Wiki3DRank: A model for measuring the relevance of knowledge objects using quantitative data from Wikidata and Wikipedia

Wiki3DRank: un modelo para medir la relevancia de objetos de conocimiento mediante datos cuanti-tativos de Wikidata y Wikipedia

Juan Antonio PASTOR-SÁNCHEZ (1), Tomás SAORÍN (1), María-José BAÑOS MORENO (2)

(1) Department of Information Studies, University of Murcia (2) Odilo / Department of Information Studies, University of Murcia, {pastor|tsp|mbm41963}@um.es

Resumen

Se presenta el modelo Wiki3DRank, que combina datos cuantitativos extraídos en tiempo real de Wikidata y Wikipedia para obtener un ranking de objetos de conocimiento a través de un valor cuantitativo que mida la relevancia de un objeto frente a otros en un determinado dominio. El modelo se basa en la distribución de los objetos de conocimiento en un espacio vectorial cuyas componentes se basan en tres variables principales: número de declaraciones en Wikidata sobre un ítem, número de artículos en las diferentes ediciones de Wikipedia y extensión en número de palabras de dichos artículos. Estas variables se asocian al nivel de descripción de los ítems de Wikidata, la difusión de los objetos de conocimiento asociados a los mismos en las ediciones de Wikipedia de diferentes idiomas y el grado de elaboración editorial de los correspondientes artículos de Wikipedia. Para demostrar la viabilidad del modelo se analizan una serie de casos de uso sobre diversos dominios: libros, películas, catedrales, terremotos, ríos y elementos químicos. A partir de los resultados obtenidos es posible concluir que Wiki3DRank es una herramienta que permite medir la relevancia de objetos de conocimientos en el contexto de un dominio de conocimiento. Se muestra el funcionamiento de una herramienta de código abierto que permite el cálculo en línea de Wiki3DRank. Los resultados obtenidos sugieren que el modelo propuesto puede aplicarse para diferentes contextos y dominios, que pueden introducirse elementos de ponderación y es posible extender el modelo mediante la introducción de nuevos componentes basados en otras características de los datos enciclopédicos de los objetos de conocimiento, al mismo tiempo que se mantiene el sistema de cálculo vectorial de base.

Palabras clave: Wiki3DRank. Rankings. Wikidata. Wikipedia. Conocimiento enciclopédico. Análisis de dominios. Objetos culturales.

Abstract

This research introduces the Wiki3DRank, a model combining real-time extracted quantitative data from Wikidata and Wikipedia to obtain a ranking of knowledge objects through a quantitative value that measures the relevance of one object compared to others in a specific domain. The model is based on the distribution of knowledge objects in a vector space, whose components are based on three main variables: the number of statements on Wikidata about an item, the number of articles in different Wikipedia editions, and the length in number of words of these articles. These variables are associated with the level of description of the Wikidata items, the dissemination of the referred knowledge objects in Wikipedia editions in different languages, and the degree of editorial elaboration of the corresponding Wikipedia articles. To demonstrate the viability of the model, a series of use cases across various domains are analysed: books, movies, cathedrals, earthquakes, rivers, and chemical elements. From the results obtained, it is possible to conclude that Wiki3DRank is a tool that allows measure the relevance of knowledge objects in the context of a knowledge domain. The operation of an open-source tool that enables the online calculation of Wiki3DRank is presented. The results suggest that the proposed model can be applied to different contexts and domains and that it`s ease to expand it by adding elements of weighting and extending the model with new components based on other characteristics of the encyclopaedic data of the knowledge objects, while the base vector calculation system is maintained.

Keywords: Wiki3DRank. Rankings. Wikidata. Wikipedia. Encyclopaedic knowledge. Domain analysis. Cultural objects.

1. Introduction

This work proposes a method for calculating a ranking applicable to knowledge objects recorded in Wikidata and Wikipedia. Rating, reviewing, and commenting are social phenomena in themselves that are part of public discourse and are manifested through both classic and current media (Black, 2007). The culture of ranking has a long tradition, but its contemporary form has a significant impact because of a progressive process of quantifying social interactions. Examples include best-seller lists, essential items, top

tennis players in the circuit, cities with the best quality of life, and so on. Ratings and scores coexist with critiques, reviews, and cultural studies. All of these act as a kind of journalistic genre that serves as a pattern for promoting or accessing culture in a very broad sense. Within the framework of publishing industries, audiovisual content, and entertainment, the importance of appearing in lists, rankings, and selections is very pronounced. The aim of this article is to explore the generalization of the ideas presented in a previous article limited to literary works (Pastor-Sánchez; Saorín; Baños-Moreno, 2023), focusing now on what we call knowledge objects. These can preliminarily be defined as those entities of any kind that gain enough notoriety to merit an article in any of the editions of Wikipedia. The concept of a 'Wikipedia article' entails a difficulty, as it is actually composed of a variable number of articles in many languages about the same element, event, or concept. The articles from different editions of Wikipedia correspond to a knowledge object that gathers both information and factual data relevant for its explanation and understanding. In the first case (information), it would be a merger of all the written information about an object in all the languages of Wikipedia, and in the second case (factual data), the data synthesized in Wikidata for that individualized element.

There is a large bibliography focused on analyzing the quality of encyclopedias and articles, based on multiple factors related to the collaborative editing effort on articles (Moás; Teixeira Lopes, 2023). Often, aspects of audience interest generated by each article are also incorporated. On the contrary, approaches based on ratings and external evaluations are scarce. One of the most common methodologies is network analysis, which is a long-term trend in the renewal of research in humanities and the cultural field, a phenomenon termed as 'Network turn' (Ahnert; Ahnert; Coleman, and Weingart, 2020). However, given that there are various Wikipedias for each language, each constituting its own graph, their analysis becomes problematic from the perspective of universal knowledge objects.

Wikipedia covers all topics and serves as a cartography of the current state of knowledge: it is a map of concepts continuously enriched. Therefore, it offers a unique entry point to the inquiry about the ranking of how these objects are treated at the informational level. Studies on thematic coverage in Wikipedia have revolved around various fields, such as science, biographies, cultural heritage, mass culture, or social current affairs (Hill and Shaw, 2020; Reznik and Shatalov, 2016; Minguillón et al., 2017). Wikipedia competes with a number of specialized information sources in each field, such as film databases, music history repertoires, or library catalogs. The textual discourse of Wikipedia has been strengthened since the launch in 2012 of Wikidata. This is an initiative that provides infrastructure for storing structured data derived from the content of articles from any different editions of Wikipedia.

Due to the current society's interest in rankings, it is common to find lists based on quantifiable intrinsic properties: the length of rivers, book sales, the wealth of billionaires, the population of cities, or the weight of chemical elements. There are also multiple rankings on socially interesting topics such as books, movies, politicians, athletes, or events. These kinds of lists are periodically updated by the media, and even constitute a global editorial series of the '1001 ... that you must ... before you die.' Rankings that attempt to measure the importance of elements in areas such as cathedrals, naval battles, archaeological sites, World Cups, popes of the Catholic Church, or aromatic plants are less common. However, Wikidata can be understood as a system of knowledge objects and an information system whose structural characteristics allow access to its data. Therefore, processing this data could allow the generalization of a ranking calculation method, which in this work is called *Wiki3DRank*, focused on defining indicators that are simple to calculate and explain.

Many proposals have been made for the automatic evaluation of aspects of Wikipedia content quality based on quantitative methods. These proposals constitute a subfield of study on Wikipedia in their own right (Nielsen, 2019). Some authors exploit network analysis metrics, others use the metrics available for the content of the articles themselves: number of words, number of references, length, incoming links, etc., complemented by the study of editor activity, reputation, and collaboration networks. Similarly, in Wikidata, research is conducted to establish the quality and completeness of data (Shenoy, 2022). This is a field that also generates applied research. The most well-known work on ranking is by Skiena and Ward (2013), in which they compare historical figures differentiating between celebrity (current popularity) and gravitas (established popularity). Similarly, the Networked Pantheon database applies centrality measures and the like in the resulting Wikipedia biography graph (Beytía; Schobin, 2020).

A relevant similar research project to the research presented here, is WikiRank (Lewoniewski et al., 2019). Currently, it maintains an online service (www.wikirank.net) that allows obtaining

multilingual article rankings. This service is an example of how to define article rankings segmented by content types using aggregate indicators they call 'popularity', 'Authors' Interest', and 'Citation Index'. It is based on the periodic processing of Wikipedia dumps and also allows access to historical data on ranking evolution.

It is possible to observe a specific version of Wikipedia and also by thematic categories to make comparisons between articles within thematic groupings obtained from the exploration of categories, Wikidata classes, and the Dbpedia ontology. Each article receives a value from 1 to 100, based on the analysis of the most frequent characteristics used in studies on automatic quality assessment in Wikipedia: completeness, credibility, objectivity, readability, relevance, style, timeliness. Data such as article length, number of references, reference density, number of images, and number of sections are taken from the Wikimedia XTools tool, and a synthetic measure is obtained from normalized values for each characteristic and adapted to each version of Wikipedia. The maximum score (100) is assigned in those characteristics where an article exceeds the median of the corresponding language. Subsequently, an arithmetic mean of the evaluated characteristics is made and finally modulated bearing in mind the existence of quality control templates in the article. Finally, quality is measured combining all these components. WikiRank also measures popularity, with synthetic measures on page views and number of editors. Each thematic block shows the most popular articles, their coverage in the analyzed languages, and which version has the highest quality. From each article, it is possible to get a view of how its popularity evolves over time, globally and within each Wikipedia. It is also possible to obtain indicators of the number of links of each article within its own Wikipedia, and a cumulative global value (Citation Index).

There is a certain likeness between the names *WikiRank* and the *Wiki3DRank* model proposed in this work. However, it is necessary to highlight that *Wiki3DRank* focuses on the use of a synthetic measure of encyclopedic knowledge objects. This measure is based both on the analysis of characteristics of the corresponding Wikidata item and its correspondent articles in Wikipedia. In addition, a conceptual model is adopted based on the representation of objects within a vector space whose dimensionality, as discussed in the discussion section, can be adapted to different analysis scenarios.

Therefore, this work defines a series of objectives and a working methodology to determine the necessary data and the way in which they should be obtained, processed, and used to obtain a measure, *Wiki3DRank*, that allows identifying and weighing multidomain knowledge objects derived from the combined use of Wikidata and Wikipedia.

2. Knowledge objects: from Wikipedia articles to Wikidata items

Encyclopedias have a long tradition in response to the need to bound the basic knowledge available at a given time and present it in an accessible format (Brown, 2011): compact, oriented towards the precise explanation of the various aspects of a concept: emergence, evolution, applications, connections. It is interesting to consider Wikipedia from two important viewpoints for our purpose, its thematic coverage and content extension.

From the point of view of its coverage, Wikipedia has achieved a breadth of topics never seen before. It also stands out for its fast response capacity to incorporate information about new events. Its growth is continuous because reality itself generates new data and entities deserving attention. Furthermore, the combination of its digital format and distributed editorial policy has facilitated 'inclusionism', which greatly expands the range of what is admitted as encyclopedic relevance or notability (McDowell; Vetter, 2022, pp. 46-70). In a digital context and with a large mass of editors, it is possible to assume articles on many more topics. At the same time, it allows for a high level of granularity, since each specific part of a topic can be addressed in its own article.

From the point of view of content extension, Wikipedia articles show a closer resemblance to specialized encyclopedia articles than to generic ones. This is because the articles tend to reach a considerable length, are divided into sections, include notes, and are densely connected with other concepts in the encyclopedia. Although the encyclopedic ideal is to present a topic sufficiently, the elasticity of the digital page allows editors to add relevant information to offer a broad overview of the subject of interest, from its different angles. The requirement for verifiability means that articles also contain a basic bibliography for orientation on each topic, as well as notes with references to specialized publications.

Many studies on Wikipedia focus on the English version, assuming that, as it contains the largest number of articles, it reflects almost all global knowledge. However, it is necessary to remember that each Wikipedia is an independent project, and a vast number of articles on topics not covered in English have been identified (Miquel-Ribé, 2019). Of course, there is a significant degree of overlap in articles from different Wikipedia

editions on the same subject, theme, or concept. But the differences due to lack of coverage between languages are also important. The launch of Wikidata as a factual database connecting all Wikipedias highlights the existence of a global concept map resulting from the sum of all encyclopedias, regardless of the language of origin.

A third aspect appears here, in addition to those already mentioned related to the convergence and extension of topics, which is the Wikidata-Wikipedia artifact as a knowledge organization system. Understanding this system increasingly requires a more agile use of domain analysis techniques (Smiraglia, 2015). An encyclopedia, especially Wikipedia, is both a scientific-cultural vocabulary and an inventory of authoritative names for individuals and groups, as well as an enormous attention devoted to the recording of social events. The native categorization in Wikipedia is a haphazard mix of browsing, description, and grouping, but it is not a taxonomy well suited for exploring a domain of knowledge. However, while the classification system in Wikidata adheres more closely to the standard criteria for a correct taxonomy, it exhibits many inconsistencies in its hierarchical structure and class assignments.

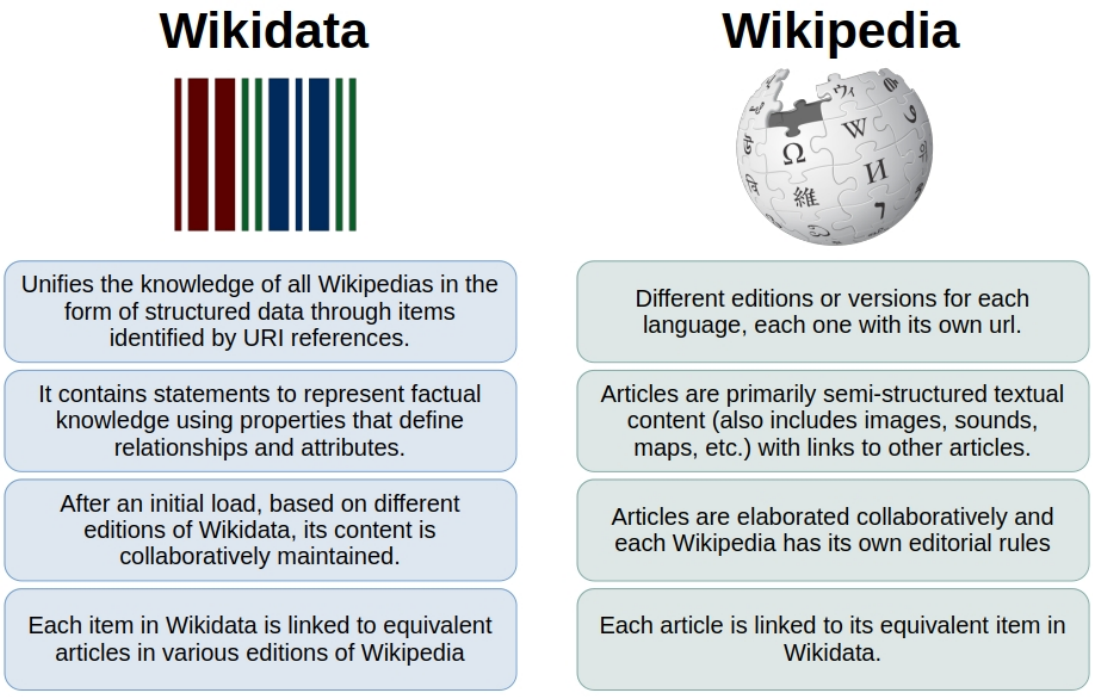

Figure 1. Summary of the main differences and the connection between Wikidata and Wikipedia editions

The concept of 'knowledge object' shaped in this article is clearly linked to the concept of an item in Wikidata. Articles arise in a certain language, and a unique element identifier is assigned to them in Wikidata, which will be used to link articles that arise in other languages with this previous item code. An item, for example Q63167656 for the article about the *Notre-Dame Cathedral fire in 2019*, connects articles in 58 Wikipedias, thus building an individualized entity for a relevant concept, in this case, of the type event.

Wikidata is a knowledge graph that uses its own data model compatible with RDF. Its main elements are items that represent a real object, a concept, or an event. Each item is associated with a unique identifier starting with the letter 'Q'. For example, the book 'One Hundred Years of Solitude' by Gabriel García Márquez is the item Q178869, though it is linked to 74 articles in different Wikipedias (Spanish, Japanese, Italian, Russian, etc.). In turn, each item is described by properties whose designations begin with the letter 'P'. The properties define relationships between elements or refer to literal values (strings, numbers, dates). For instance, for the aforementioned book, it is declared that its author (P51) is item Q5878 (the writer García Márquez) and its publication date (P577) is 1967. Wikidata does not have explicitly differentiated classes from the rest of the elements. Instead, some elements act as classes within a taxonomy of classes and subclasses connected through the property P279 (subclass of). The membership of items to classes is achieved through the property P31 (instance of). That is, Wikidata can be understood

to some extent as a 'collaborative ontology' containing both primary data and the schema used to formalize the organization of knowledge (Piscopo and Simperl, 2018). Within each item, there is a section called 'Identifiers', which establish connections with records and databases of all kinds, such as with the international authority control system VIAF (Bianchini and Sardo, 2022). Not all Wikidata items have an article in some edition of Wikipedia, as they can be created as data and may not generate enough interest to require an explanatory article. Although there are elements without an article, this work focuses only on those that do, thus establishing a limit on knowledge objects that have achieved encyclopedic relevance.

A descriptive record in Wikidata may include the following types of information:

- *Labels, aliases, and descriptions:* texts in various languages that allow naming items, aliases (synonyms), and descriptions with abbreviated definitions. Not all items have labels and descriptions in the same languages.
- *Properties or attributes:* the descriptor for a value (either literal or another item) of a claim or statement. They have an identifier that begins with P (e.g., P23). Some properties allow defining instance-class relationships (P31 instance of) or taxonomies with subclass-class schemes (P279 subclass of).
- *Statements:* data about a specific item. Formed by a claim with its corresponding qualifiers, references, and ranks.
- *Claims:* data for a specific property about a particular item, generally in the form of a link to another Wikidata entity. Formed by property-value pairs.

Statements, in turn, can be specified through:

- *Qualifiers:* a claim that says something about a specific statement to nuance or detail it (reification). Also formed by property-value pairs.
- *References:* describe the source of a claim and can be an external link or another Wikidata item.
- *Rank:* an indicator that identifies the most relevant statement compared to others when there are several on the same property.

The underlying idea of this work's proposal is that Wikipedia can be considered both discourse and data. The Wikipedia articles distributed in each of the different language editions offer diverse perspectives on knowledge. The data, multilingual and gathered in Wikidata, combines all editions of Wikipedia and shapes a global inventory of facts or concepts.

3. Wiki3DRank calculation methodology

The aim of this work is to propose a method for calculating the ranking of knowledge objects from Wikipedia and Wikidata, named *Wiki3DRank*. This metric uses the contents of Wikipedia or Wikidata as data sources to measure something related to the overall attention on the knowledge of an external object; therefore, it does not attempt to measure the quality of Wikipedia articles or Wikidata descriptions.

Its operation is largely based on previous work (Pastor-Sánchez; Saorín; Baños-Moreno, 2023), although it represents a refinement from the standpoint of its conceptual definition. In the mentioned work, a literary canon was defined based on the number of statements of its Wikidata item, the number of Wikipedia editions in which the item had an article, and the sum of words in those articles, complemented with clustering calculations to delineate coherent subsets.

For the calculation of *Wiki3DRank*, three fundamental indicators are defined, namely:

- N_{Props} : The number of properties used to describe a Wikidata item, excluding those used in the identifiers section. This indicator reflects the depth and breadth in the process of describing an item.
- N_{Wikis} : The number of Wikipedia editions in which a Wikidata item has a corresponding article. This indicator represents the global reach that a particular item has within the context of various languages.
- N_{Words} : It is calculated as the sum of the number of words in the content of all the articles from the different editions of Wikipedia linked to the Wikidata item. This indicator measures the volume of editorial work carried out in the process of writing the articles.

It is necessary to consider that the numerical ranges in which these indicators vary are very disparate. The magnitudes measured by N_{Props}, N_{Wikis} and N_{Words} are of different natures. For example, the values that N_{Words} can reach by summing all the words of equivalent articles of an item in all editions of Wikipedia are considerably higher than those that N_{Props} and N_{Wikis} can reach. On the other hand, the collaborative nature of both Wikipedia and Wikidata implies that the global community of editors pays more attention to a relatively small set of items and their corresponding articles, while others have much less development.

Therefore, *a priori*, the distributions of the three indicators have a strong positive asymmetry characteristic of the 'long tail' phenomenon widely

studied in many social processes, especially in communities and digital platforms (Anderson, 2014). This is because there is a large number of items with low or intermediate values for the three indicators and a small number of items that have high values for them. For both reasons, a logarithmic transformation has been chosen to combine the three indicators. This allows normalizing the distribution of the three indicators and working with indicators whose original range of values is very different and comparing different data samples. Logarithmic transformation was chosen (instead of alternatives such as Z-score, Min-Max or Robust Scaling) because this normalization can be performed independently on each Wikidata item, without depending on the values achieved by other items in a dataset. This feature allows the isolated calculation of Wiki3DRank in real time in a fast and easy way.

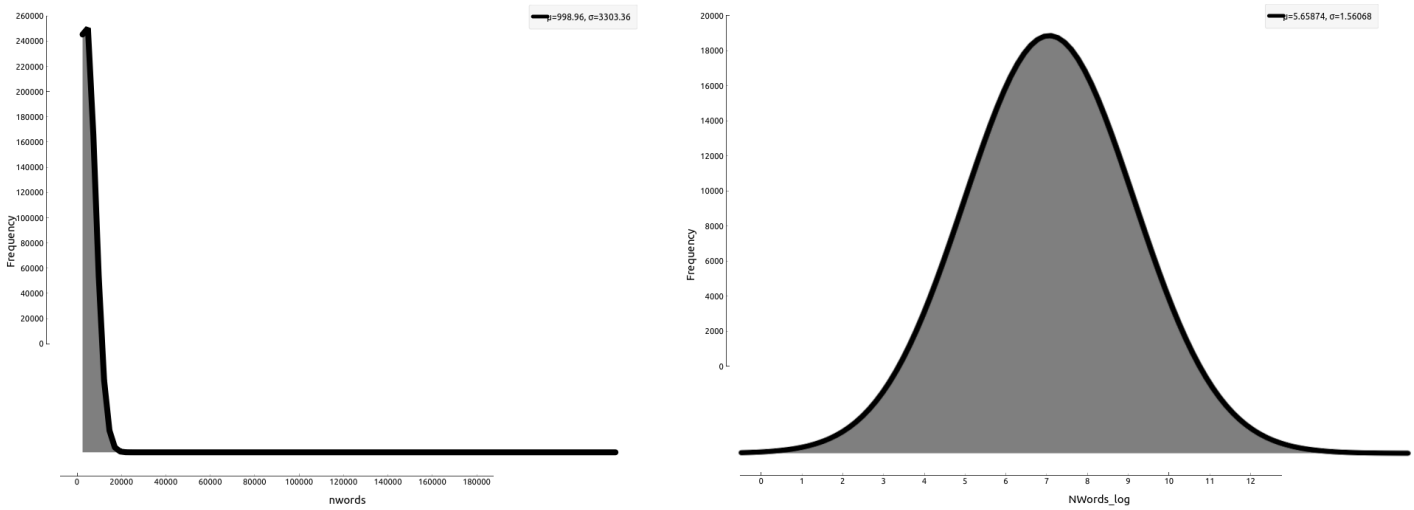

Figure 2. Original distribution of N_{Words} in a dataset of film Wikidata ítems (left) and distribution after logarithmic transformation log(1+N_{Words})

Figure 2 shows the original distribution of N_{Words} from a dataset of movie items from Wikidata and the one obtained after the logarithmic transformation. Taking into account the aforementioned logarithmic transformation, a first proposal was developed to calculate *Wiki3DRank* for each item as the aggregation of three components, as shown in the following equation:

$$a = log\ (1 + N_{Wikis})$$

$$b = log\ (1 + N_{Props})$$

$$c = log\ (1 + N_{Words})$$

$$Wiki3DRank = a + b + c$$

The results obtained from the previous work on the literary canon validated the ranking calculation method. In this regard, it was demonstrated that the combined use of the three indicators allowed for more coherent and precise results than the isolated use of one of the indicators or reducing dimensionality through principal component analysis (PCA).

Following this approach, this work adopts a more general perspective, proposing a model in which each item is represented as a vector, whose components would initially be the three mentioned indicators. Consequently, *Wiki3DRank* could be calculated as the module of this vector. The use of a vector space for the representation of these encyclopedic objects would allow the application of numerous vector composition techniques, similarity calculations, clustering, dimensionality reduction, or distance calculations.

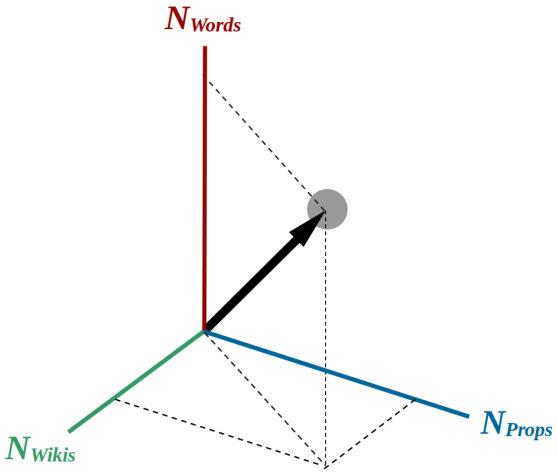

Figure 3. Representation of a Wikidata item as a vector

Therefore, in this work, it is proposed to use the vector module for the calculation of *Wiki3DRank*. In this way, for each item would be obtained a numeric magnitude:

$$a = log\ (1 + N_{Wikis})$$

$$b = log\ (1 + N_{Props})$$

$$c = log\ (1 + N_{Words})$$

$$Wiki3DRank = \sqrt{a^2 + b^2 + c^2}$$

This new approach of *Wiki3DRank* allows representing Wikidata items within a vector space. It also offers a mechanism of generalization which, as shown in the discussion section, allows for the incorporation of new components into the item vectors. One aspect to note is that the proposed method allows items to be evaluated independently, one by one, without the need to reprocess the entire dataset. In other words, the calculation of a *Wiki3DRank* for one item does not depend at any point on that obtained for other items. This allows a value to be obtained that can be compared with that obtained for other items and thus establish a ranking as necessary.

To implement the calculation of *Wiki3DRank*, it is necessary to access the data on which it is based, N_{Wikis}, N_{Props} and N_{Words}, for each specific item that needs to be evaluated. For all of these, there are sufficient online data sources. In the case of N_{Wikis} and N_{Props}, the Wikidata Query Service (WDQS, https://query.wikidata.org) can be used to obtain the corresponding data for each item, through SPARQL queries. For N_{Words}, access will be through the Xtools API (https://xtools.wmcloud.org/api). In the case of N_{Props}, statements about properties that are mere identifiers, according to the Wikidata data model, have not been taken into account. Other complementary data that have been incorporated into the script to enrich the *Wiki3DRank* query application are detailed, and these are discussed in the discussion section.

4. Results

This work presents two types of results. Firstly, a case analysis of the multidomain application of *Wiki3DRank* is carried out by evaluating the ranking data when applied to creative works, scientific objects, geographical reality, events, and architectural monuments. Secondly, a web application that allows the online and real-time calculation of *Wiki3DRank* is presented. The datasets and scripts used for the generation and processing of this use cases are available in the Zenodo repository, the source code of the *Wiki3dRank Calculation* application is available on GitHub, and the web application is hosted on the web servers of the University of Murcia, at the addresses indicated at the end of the work.

4.1. Use cases analysis

This section presents different use cases of rankings for knowledge objects, with the aim of providing an initial approach to their use in practical contexts. To better capture the behavior of the proposed measure, the following specific cases from domains with different characteristics have been selected: Literary works, movies, chemical elements, rivers, earthquakes, and cathedrals.

For the exploratory purpose of this part of the research, the concept of knowledge domain is considered without establishing a formal definition and seeking a generalist approach. Literary works and movies are understood in the context of creative works, as cultural achievements that are distributed massively and contain unique content linked to authorship and originality. Chemical elements constitute universally well-defined knowledge in basic sciences, stable and limited to a few items gathered in the periodic table.

Rivers are a type of knowledge object present all over the planet, abundant and studied from physical geography and other disciplines. Earthquakes are unexpected events with great social impact, with a long history and significant differences in intensity and consequences, while cathedrals are material elements characteristic of Christian culture, constituting a very recognizable type of monument and object of attention not only from art history but also from other fields such as tourism or heritage interpretation. In all of them, properties or attributes such as time factor, objectivity, interpretation, language, media impact, change, materiality, the scope from which they are studied, or their universality are combined differently. It should be noted, however, that cases have been chosen where, in principle, it is easy to delineate what falls within that category. Nevertheless, it should be considered that any collection activity requires, in its initial phase, to operationally define what does and does not fall within a category. Objects are not pure, but are accessed through a viewpoint; an example of this would be tributaries of rivers, parts of an architectural complex, or works in series or sagas.

For the selection of items, the querying power of Wikidata has been utilized, using the 'Instance of' (P31) classification. Although there are concerns for Wikidata's categorization, with issues about level of detail in its assignment, as well as a lack of rigor in the definition of classes and subclasses (Piscopo, 2019), direct querying by common classes allows for the retrieval of significant and precise sets of elements. For each use case, the most

frequent domain item has been retrieved directly, without having to resort to complex recursive queries. Some relevant objects may be left out, but for the illustrative purposes of this research, it is good enough. The following global results are obtained:

	Literary Works	*Movies*	*Chemical elements*	*Rivers*	*Earthquakes*	*Cathedrals*
Domain Item	Q7725634	Q11424	Q11344	Q4022	Q7944	Q56242215
Nº items	118497	267177	166	411443	2217	855
Statements	1070183 (*)	7760860	13384	2973093	16889	19153
Not ID Statements	794293 (*)	4156620	5230	2154220	12467	13884
Wikipedias articles	249263	1013165	18116	723662	9237	8344
Correlation N_{Wikis}-N_{Props}	0.534 0.383	0.769 0.693	0.802 0.935	0.745 0.531	0.581 0.373	0.827 0.809
Correlation N_{Wikis}-N_{Words}	0.854 0.545	0.823 0.786	0.776 0.949	0.766 0.645	0.876 0.744	0.891 0.891
Correlation N_{Props}-N_{Words}	0.496 0.327	0.630 0.661	0.820 0.929	0.567 0.432	0.495 0.293	0.843 0.843

Table I. Summary of item domains from the use case analysis (Data: January 2024)

In Table I, it is marked with (*) that during the data collection phase an anomaly was detected in the item Q213019 corresponding to the literary work 'The War of the Worlds' by George Orwell. This anomaly consists of the very recent introduction of 6,400 declarations of translations or editions of the work. The declarations were created by a bot between January 13 and 17, 2023. For this work, it has been decided not to consider these declarations because they represent an extreme value that significantly alters the statistical data. The Table includes data excluding the mentioned declarations in the domain of Literary Works.

The top 20 elements ordered by ranking in each domain would be as follows:

Chemical elements			*Rivers*		
Item	Label	Wiki3DRank	Item	Label	Wiki3DRank
Q897	gold	14.08077	Q1653	Danube	14.33553
Q677	iron	13.98320	Q584	Rhine	13.80449
Q753	copper	13.98121	Q3783	Amazon	13.53747
Q629	oxygen	13.97657	Q626	Volga	13.53228
Q556	hydrogen	13.83113	Q3392	Nile	13.52990
Q568	lithium	13.82849	Q5089	Ganges	13.44008
Q663	aluminium	13.73210	Q1644	Elbe	13.24763
Q623	carbon	13.70546	Q5413	Yangtze	13.21684
Q560	helium	13.69960	Q7348	Indus River	13.12271
Q1090	silver	13.65123	Q2251	Columbia River	13.07417
Q627	nitrogen	13.50882	Q7355	Yellow River	13.06024
Q925	mercury	13.41739	Q5419	Missouri River	13.04587
Q708	lead	13.41456	Q3503	Congo	12.97481
Q1098	uranium	13.41390	Q973	Ob	12.88339
Q758	zinc	13.40793	Q3542	Niger River	12.79182
Q871	arsenic	13.29793	Q40855	Dnieper	12.71570
Q716	titanium	13.28139	Q19686	River Thames	12.68431
Q682	sulfur	13.25536	Q1265	Colorado River	12.63784
Q674	phosphorus	13.20026	Q41986	Meuse	12.56699
Q725	chromium	13.17737	Q78707	Yenisey	12.55018

Table II. List of the top twenty results by Wiki3DRank for chemical elements and rivers (Data: January 2024)

In Table II, it can be seen how chemical elements, being a concept of basic science and a closed set of items, fit less into the ranking, with very narrow differences, while in rivers, there appears to be a close relationship between their length, and thus their impact on the territory, and their position in the ranking.

The data regarding earthquakes and cathedrals are in Table III.

Earthquakes			*Cathedrals*		
Item	*Label*	*Wiki3DRank*	*Item*	*Label*	*Wiki3DRank*
Q43777	2010 Haiti earthquake	12.25761	Q2981	Notre-Dame de Paris	13.57579
Q122351413	2023 Marrakesh-Safi earthquake	11.80661	Q5943	St. Stephen's Cathedral	12.90083
Q151835	2010 Chile earthquake	11.69395	Q4176	Cologne Cathedral	12.87427
Q19830062	April 2015 Nepal earthquake	11.66956	Q205136	Cathedral of Santiago de Compostela	12.86502
Q191055	1755 Lisbon earthquake	11.63256	Q180274	Notre-Dame de Chartres	12.62891
Q211386	1906 San Francisco earthquake	11.40086	Q106934	Notre-Dame d'Amiens	12.55677
Q1798567	1985 Mexico City earthquake	11.35079	Q18068	Milan Cathedral	12.51069
Q152033	2008 Sichuan earthquake	11.30518	Q1123180	Toledo Cathedral	12.46980
Q212618	1960 Valdivia earthquake	11.24643	Q231606	Catedral de Sevilla	12.46332
Q207918	2009 L'Aquila earthquake	11.22820	Q33200	Mosque-Cathedral of Cordoba	12.40758
Q214866	Great Hanshin earthquake	11.12279	Q84090	Archbasilica of St. John Lateran	12.39620
Q1348910	1908 Messina earthquake	11.08844	Q610961	Mexico City Metropolitan Cathedral	12.38517
Q56768333	2018 Sulawesi earthquake and tsunami	11.05736	Q5949	St. Vitus Cathedral	12.28075
Q112666390	June 2022 Afghanistan earthquake	10.97984	Q389210	Pamplona Cathedral	12.19373
Q191293	1556 Shaanxi earthquake	10.94832	Q184407	Basilica of Saint-Denis	12.17310
Q104535090	2020 Petrinja earthquake	10.89959	Q17155	Cathedral of the Holy Cross and Saint Eulalia	12.14300
Q151850	February 2011 Christchurch earthquake	10.89246	Q745460	Cathedral of Our Lady of Strasbourg	12.14084
Q462195	1976 Tangshan earthquake	10.85084	Q744420	Burgos Cathedral	12.09830
Q115322003	2022 Cianjur earthquake	10.82147	Q206823	Reims Cathedral	12.05725
Q189079	2011 Van earthquake	10.77991	Q22720	Speyer Cathedral	11.93270

Table III. List of the top twenty results by Wiki3DRank for earthquakes and cathedrals (January 2024)

It is observed that in the recorded earthquakes, their treatment as a phenomenon with cultural significance as opposed to mere geophysical aspects is perceived. Contemporary earthquakes have greater media coverage, but major historical disasters maintain their relevance. Cathedrals, logically, stand out for their monumental and touristic value, particularly those from a specific period in European and overseas Christian history.

Meanwhile, Table 4 shows the results for Literary Works and Movies.

Literary works			*Movies*		
Item	*Label*	*Wiki3DRank*	*Item*	*Label*	*Wiki3DRank*
Q428	Qur'an	14.61518	Q44578	Titanic	14.12210
Q9184	Book of Genesis	14.02116	Q24871	Avatar	13.79666
Q8275	Iliad	13.88599	Q17738	Star Wars: Episode IV – A New Hope	13.76991
Q35160	Odyssey	13.61286	Q163872	The Dark Knight	13.73499
Q480	Don Quixote	13.51878	Q47703	The Godfather	13.70764
Q43361	Harry Potter and the Philosopher's Stone	13.46623	Q104123	Pulp Fiction	13.66943
Q74287	The Hobbit	13.44176	Q23781155	Avengers: Endgame	13.65912
Q6511	Ulysses	13.40296	Q102438	Harry Potter and the Philosopher's Stone	13.61317
Q8272	Epic of Gilgamesh	13.30937	Q2875	Gone with the Wind	13.61120

Q8258	One Thousand and One Nights	13.29926	Q134430	Snow White and the Seven Dwarfs	13.56524
Q92640	Alice's Adventures in Wonderland	13.29783	Q23780914	Avengers: Infinity War	13.46630
Q208460	Nineteen Eighty-Four	13.29518	Q182218	The Avengers	13.44259
Q161531	War and Peace	13.27030	Q91540	Back to the Future	13.42724
Q8279	Shahnameh	13.23405	Q103474	2001: A Space Odyssey	13.42255
Q60220	Aeneid	13.21705	Q18407657	Captain America: Civil War	13.41430
Q150827	Frankenstein; or, The Modern Prometheus	13.14119	Q23780734	Black Panther	13.36900
Q19786	Old Testament	13.13063	Q14171368	Avengers: Age of Ultron	13.36220
Q165318	Crime and Punishment	13.11137	Q134773	Forrest Gump	13.35778
Q178869	One Hundred Years of Solitude	13.10061	Q132689	Casablanca	13.33334
Q46758	Harry Potter and the Deathly Hallows	13.09859	Q483941	Schindler's List	13.32423

Table IV. List of the top twenty results by Wiki3DRank for literary works and movies (Data: January 2024)

In Table IV, a certain balance is observed in literature between historical periods, cultures of origin, and genres. The presence of mythological works and religious texts is interesting. In the case of cinema, with a history of just over a century, there is a clear predominance of American cinema. A phenomenon of predominance of very recent movies is also detected, and it is noteworthy that among the top 20 results, six movies from the Marvel franchise are present.

4.2. Web app for real-time Wiki3DRank Calculation

The model used for the representation of *Wiki3DRank* allows for the online and real-time retrieval of data for its calculation. Calculating *Wiki3DRank* for an item does not require processing massive data dumps from Wikipedia or Wikidata. It is possible, through queries to WDQS and Xtools, to obtain data in a relatively simple and fast way.

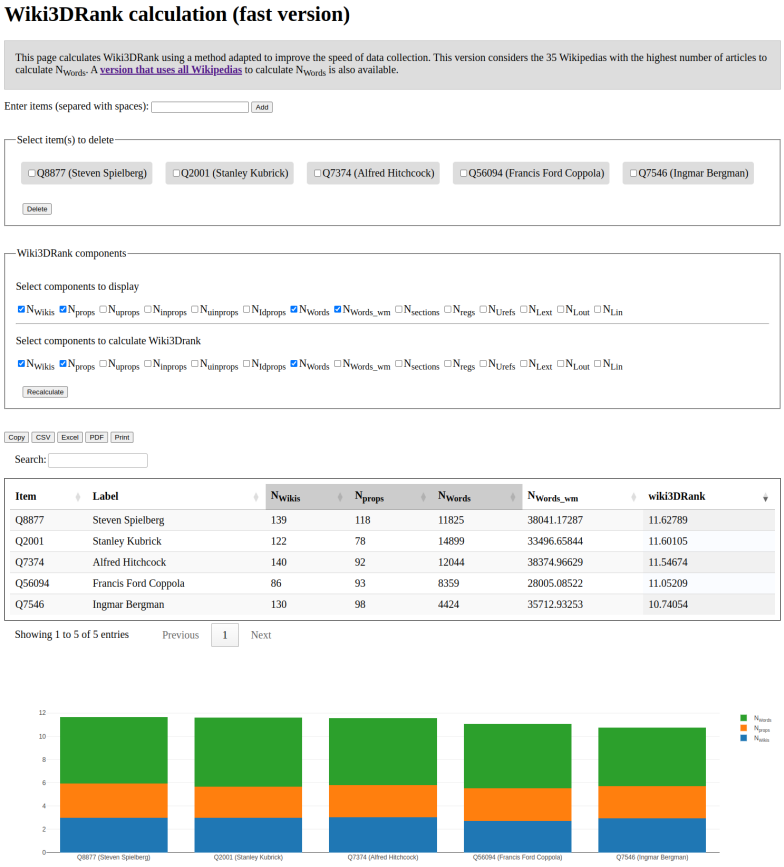

Figure 4. Example of use of Wiki3DRank Calculator for items of various film directors (January 2024, https://gicd.inf.um.es/wiki3drank)

As part of the results of this work, a ready-to-use application for the calculation of *Wiki3DRank* has been released. This application has been developed in Python (data retrieval) and PHP (*Wiki3DRank* calculation and results visualization). The operation is straightforward: the user only needs to enter one or several Wikidata item codes, and the application takes care of retrieving the data, performing the calculations, and displaying the results (see Figure 4).

The operation is straightforward: the user only needs to enter one or several Wikidata item codes, and the application takes care of retrieving the data, performing the calculations, and displaying the results (see Figure 2). An interesting feature is that it allows the separate selection of the calculation components that will be shown and those that will be used for the calculation of the Wiki3DRank. Given our demonstrative purpose, several alternative ways of calculating the Wiki3DRank are offered, which will be explained in the following section of this work.

Data on no more than 20 items can be retrieved simultaneously. They are displayed in a detailed table and a stacked column chart, both of which are exportable, and additional components and a faster retrieval version are available for flexible and efficient Wiki3DRank calculations. Users can explore the changing values of the metrics when optional features such as the number of incoming and outgoing links to articles or incoming and outgoing relationships are included in the calculations.

5. Discussion

The presented results are quite explicit in their approach, both from the standpoint of execution and calculation. Throughout, we have aimed to maintain simplicity in the process, making it easily replicable, explainable and observable. This section focuses on the discussion of three aspects of very different nature but which appear relevant to the research topic and professional community: a) the efficiency of calculation execution; b) the incorporation of additional components into *Wiki3DRank*; c) the refinement of results through domain properties.

5.1. Efficiency in Data Retrieval

One of the most relevant aspects to discuss regarding the calculation method is the efficiency of obtaining the N_{Words} indicator. In this work, we have showed how this indicator is calculated by summing the number of words in each article related to the item in question, retrieved through XTools. The primary drawback of this method is the need to establish a connection to the XTools API for each article in each Wikipedia. However, this process is constrained by the XTools server, which limits the use of asynchronous connections.

The experience with the *Wiki3DRank Calculation* tool has shown that the optimal number of concurrent connections is 35. This means that for items with a high number of articles, multiple asynchronous connections are required. This results in a certain level of delay in data retrieval and diminishes the quality of the user experience. A practical alternative is to calculate N_{Words} for each item while limiting the number of Wikipedia articles to 35, allowing all data to be retrieved in a single connection. The fast version of *Wiki3DRank Calculation* adopts this method, but always selects the 35 Wikipedias with the highest number of articles where the item has an equivalence. In this case, the indicator is referred to as N_{Words_fast}, and the value of the Wiki3DRank ranking is denoted as $Wiki3DRank_{fast}$.

Another different approach that could be interesting is the use of an alternative measure to the number of words in each article. This alternative measure would consider the overall editing effort at the Wikipedia article level, rather than considering articles individually. In this way, it would sum up the average words of all articles in each Wikipedia edition. Therefore, this indicator, referred to as N_{Words_wm}, would be calculated by summing the average words per article from each of the Wikipedias where the Wikidata item has an equivalent article. The total number of articles and words is obtained from the statistics page of each Wikipedia, and the data is stored on the server and can be periodically updated through a script that stores them as a JSON file. The calculation of words per article for each Wikipedia does not vary significantly, so the data would be obtained from loading a static file that could be updated periodically. The ranking value calculated using this method is referred to as $Wiki3DRank_{wm}$.

Based on the case analysis, a study was conducted comparing the results of the original ranking that uses N_{Words} with the results in which Wiki3DRank is calculated with N_{Words_wm} and N_{Words_fast}. The data in Figure 5 shows the ranking coincidence based on the sample size of items inversely ordered by the value of *Wiki3DRank*. It can be seen that in the three selected datasets (rivers, movies, and literary works), the values obtained with $Wiki3DRank_{fast}$ are much more accurate than those obtained with $Wiki3DRank_{wm}$. Therefore, it could be concluded that by using only the top thirty-five Wikipedia editions with the highest number of articles, values very close to the original *Wiki3DRank* are obtained while achieving

greater efficiency in data retrieval for calculation by being able to retrieve all the data of words in articles from Wikipedia editions to calculate N_{Words_fast} with a single connection from XTools.

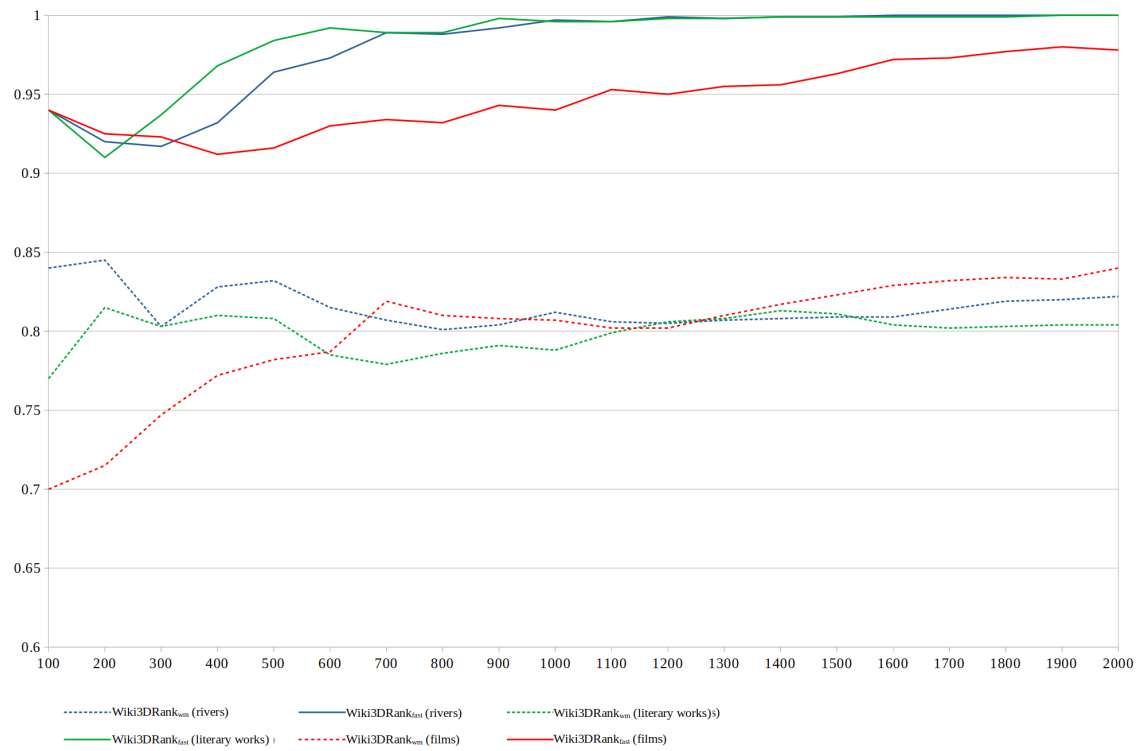

Figure 5. Degree of coincidence (Y-axis) in the top N positions of the Ranking (X-axis) of Wiki3DRank compared to $Wiki3DRank_{fast}$ and $Wiki3DRank_{wm}$ (Data: January 2024)

However, while there may appear to be significant differences, a more in-depth study would be necessary because the number of common works in the lists obtained with both methods is higher when the number of items is increased. When selecting lists with the top 150 items, it can be observed that both lists share 80.6% of the works, with 500 items, they share 80.8%, and with 1000 items, both lists share 78.8% of the works.

5.2. Increase in the number of components: from 3D to 360º

Another aspect that is suggestive is the incorporation of additional components into the vector. This would involve an expansion of the model that takes into account the following indicators in addition to those mentioned earlier. Table V provides a complete description of them:

Indicator	*Description*	*Source*	*Method*
N_{Uprops}	Frequency of use of different properties in the same item, excluding external identifier types.	Wikidata	SPARQL WDQS
$N_{Inprops}$	Number of incoming relations to an item.	Wikidata	SPARQL WDQS
$N_{Uinprops}$	Number of incoming relations from different items.	Wikidata	SPARQL WDQS
$N_{Idprops}$	Number of statements with external identifier properties in the same item.	Wikidata	SPARQL WDQS
$N_{Uidprops}$	Frequency of use of different external identifier properties in the same item.	Wikidata	SPARQL WDQS
$N_{Section}$	Total number of sections in all articles of an item.	Wikipedia	XTools Prose
N_{Refs}	Total number of references in all articles of the same item.	Wikipedia	XTools Prose
N_{Urefs}	Total number of unique references in all articles of the same item.	Wikipedia	XTools Prose
N_{Lext}	Total number of external links in all articles of the same item.	Wikipedia	XTools Links
N_{Lout}	Total number of outgoing links to other articles in all articles of the same item.	Wikipedia	XTools Links
N_{Lin}	Total number of incoming links from other articles to different articles of the same item.	Wikipedia	XTools Links

Table V. Aditional indicators for Wiki3DRank calculation

In the presented application, *Wiki3DRank Calculation*, the use of all these variables to calculate the relevance measure has been included as an optional element for the user. This allows for the analysis and evaluation of the results obtained on small samples. To demonstrate its generic validity, further research is necessary, which is beyond the scope of this publication.

In this regard, instead of incorporating all available elements to obtain a single metric, it seems more efficient to achieve analogous results in practice through minimal set of data, facilitating their interpretation and validation. In disciplines related to informetrics, more is not always better, nor does it provide more clarity to evaluate or understand information resources (Torres-Salinas; Robinson-García; Jiménez-Contreras, 2023). Beyond a certain point, the foreseeable marginal increases from incorporating more variables do not seem to result in an appreciable improvement in quality, but rather complicate the explanation and understanding of the proposed measure.

5.3. Refinement through Domain Properties: The Case of Creative Works

The difficulty of having a universal measure that satisfies the characteristics of all cases is undeniable. It is expected that measuring relevance through Wikidata-Wikipedia data will be more meaningful in certain domains. In the previous section, two examples related to creative works were presented, which are cultural objects or objects of knowledge for which there is an extensive collection of cataloging, compilation, dissemination, and assessment instruments. In this domain, specifically in films and literary works, the results show a certain "presentism" that appears to favor recent works. This phenomenon could be explained by greater attention from the editing community to recently socially impactful works (such as the release of major films, bestsellers, promotional campaigns, global media consumption patterns, etc.). To account for this effect, it may be interesting to add an additional component to represent the age of the works.

This new component would give higher weight to works with a longer publication or creation history. It is important to emphasize that while the previously mentioned components are directly incorporated into the vector calculation, in this case, it involves defining and justifying how it will be integrated into the calculation.

The first step is to identify candidate domain properties and analyze their implementation, especially the RDFS subproperty relationships, and their usage by the community (extension of their use, quality of the data entered) for the type of objects in the domain. In the case of literary works, previous research indicates that "presentism" is well-balanced in the metric (Pastor-Sánchez; Saorín; Baños-Moreno, 2023).

However, in the audiovisual domain, which is governed by more pronounced criteria of fast and mass consumption, it does seem to distort approaching more consensus-based results on the evaluation of works.

The properties that would be used to obtain the age of the work are P577 (publication date) and alternatively P571 (creation date). These properties reflect the instantiation of a work for dissemination, which is a fundamental element in any metadata schema, such as Dublin Core or Schema.org. The date properties are organized through a relative scheme of subclass and subproperty relationships, but their logic is not very rigorous.

In this work, it is proposed for this case to calculate the difference between the current date and the date corresponding to the work. The result of this calculation would be used to obtain a new component N_{Date} and would be incorporated into the calculation vector of Wiki3DRankDate.

In this way, the original calculation equation could be as follows:

$$a = log\ (1 + N_{Wikis})$$

$$b = log\ (1 + N_{Props})$$

$$c = log\ (1 + N_{Words})$$

$$d = lo\ g(1 + N_{Date})\,;\, N_{Date} = Year_{current} - Year_{pub}$$

$$Wiki3DRank = \sqrt{a^2 + b^2 + c^2 + d^2}$$

An example of the application of *Wiki3DRank*$_{Date}$ can be seen in Table 6, where it can be observed that some film items considered classics in cinema appear in the top twenty positions. With the original calculation of *Wiki3DRank*, these items were ranked lower (see Table VI, in the next page).

This preliminary approach suggests that the results maintain a strong component of current relevance, not to mention a distinctly American bias that barely captures the global nature of cinema as an artistic medium, beyond being an entertainment industry. It is important to note that in some cases, there is a need to set a limit on N_{Date}. In the domain of creative works or events, very distant dates can be found. While cinema has a history of just over a century, literature, painting, and other arts have a much longer tradition.

It would be necessary to consider the imbalance introduced by ancient or centuries-old classics. In these cases, the unlimited use of N_{Date} can have the opposite effect on the ranking results. That is, it could be the case of finding items linked to distant dates in the top positions because they would be excessively weighted upwards.

For this reason, a more thorough and specific study would be needed to correctly construct this modulator, assessing the maximum percentage of contribution of N_{Date} to the calculation of *Wiki3DRank*.

Item	*Label*	*Wiki3DRank*	*Item*	*Label*	*Wiki3DRank$_{Date}$*
Q44578	Titanic	14.12210	Q44578	Titanic	14.26255
Q24871	Avatar	13.79666	Q17738	Star Wars: Episode IV – A New Hope	14.01840
Q17738	Star Wars: Episode IV – A New Hope	13.76991	Q2875	Gone with the Wind	14.00175
Q163872	The Dark Knight	13.73499	Q134430	Snow White and the Seven Dwarfs	13.92978
Q47703	The Godfather	13.70764	Q47703	The Godfather	13.90140
Q104123	Pulp Fiction	13.66943	Q132689	Casablanca	13.82313
Q23781155	Avengers: Endgame	13.65912	Q103474	2001: A Space Odyssey	13.71408
Q102438	Harry Potter and the Philosopher's Stone	13.61317	Q102438	Harry Potter and the Philosopher's Stone	13.66237
Q2875	Gone with the Wind	13.61120	Q103569	Alien	13.62753
Q134430	Snow White and the Seven Dwarfs	13.56524	Q184843	Blade Runner	13.61312
Q23780914	Avengers: Infinity War	13.46630	Q104123	Pulp Fiction	13.59927
Q182218	The Avengers	13.44259	Q24871	Avatar	13.56592
Q91540	Back to the Future	13.42724	Q41483	The Good, the Bad and the Ugly	13.55800
Q103474	2001: A Space Odyssey	13.42255	Q91540	Back to the Future	13.55078
Q18407657	Captain America: Civil War	13.41430	Q23781155	Avengers: Endgame	13.52241
Q23780734	Black Panther	13.36900	Q24815	Citizen Kane	13.52018
Q14171368	Avengers: Age of Ultron	13.36220	Q180098	Ben-Hur	13.51389
Q134773	Forrest Gump	13.35778	Q483941	Schindler's List	13.49594
Q132689	Casablanca	13.33334	Q42051	Star Wars: Episode III – Revenge of the Sith	13.43629
Q483941	Schindler's List	13.32423	Q134773	Forrest Gump	13.43046

Table VI. Comparison of the top twenty results between Wiki3DRank and Wiki3DRank$_{Date}$ for the film dataset (Data: January 2024)

6. Conclusions and future research

Throughout this work, we have presented a methodology for calculating *Wiki3DRank* that is not only simple in its formulation based on the use of a vector space but also easy to read and interpret. When applied to various use cases involving objects of knowledge, it is evident that its operation on objects that have some "social impact" (diffusion, audience, value, population, territory, social repercussions, etc.) reflects notability or relevance. Measurement has a very visible value in the social business of attention, but it is often overlooked that it can also be valuable in knowledge organization systems. The concepts in a taxonomy or the elements in an authority list do not all have the same significance. The existence of easily accessible tools, transparent and reproducible procedures, as well as standardized metrics provided by trusted providers, for measuring different aspects of the knowledge society contributes to the development of the language of digital humanities and the information science field.

Regarding the three core elements used to construct the metric, N_{Wikis}, N_{Props} and N_{Words}, it would be necessary to analyze in more detail the correlations between these variables to better understand their contribution to the ranking score and to perform clustering processes. It also seems advisable to explore and validate the opportunity of using incoming and/or outgoing links or connections as another layer to understand the quality of content present in the articles, although this poses operational challenges in determining the general or relative values of each node in a graph. The approach of dimensionality reduction, selecting variables that, when combined, allow for filtering, grouping, and ranking, is viable, but it makes sense to consider that for more refined

results, domain-specific variables that are meaningful for a specific type of knowledge object should also be managed. The use of domain-specific variables, such as the date on creative works, artist awards, and even their weighting in the calculation of the final score, implies a rigorous construction and validation of composite indicators (Blasco-Blasco, Rodríguez-Castro; Tuñez-López, 2020). This could undoubtedly be an interesting avenue for future research involving the integration of data from outside the Wikimedia sphere. The use of the number of words as a variable to reflect the depth of content represents a limitation in capturing the value of information; the structure of content in sections, the presence of notes, bibliography, or illustrations would provide a richer reflection of the quality of encyclopedic content. It is also necessary to consider that there is a trade-off between data richness and processing speed when using XTools instead of MediaWiki APIs.

To delimit the scope of objects analysis, it is noted that there are difficulties in using Wikidata's taxonomy of classes to accurately and exhaustively select resources of the same type for the purpose of analyzing related elements within the same domain or field. This is a different problem from the mere calculation but complicates its applicability for sectorial studies because the ranking measure makes sense within a robust set of comparable elements.

Network analysis methodologies are not the only ones applicable to large connected datasets. In many cases, there are less costly approaches that reduce access barriers or provide interpretable data with sufficient clarity. It is possible to obtain results without processing and reprocessing an entire dataset. This strategy facilitates the assignment of metric values and their updating, as the value does not depend on the state of all elements but is derived from the calculation of individually validated and consensus-based properties. Obtaining values on-the-fly can be useful for enriching other information discovery systems, in the recommendation and result filtering process, and even as a way to reproduce approximate measures of centrality without having access to the complete graph.

Another limitation to consider is that objects of knowledge within the realm of local culture, which often appear in only one Wikipedia language edition, are penalized in *Wiki3DRank*. The model used assumes that knowledge is universal and is reflected in as many languages as possible. Global objects, by their objective nature, size, and impact, have an obvious advantage over aspects linked to a specific culture or region. It is worth mentioning that the enormous attention given to the term "Big Data" tends to overshadow very useful approaches based on large datasets (Long data) and the efficient use of a few representative variables. The economy of means contains its own value proposition for gaining insights into its objects of study, as opposed to deploying analytical and data arsenals operated by internet giants. Ocean-sized data—internet data—often acts as a substitute for unavailable or highly imperfect data on the phenomena we want to observe. There isn't always data, or it is very imperfect, despite the collective delusion that automatically believes otherwise (Borgman, 2017). The mandate of "smart" must be understood with subtlety, as the need for sufficient dimension and a sufficient understanding of the variety of interpretive registers (Halpern; Mitchell, 2022).

The cumulative definition applied by *Wiki3DRank* practically means that an element cannot decrease in value over time (except in cases where articles are deleted or condensed, which is uncommon). It would be worth conducting simulations to assess the cost of new elements competing with established ones or the potential distortion of the ranking when many elements have very similar values.

It's also important to consider that the awareness of a measurement system can facilitate its manipulation. Quantifying the ranking implies that actions can be taken to generate the type of data that increases its ranking. Without examining actions of a destructive nature, in terms of "overfeeding" a resource to improve its ranking, our superficial estimation suggests that massive editing actions in many languages are costly to orchestrate, and enriching records by creating new statements has a very controlled effect. However, two exceptions can have more pronounced dimensions: firstly, it has been observed that the occasional use of Wikidata as a comprehensive catalog of editions of a work can generate a volume of data that impacts the ranking (this was observed in the item "The War of the Worlds" by George Orwell, with 6000 statements of property P655 [has edition]). Secondly, the widespread availability of multilingual generative AI engines can simplify selective bombardments of large amounts of text across many Wikipedia editions, which can also affect the ranking. The first situation can be taken into account by monitoring the activities of the interest group working towards making Wikidata a global bibliographic and cataloging database; the second situation opens up a much less easy-to-define and track scenario.

Finally, it's worth noting the clarity and simplicity of the proposed vector calculation, which allows for the addition of components that are incorporated

into the calculation mechanics without the need to construct complex composite metrics.

The authors of this work intend to make continuous improvements to the presented web application, in addition to providing the research community with the source code, data, and scripts used in the study. This will enable the exploration of *Wiki3DRank* for both large collections of items (types, classes) and ad-hoc item selections, allowing for agile comparisons.

Access to research data and scripts

Dataset and data processing scripts: https://doi.org/10.5281/zenodo.10576041

Source code for the Wiki3dRank Calculation web application: https://github.com/j-pastor/wiki3drank

Wiki3dRank Calculation web app: https://gicd.inf.um.es/wiki3drank

References

Ahnert, Ruth; Ahnert, Sebastian; Coleman, Catherine; Weingart, Scott (2020). The Network Turn: Changing Perspectives in the Humanities. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108866804

Anderson, Chris (2014). The Longer Tail Why the Future of Business is Selling Less of More. New York: Hachette Books.

Beytía, Pablo; Schobin, Janosch (2020) Networked Pantheon: a Relational Database of Globally Famous People. // Research Data Journal for the Humanities and Social Sciences. 5, 50-65. https://doi.org/10.1163/24523666-00501002

Bianchini, Carlo; y Sardo, Lucia (2022). Wikidata : a new perspective towards universal bibliographic control. // JLIS. 13:1, 291-311. https://doi.org/10.4403/jlis.it-12725

Blank, Grant (2007). Critics, Ratings, and Society. Lanham: Rowman and Littlefield.

Blasco-Blasco, Olga; Rodríguez-Castro, Marta; Túñez-López, Miguel (2020). Composite indicators as an innovative methodology for Communication Sciences: implementation for the assessment of European public service media". // Profesional de la información. 29, n. 4, e290437, 2020. https://doi.org/10.3145/epi.2020.jul.37

Borgman, Christine L. (2017). Big data, little data, no data. Cambridge, Massachusetts: The MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

Brown, Andrew (2011). A brief history of encyclopaedias: from Pliny to Wikipedia. Londres: Hesperus.

Halpern, Orit; Mitchell, Robert (2022) The smartness mandate. Cambridge, Massachusetts: The MIT Press. https://doi.org/10.7551/mitpress/14623.001.0001

Hill, Benjamin Mako; Shaw, Aaron (2020). The Most Important Laboratory for Social Scientific and Computing Research in History. // Reagle, Joseph; Koerner, Jackie (eds.). Wikipedia @ 20: Stories of an Incomplete Revolution. Cambridge, Massachusetts: The MIT Press. https://doi.org/10.7551/mitpress/12366.001.0001

Lewoniewski, Włodzimierz; Węcel, Krzysztof; Abramowicz, Witold (2019). Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics. // Computers. 8:3, 60. https://doi.org/10.3390/computers8030060

McDowell, Zachary J.; Vetter, Matthew A (2022). Wikipedia and the Representation of Reality. New York: Routledge. https://doi.org/10.4324/9781003094081

Minguillón, Julia; Lerga, Maura; Aibar, Eduard; Lladós-Masllorens, Josep; y Meseguer-Artola, Antoni (2017). Semi-automatic generation of a corpus of Wikipedia articles on science and technology. // El Profesional de la Información. 26:5, 995-1004. https://doi.org/10.3145/epi.2017.sep.20

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

Moás, Pedro Miguel; Teixeira Lopes, Carla (2023). Automatic Quality Assessment of Wikipedia Articles: A Systematic Literature Review. // ACM Computing Surveys. 56:4, article 95. https://doi.org/10.1145/3625286

Nielsen, Finn Årup (2012). Wikipedia Research and Tools: Review and Comments. http://doi.org/10.2139/ssrn.2129874

Piscopo, Alessandro; y Simperl, Elena (2018). Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata. // Proceedings of the ACM on Human-Computer Interaction. 2:CSCW, Article 141. https://doi.org/10.1145/3274410

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

Shenoy, Kartik; Ilievski, Filip; Garijo, Daniel; Schwabe, Daniel; Szekely, Pedro (2022). A study of the quality of Wikidata. Journal of Web Semantics. 72, 100679. https://doi.org/10.1016/j.websem.2021.100679

Skiena, Steven; Ward, Charles B. (2014). Who's bigger? Where historical figures really rank. Cambridge: Cambridge University Press.

Torres-Salinas, Daniel; Robinson-García, Nicolás; Jiménez-Contreras, Evaristo (2023). The bibliometric journey towards technological and social change: A review of current challenges and issues. // Profesional de la información. 32:2, e320228. https://doi.org/10.3145/epi.2023.mar.28

Este artículo es la versión en inglés del artículo siguiente publicado en el mismo número.

Enviado: 2024-03-14. Segunda versión: 2024-05-21.
Aceptado: 2024-05-23.