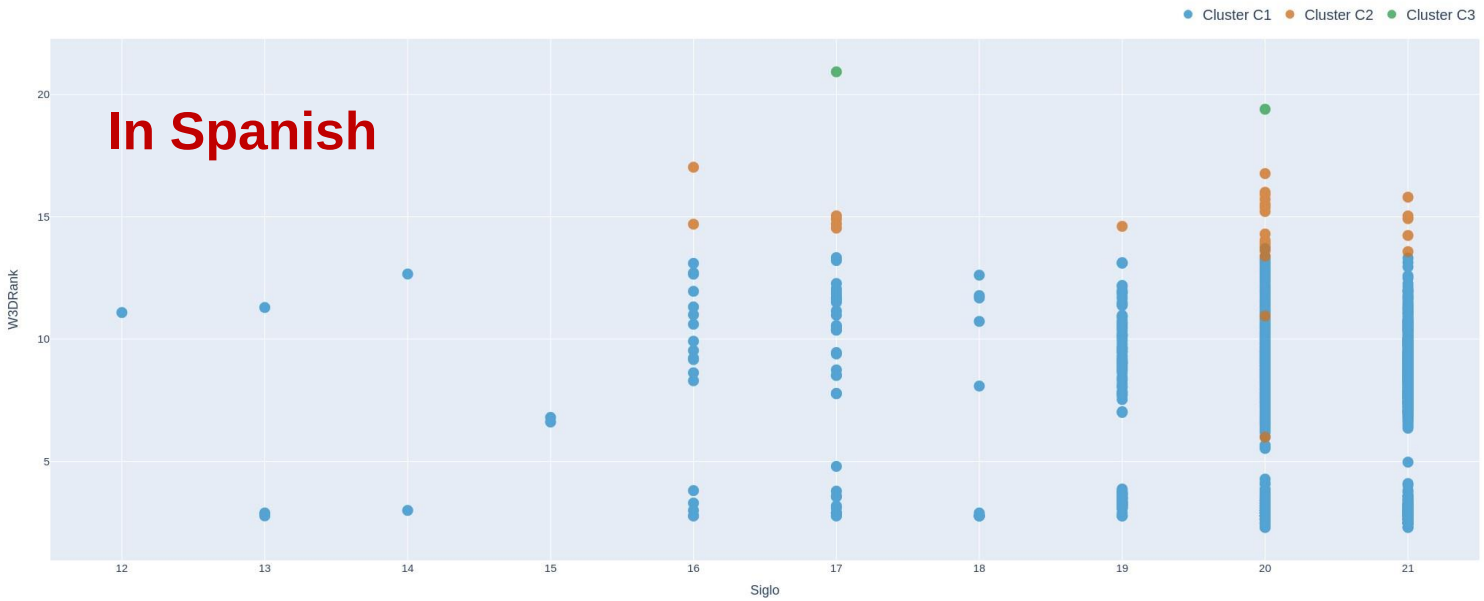


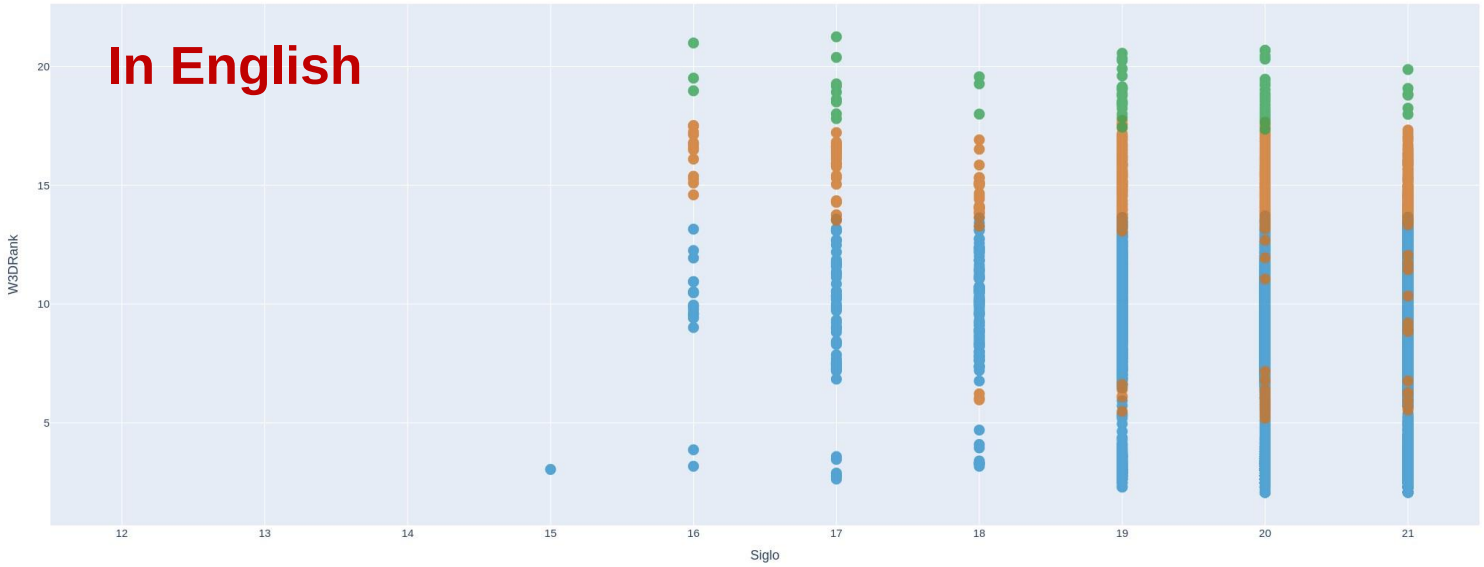


- Dark centuries in the **main clusters**.
- Literary tradition still resists in the **secondary cluster.**
- Modern age and Printing press era
- **Birth of national languages.**
- Reflexions about coverage, production, markets, empires, coloniamism, translations, globalization.
- Current Works compete with remarkable historical literary Works.
- Awareness of risk of present-bias in global literary markets

Times of rise and fall of languages

UNIVERSIDAD DE **MURCIA**





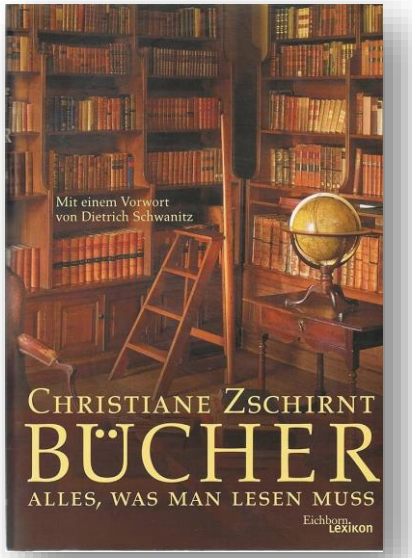
Comparing the presence of canon Works along centuries.

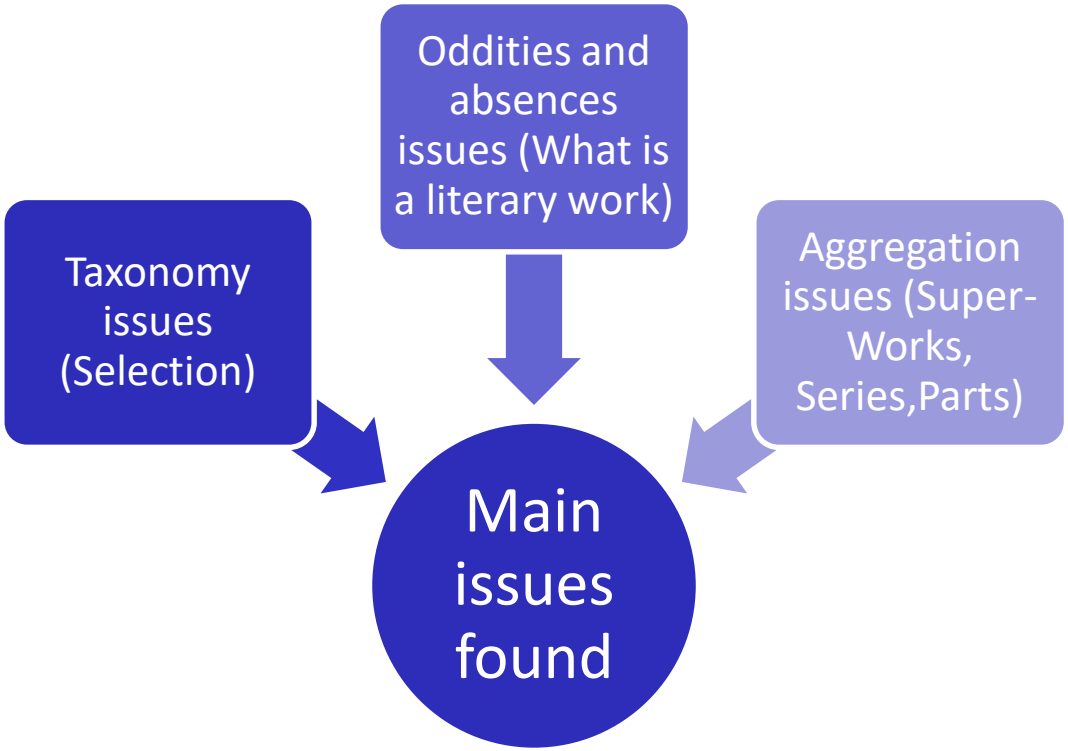
Here comes Shakespeare, here comes Dickens…

Reflexions about coverage, production, markets, empires, translations, coloniamism, globalization

Discussion and issues

- The results obtained are quite **common-sense**, but their provenance isn't intermediated by a single institution, scholar or country.
- Not so different from the mainstream selections made by **Christiane Zschirnt** or **Peter Boxall**.





Taxonomy issues

UNIVERSIDAD DE MURCIA

- Wikidata's collaborative taxonomy of classes is a mess.
- There are three main concepts that cover the vast majority of "literary Works", but they don't conform a coherent taxonomy and they also **overlap**.
- We estimate the total of "books" in Wikipedia(s) is between 370.000 - 401.158**,** that represents near a million articles.
- But to obtain a subset without **noise** and **silence** is complex. Overcome the 80/20 distribution is possible but complex, because of the oddities when assigning subclases and classifying ítems and curational issues.

Class (including subclasses)	Identifier in Wikidata	Numer of items
Literary work	Q7725634	244.929
Book	Q571	16.485
Written work*	Q47461344	139.744

Continuous wikidata editing and changes in criteria when assigning classess may produce high variability

Data as 7th october, 2023.

* Direct query, not recursive due to limitations of Wikidata Query Service

Recall and Precision

- The paper published uses the class “Literary work” (Q7725634) strictly, without including its subclasses, neither the other classes stated before.
- There are important literary Works that, at the moment of retreiving data, weren’t classyfied as “Literary work”; for example, **Odissey**, by Homer or **The Divine Comedy**. It can’t come into competition.
- If preliminary exploration of the Rank proposed validates, it will be neccessary to improve exhaustivity when selecting items.
- But, an overview of the resulting list, points at the presence of questionable ítems, that do not fit in the usual concepts of “literary work” or “literary value”.
    - Mein kampf (Q48244), by Adolf Hitler (60^{th})
    - Guinness World Records (Q41675) (114^{th})
    - The Wealth of Nations (Q233562), by Adam Smith (126^{th})
    - Chapters or parts of the Bible (1^{th}, 14^{th},18^{th}, 27^{th}, 43^{th}…)

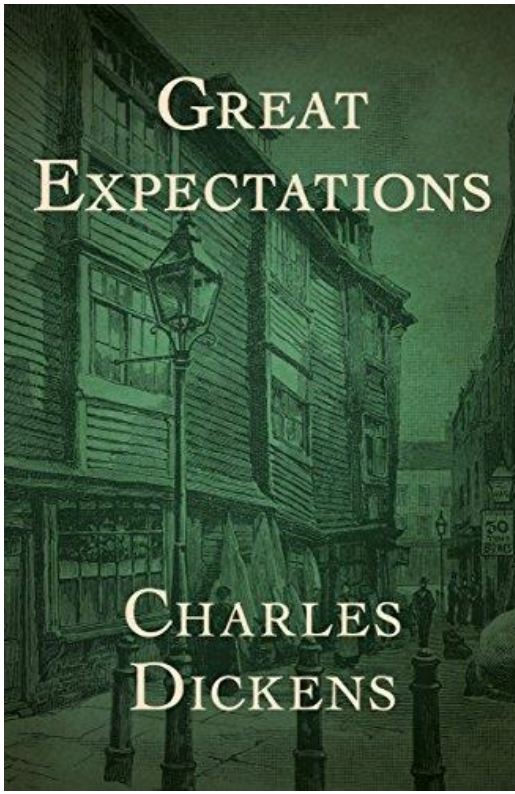
Aggregated items issues

- In Library and Bibliographic studies there is a concern about **what is really a work**.
- In the *Bibframe framework* or the *Library Reference Model* there is enough room for arbitrary decissions about what is a work or a expression or an instance.
    - **Compound books** as “The Bible”.
    - **Short stories** published that aren’t *per se* a book. Not only Borges’ tales, but also folk tales as “The Little Mermaid”.
- Books published in different ways, such as “The Lord of the Rings”, as a **single book or a trilogy**, where each part is an autonomous book.
- Books **series or sagas**, such as Harry Potter of Sherlock Holmes.
- There is a frustrating continuum between characters, fictional universes, and transmedia relations that difficults the automatic identification of a work for the purpose of this research.

Expectations of a good literary canon

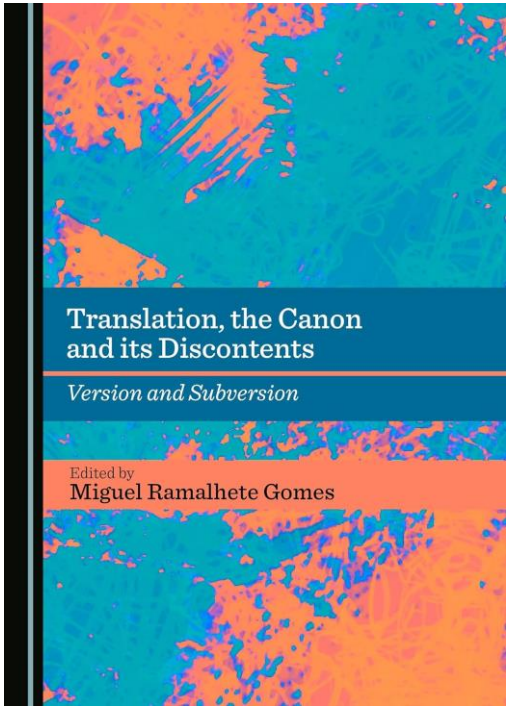
UNIVERSIDAD DE MURCIA

- Must include only “literary work”, based in a purposed previous definition(s).
- Must aggregate the value of books published in parts, series or sagas.
- Must capture a wide global point of view with cultural diversity.
- Also temporal diversity, balancing the long-term and the short-term.
- Works included should be relatively stable in time.
- It also should change timely; canon isn’t static but changes slow with short steps.



Final thoughts

- Almost every important cultural object has its own Wikipedia article in many languages.
- This research captures something related to **international diffusion** and translation history, that could be understood as *“translated canon”* useful in the field of World Literature Studies.
- The rank score of literary works of the same original language actually reflects the **attention** they receive from the **outside**; foreing cultures, other markets and other languages.
- That is to say; Our method adopts a globalized point of view, and can’t be used in regional literature without taking into account its **reception** outside it.

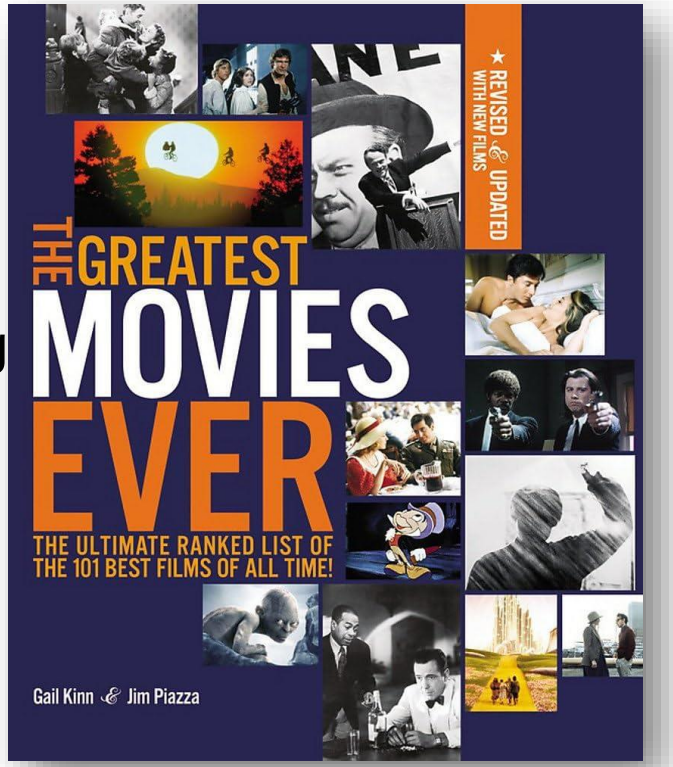


Generalize for other cultural sectors

UNIVERSIDAD DE MURCIA

Do the same parameters that works for books work for …?

- 20th century media
  - **Films**, TV Series
  - Recorded music, Songs, Albums
  - Comic, Videogames
- Ancient and all-time Cultural Objects throug the lens of the Creative Works practice
  - Fine Arts
  - Paintings
  - Monuments
  - Buildings
- Historial events?
- Fictional characters?



Movies: Preliminary Case of Use

UNIVERSIDAD DE MURCIA

Ejemplo de caso de uso: películas

Item	Título	Fecha	N_{Wikis}	N_{Props}	N_{Words}	Wiki3DRank
Q44578	Titanic	1997	109	228	178057	14,063594
Q17738	Star Wars: Episodio IV - Una nueva esperanza	1977	77	132	181040	13,7645652
Q163872	The Dark Knight	2008	81	143	162244	13,7128748
Q104123	Pulp Fiction	1994	79	169	149086	13,6923747
Q23781155	Vengadores: Endgame	2019	78	229	116645	13,593429
Q2875	Lo que el viento se llevó	1939	74	169	130155	13,5537014
Q102438	Harry Potter y la piedra filosofal	2001	81	209	112958	13,5417445
Q47703	El padrino	1972	97	146	118004	13,502344
Q23780914	Avengers: Infinity War	2018	77	154	123154	13,4834549
Q182218	The Avengers	2012	81	154	117324	13,4576116
Q91540	Back to the Future	1985	81	114	126913	13,4173687
Q18407657	Captain America: Civil War	2016	60	253	97908	13,4023385
Q14171368	Avengers: Age of Ultron	2015	71	191	100160	13,3609741
Q23780734	Black Panther	2018	62	168	111087	13,3589356
Q184843	Blade Runner	1982	65	106	130152	13,3444253

Item	Título	Fecha	N_{Wikis}	N_{Props}	N_{Words}	$Wiki3DRank_{CW}$
Q44578	Titanic	1997	109	228	178057	14,4360614447
Q17738	Star Wars: Episodio IV - Una nueva esperanza	1977	77	132	181040	14,2871183368
Q2875	Lo que el viento se llevó	1939	74	169	130155	14,259556799
Q104123	Pulp Fiction	1994	79	169	149086	14,100347707
Q47703	El padrino	1972	97	146	118004	14,0631627043
Q132689	Casablanca	1942	77	116	116815	14,0399754907
Q163872	The Dark Knight	2008	81	143	162244	13,9777134078
Q91540	Back to the Future	1985	81	114	126913	13,9017199263
Q102438	Harry Potter y la piedra filosofal	2001	81	209	112958	13,8900463361
Q103474	2001: A Space Odyssey	1968	79	116	107534	13,8623464046
Q184843	Blade Runner	1982	65	106	130152	13,8515090942
Q24815	Ciudadano Kane	1941	66	96	97473	13,7835263012
Q41483	Il buono, il brutto, il cattivo	1966	74	90	108352	13,7744669768
Q483941	La lista de Schindler	1993	77	159	98823	13,7325033616
Q134773	Forrest Gump	1994	89	241	74480	13,6954607692

Wiki3DRank captures someway multilingual diffusion (N_{wikis}), editorial effort and attention (N_{Words}) and depth of metadata description (N_{Props})

Realms for Creative Works suffer from certain degree of **presentism**. Media content with big launch campaign receive much more attention.

To weigh this effect it's possible to add an aditional component that **reflects how old a Creative Work is**.

Aged Works weight upwards in **$Wiki3dRank_{cw}$**. Althoug is necessary to fine tuning this component to avoid overvalue the very ancient Works.

$$d = \log\left(Año_{Actual} - Año_{Publicación} + 1\right)$$

$$\overrightarrow{V_{CW}} = (a, b, c, d)$$

$$Wiki3DRank_{CW} = |\overrightarrow{V_{CW}}| = \sqrt{(a^2 + b^2 + c^2 + d^2)}$$

Further research

UNIVERSIDAD DE **MURCIA**

- Context parameters: meaningful creative work properties: date, editions, …
- Include inbound and outbound links and relations data.
- Rank creative works in a broad sense.
- Generalize the study for any other cultural and knowledge objects.

And related research paths:

- Ongoing project research on fiction connection and recommendation (Discovery).
- Formal describing of fictional universe content for analysis and discovery (Metadata).