# The World Literature Knowledge Graph

Marco Antonio Stranisci[1][0000−0001−9337−7250]⋆, Eleonora Bernasconi[2][0000−0003−3142−3084]⋆, Viviana Patti[1][0000−0001−5991−370X], Stefano Ferilli[2][0000−0003−1118−0601], Miguel Ceriani[2,3][0000−0002−5074−2112], and Rossana Damiano[1][0000−0001−9866−2843]

[1] University of Turin, Corso Svizzera 185, Turin, Italy
[2] University of Bari Aldo Moro, Via Orabona 4, Bari, Italy
[3] ISTC-CNR, Via S. Martino della Battaglia 44, Roma, Italy
{marcoantonio.stranisci,viviana.patti,rossana.damiano}@unito.it
{eleonora.bernasconi,stefano.ferilli,miguel.ceriani}@uniba.it

**Abstract.** Digital media have enabled the access to unprecedented literary knowledge. Authors, readers, and scholars are now able to discover and share an increasing amount of information about books and their authors. However, these sources of knowledge are fragmented and do not adequately represent non-Western writers and their works. In this paper we present The World Literature Knowledge Graph (WL-KG), a semantic resource containing 194,346 writers and 971,210 works, specifically designed for exploring facts about literary works and authors from different parts of the world. The knowledge graph integrates information about the reception of literary works gathered from 3 different communities of readers, aligned according to a single semantic model. The resource is accessible through an online visualization platform, which can be found at the following URL: https://literaturegraph.di.unito.it. This platform has been rigorously tested and validated by 3 distinct categories of experts who have found it to be highly beneficial for their respective work domains. These categories include teachers, researchers in the humanities, and professionals in the publishing industry. The feedback received from these experts confirms that they can effectively utilize the platform to enhance their work processes and to achieve valuable outcomes.

**Keywords:** Knowledge Graph · World Literature · Information Visualization

## 1 Introduction

The impact of digital media on the literary ecosystem has led to a transformation of reading [23] and researching [24] practices. Digital media represent an unprecedented opportunity for studying the World Literature [10]. Digital platforms are not only open windows on different parts of the world, but also privileged viewpoints on how communities of readers receive and share literary works [34].

⋆ Contributed equally to this work.

The opportunities emerging from this transformation are, however, limited by a series of issues. The knowledge stored in these archives is vast, but fragmented: only a minimal part of writers and works is mapped from one source to another and many archives do not rely on a semantic model. This hinders the study of the writers and their reception by different groups of readers. Furthermore, it has been proved that some of these resources are characterized by the underrepresentation of non-Western people. It is the case of Wikidata [1] and Wikipedia [13] that are both affected by an ethnic and gender bias. In a recent work [33] the analysis of 48,789 biographies from Wikipedia extends the findings from previous work indicating that representational biases are present in an allegedly objective source such as Wikipedia along intersectional axes [9], namely ethnicity and gender.

The World Literature Knowledge Graph (WL-KG) is a knowledge base developed for tackling these issues. The resource includes 194,346 writers and their works gathered from three sources of knowledge: Wikidata, Open Library, and Goodreads. Such a collection relies on a common ontology network [35] specifically developed with the aim of emphasizing the ethnic origin of writers and the readers' response about them and their works.

The WL-KG is intended to support two main types of tasks: (i) the analysis of the underrepresentation of non-Western writers; (ii) the reception of works by different communities of readers. These tasks, in turn, can support the implementation of applications like recommender systems [28], and discovery tools [27], which may take advantage from the more balanced representation of literary world provided by the knowledge base. The WL-KG is also intended as a tool for all professionals that work in the literary field (e.g., researchers in the humanities and publishers) and operate in multicultural contexts (e.g., teachers, educators, activists). In order to make the resource accessible to these target, it is hosted on a visualization platform [4] that allows for a graph-based exploration of the KG. Both the platform and the WL-KG were tested by 3 categories of experts who evaluated them along three dimensions: completeness, accuracy, and usability. Results showed that our resource may be considered as an alternative to traditional literary search tools, especially for the discovery of new writers.

This paper is structured as follows. In Section 2 related work and theoretical background are presented. Section 3 describes the semantic model on which the WL-KG relies. Section 4 describes the creation of the resource, while Section 5 illustrates its implementation in a visualization platform. Section 6 reports on the evaluation of the resource.

## 2 Background and Related Work

In this section we first briefly describe the World Literature theoretical framework. Then, we review the related work in two fields: semantic resources designed for literary studies and Linked Data visualization platforms.

### 2.1 Theoretical Framework

World Literature is a recent approach to literary studies that emphasizes the idea of works as windows on different parts of the world [10]. In such a perspective, national and chronological boundaries must be overcome and a crucial step of the analysis is how works transcend their local contexts to be globally received [19, 2]. Such a framework gained prominence in last years thanks to the availability of an unprecedented knowledge about writers and their works enabled by social media: this paved the way for the development of distant reading approaches [22] as well as digital humanities studies of digital platforms [18]. The centrality of reception and the emphasis on a non-Western-centric approach are two features from this theory that were adopted for modeling the WL-KG. In fact, our resource can be used not only for discovering writers and works from the world, but also to analyze how communities of readers increase or decrease their underrepresentation, and to devise ways to contrast it.

### 2.2 Semantic Technologies for Literary Studies

Several digital resources that provide information about literary works and writers are available online. Wikidata [40] is a general-purpose KG which includes knowledge about writers and their works. Other archives are domain-specific: Goodreads is a social cataloging website owned by Amazon, where readers share their impressions about books. Open Library is a project of the Internet Archive[4] where users can borrow books. Among these three archives, only Wikidata relies on the Linked Open Data paradigm. Open Library exposes its data through APIs, while Goodreads dismissed its APIs in 2020. This leads to issues in data gathering and mapping, since there is no unified model to align these resources.

Some digital archives are monographic and curated by teams of experts. It is the case of The European Literary Text Collection[5] [30], a multi-lingual dataset of novels written from 1848 to 1920; DraCor[6] [14], a collection of plays corpora in multiple languages; MiMoText[7], a parallel corpus of French and German novels published from 1750 to 1799.

Other resources are more oriented to explore the intersection between people and society. The Japanese Visual Media Graph[8] [25] gathers data about Japanese visual media (including manga and visual novels) from communities of fans. The Orlando Textbase[9] [31] is a KG developed to explore feminist literature. WeChangeEd[10] [38] is a KG of 1,800 female editors born between 1710 and 1920 aligned with Wikidata.

[4] https://archive.org
[5] https://www.distant-reading.net/eltec
[6] https://dracor.org
[7] https://mimotext.github.io
[8] https://jvmg.iuk.hdm-stuttgart.de
[9] https://www.artsrn.ualberta.ca/orlando
[10] https://www.wechanged.ugent.be

The WL-KG is the first resource designed to study the intersection between literary production and ethnic information about writers. There are research projects that analyze the world of literature according to Wikipedia [18], but this is the first attempt to release a resource which could be at the same time a platform to foster digital humanities and literary studies and a benchmark dataset for analyzing the knowledge gaps that affect an authoritative source like Wikidata in the literary domain.

### 2.3 Visualization Platforms

Many works deal with interfaces for visualising Linked Data [11, 26, 16, 21, 8, 37, 17, 39], but only some focus on exploring and disseminating domains related to digital humanities, primarily digital libraries [3, 12]. The interaction paradigm and the information reduction strategies are the two main characteristics of an interface for visualising Linked Data.

ARCA [4] is a modular system that deals with knowledge extraction from a digital library, visualisation, and collaborative validation of automatically extracted associations between concepts and books [5]. ARCA uses two different interaction paradigms: the node-link paradigm for visualising resources extracted and linked to the DBpedia knowledge base [11], and the tabular paradigm for the visualisation of additional metadata related to books. As an information reduction strategy, ARCA allows for incremental visualisation of resources.

On the other hand, Yewno Discover [6] allows node-link visualisation of concepts contained in a digital library. Unlike ARCA, Yewno has a static and non-incremental visualisation of resources but uses ranking algorithms to filter the displayed content.

Another tool is ResearchSpace [12], an open-source platform that facilitates working with digital cultural heritage data in a Linked Data environment, enabling improved discoverability and reuse of data. The platform includes a node-link interaction paradigm, which employs incremental visualization for knowledge exploration. Additionally, it allows for collaborative annotation of texts or images.

Thanks to the flexibility and modularity of the ARCA system, we have chosen to build upon it by creating an extension called SKATEBOARD (Semantic Knowledge Advanced Tool for Extraction Browsing Organization Annotation Retrieval and Discovery), as described in Section 5. This extension has been customized and updated to meet the specific needs of users interacting with the World Literature Knowledge Graph.

## 3 The Semantic Model

In this section the semantic model adopted for the WL-KG is described. After a general introduction of the model and the authoritative ontologies to which it is

[11] `https://dbpedia.org`

aligned, we focus on two aspects that we modeled through our ontology network: the interaction between writers and their ethnic origin and the representation of the publishing history of the works.

### 3.1 The UR-Ontology Network

The ontology network serves two main functions: modeling ethnic-based underrepresentation of writers; mapping different digital libraries under a unique data model. Data in the WL-KG are modeled according to the Under-Represented Ontology Network (UR-O) composed of two modules: a revised version of the Under-Represented Writers Ontology (URW-O)[12] [35] and a module for the encoding of works: the Ontology of Under-Represented Books (URB-O)[13].

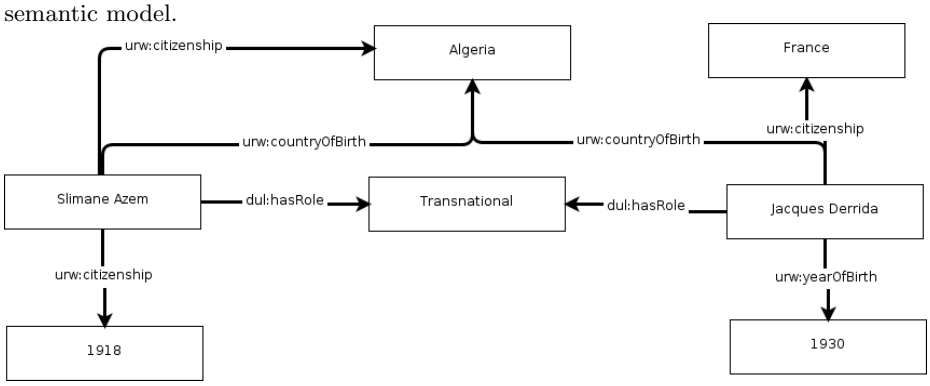
The ontology network is mapped onto three authoritative ontologies: the Functional Requirements for Bibliographic Records (FRBR) [36], the PROV Ontology (PROV-O) [20], and the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [15]. FRBR is a standard for modeling the relationship between a work (FRBR:Work), its expressions (FRBR:Expression), and manifestations (FRBR:Manifestation). From PROV-O the relationships of attribution, association, and derivation are inherited, in order to make explicit the sources from which data were gathered (**prov:wasDerivedFrom**), the people and organizations involved in specific editions of given works (**prov:wasAssociatedWith**) and their roles (e.g., publisher, translator), and the attribution of a work to its creator (**prov:wasAttributedTo**). DOLCE has been used as a reference model for encoding biographical and publishing events, which are represented as time-bounded perdurants in which entities play specific roles. This allows representing publications as events where sets of entities participate (**dul:hasParticipant**) and life events (e.g., Birth, Migration) as situations which are setting for (**dul:isSettingFor**) agents and their roles.

### 3.2 Modeling Underrepresentation

For modeling ethnic-based underrepresentation of writers we relied on two criteria derived from post-colonial studies - Gayatri Spivak's work in particular [32]. To be potentially under-represented an author must either (i) be born in a non-Western former colony country or (ii) belong to an ethnic minority in a Western country. Using the country of birth as a criterion is prone to false positives though, since many writers with Western origin were born in former colonies (e.g., George Orwell, Rudyard Kipling). In order to mitigate such issue we chose to adopt the term 'Transnational', which is broader than 'Under-Represented' since it refers to people who "operated outside their own nation's boundaries, or negotiated with them" [7]. Furthermore, we classified as 'Transnational' only people born in former colonies from Latin America and Caribbeans since 1808, and in former African and Asian colonies since 1917, to reduce the number of

[12] `https://purl.archive.org/urwriters/lode`
[13] `https://purl.archive.org/urbooks/lode`

**Fig. 1.** An example of how the concept of 'Transnational' writer is encoded in our semantic model.




people of Western origin selected by this condition. The first date marks the beginning of the Spanish American wars of independence; the second was chosen as a symbolic beginning of the decolonization process in Africa and Asia. Finally, we coupled the condition of being 'Transnational' with the citizenship of an author in order to reveal potentially false under-represented writers which may be still present in the knowledge base. As it can be observed in Figure 1, Jacques Derrida and Slimane Azem are both classified (**dul:hasRole**) as 'Transnational' in the KG, since they were born in Algeria, a former African colony, in 1918 and 1930. The specification of their citizenship (**urw:citizenship**) provides additional information about Jacques Derrida, who was not an Algerian citizen despite Algeria is his country of birth. This allows users to infer his European origins.

### 3.3 Modeling Works Publishing History

Before gathering data from Wikidata, Open Library, and Goodreads we designed a common data model for aligning literary information that the platform represents in heterogeneous shapes. Following the FRBR ontology, we defined each work in the platform as an instance of type FRBR:EXPRESSION, which is described as the "intellectual or artistic realization of a work in the form of alphanumeric, musical, or choreographic notation". We then defined the concept of URB:EDITION as a subclass of FRBR:MANIFESTATION, namely "the physical embodiment of an expression of a work". These two concepts are linked through the property **frbr:embodiment**. Such semantic relationship is wrapped in a URB:PUBLICATION pattern, which is a subclass of a DUL:EVENT. An event in DOLCE can be used as a reification to provide richer descriptions of a property. In our case this type of pattern is adopted for two reasons: (i) expressing a large number of facts about an edition (place, date, language of publishing and publisher) in a compact way; (ii) encoding roles of people who contributed to a publication without being the author of a work. A final feature of the

semantic model is the reception of works from communities of readers. Depending on the source of knowledge from which a work is derived, it may have an average rating (**urb:rated**), a number of ratings (**urb:numberOfRatings**), or a number of readers (**urb:numberOfReaders**). Figure 2 shows an example of our representation of works. 'Harry Potter e il Prigioniero di Azkaban', namely the Italian version (FRBR:EXPRESSION) of the 3rd Harry Potter book, **prov:wasAttributedTo** to J. K. Rowling, has an average rating and a number of ratings from the Goodreads community, and it has as **frbr:embodiment** the '1999 edition'. The latter in turn participates (**dul:isParticipantIn**) to a URB:PUBLICATION, a blank node entity that can be used for expressing several information: country of publication, year of publication, publisher, and translator. The translator is linked to the publication through the property **prov:wasAssociatedWith** and **dul:hasRole** 'translator'. Such representation supports a thorough exploration of the intersections between writers' biographies and their publishing history as well as a more accurate analysis of their relationships with other authors and people working in the publishing industry. It is however a verbose encoding that may affect the usage of this resource. In order to avoid this issue, we defined a set of property chains that directly link works to bibliographical information. Examples of these properties are shown in red in Figure 2.

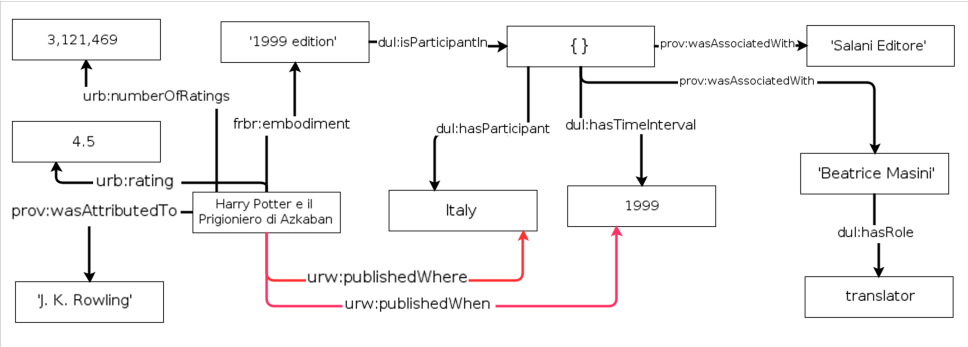



**Fig. 2.** An example of publication.

## 4 Creation of the WL-KG

In this section, we describe the process involved in creating our World Literature Knowledge Graph (WL-KG), which can be queried online through the SPARQL endpoint available at `https://kgccc.di.unito.it/sparql/wl-kg`. We first describe our strategy for mapping knowledge from Wikidata onto Open Library and Goodreads. We then introduce our strategy for evaluating the mapping. Finally, we provide some statistics about the number of literary facts collected

from each platform and about the interaction of communities of readers with works.

### 4.1 Mapping between Platforms

The data collection process started from Wikidata. From this knowledge base we gathered all the 194,346 entities of type Person (wd:Q5) with occupation (wdt:P106) writer (wd:Q36180), novelist (wd:Q6625963), or poet (wd:Q49757) born after 1808 and having information about their place of birth. For each author, we collected the ethnic group, gender, date and place of death, Wikipedia page, and all the works associated with them. We converted all geographical information gathered from Wikidata to the "ISO 3166-1 alpha 3" code[14] (e.g., IND, NGA), which is internationally recognized as a standard for referencing modern countries.

To enrich the knowledge base we first conducted a quantitative analysis of their external identifiers in Wikidata pages on writers. We focused on three of them: writers' Virtual International Authority File Name (VIAF) IDs, Open Library IDs and Goodreads IDs. A fourth platform, Library Things, was not included in the data collection process given the low number of links from Wikidata and the impossibility of automatically obtaining authors' IDs from that website. In Table 1 it is possible to observe that the 84% of writers has a VIAF ID, the 18.5% an Open Library ID, and the 4.5% a Goodreads ID. In order to increase the percentage of writers mapped to VIAF and Open Library identifiers, we adopted three heuristics:

- We retrieved all the names of the writers through the OpenLibrary APIs and kept only the entities fulfilling two conditions: (a) an exact string match between the author name in our KG and the one in OpenLibrary; (b) the same year of birth in our KG and in OpenLibrary. As a result, we obtained 19,737 additional ids.
- We scraped all writers' names from Goodreads sitemap[15] filtering out all homonyms. We then mapped all the names in our KG onto Goodreads author list, keeping only the string matches. We thus obtained 26,019 new ids.
- We searched all ISBNs related to each authors through VIAF and performed a search through ISBN on Open Library and Goodreads, that allowed retrieving 22,661 Open Library IDs and 44,142 Goodreads IDs.

### 4.2 Quality Assessment of the Mapping

After the mapping, we performed a quality assessment of a sample of links between Wikidata and Goodreads, and between Wikidata and Open Library for removing incorrect links before gathering works. Our evaluation strategy is composed of three steps. We computed the Gestalt pattern similarity [29]

[14] `https://www.iso.org/iso-3166-country-codes.html`
[15] `https://www.goodreads.com/siteindex.author.xml`

between the names of the same writer in different platforms. For instance, Esther Salaman[16] is linked to her Goodreads page[17], where she is referred as 'Esther Polianowsky Salaman'. The two strings have a Gestalt pattern score [29] of 0.7. Then, we manually checked random samples of 100 name pairs with 7 degrees of similarity: $x < 0.1$, $0.1 \geq x < 0.2$, $0.2 \geq x < 0.3$, $0.3 \geq x < 0.4$, $0.4 \geq x < 0.5$, $0.5 \geq x < 0.6$, $0.6 \geq x < 0.7$. As it can be observed in Figure 3, the percentage of correct links is directly proportional to the similarity between the name by which the writer is referred to in different platforms. In particular, the accuracy dramatically increases with a similarity between 0.5 and 0.6 (77% of correct links) reaching a 89% of accuracy with a similarity between 0.6 and 0.7.

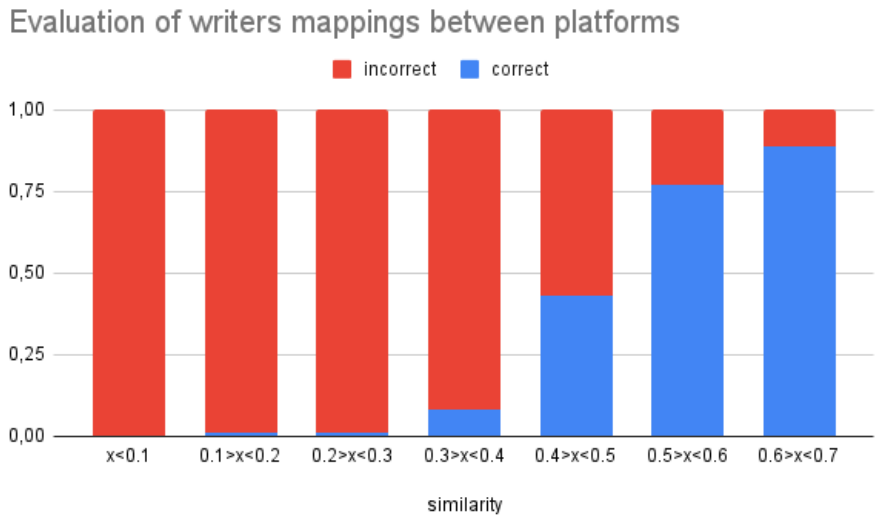



**Fig. 3.** Results of the evaluation of writers mappings between Wikidata, Goodreads, and Open Library.

Finally, we set a similarity threshold for filtering out potentially incorrect links. In order to privilege precision over recall, we set the threshold at 0.7. As a final result, we obtained 71,706 (36.8%) writers with an Open Library ID and 79,158 (40.7%) with a Goodreads ID (Table 1). The percentage of writers linked to at least one of the two platforms is 54%.

### 4.3 Data Collection and Statistics

After the augmentation of external identifiers of authors, we collected all their works in these platforms. OpenLibrary APIs allow retrieving all works, and for

[16] `http://www.wikidata.org/entity/Q4405658`
[17] `https://www.goodreads.com/author/show/618352`

| Identifier | Before Mapping | After Mapping |
|---|---|---|
| VIAF | 163,353 (84.0%) | |
| Open Library | 36,097 (18.5%) | 71,706 (36.8%) |
| Goodreads | 8,997 (4.6%) | 79,158 (40.7%) |

**Table 1.** Number of authors with an external identifier

each work it is possible to obtain all editions. Results include a set of useful publishing information, readers count, ratings, and number of ratings. Goodreads does not provide APIs, but allows for web scraping. Hence, we first collected the list of all works from writers pages, their ratings and number of ratings, then we obtained publishing information through Google Books APIs.

In order to emphasize the role of readers communities, we only kept works that had received at least one reception or that were marked as read by at least one user. Table 2 shows the number of works collected from each platform and the number of writers associated with at least one work from them. As it can be observed, Goodreads includes a higher number of works and writers with at least one work. Furthermore, both Open Library and Goodreads show a higher percentage of 'Transnational' writers than Wikidata: 12.6% and 11% against 8.6%.

| Source | N. of writers with $\geq 1$ works (% transn.) | N. of works |
|---|---|---|
| Wikidata | 22,515 (8.6%) | 117,798 |
| Open Library | 24,370 (12.4%) | 226,108 |
| Goodreads | 60,201 (11.0%) | 627,214 |
| Total | 71,443 (10.6%) | 971,120 |

**Table 2.** Number of works for each platform

The analysis of readers communities may also be observed through the lens of the number of interactions between readers and works. While Wikidata does not include users evaluation of literary works, it is possible to obtain this information from Goodreads and Open Library. Both expose the number of ratings and the average rating, while the latter also exposes the number of readers. Table 3 shows the number of interactions between readers and literary works in the two platforms. As it can be observed, absolute numbers are incomparable: there are 112.708 ratings in Open Library against 1.7 billions in Goodreads. The percentage of ratings about Transnational works is higher on Open Library (6%) than in Goodreads (4.9%), while both platforms show a slightly higher average rating of Transnational writers.

Summarizing, aligning literary facts from different platforms in a unique semantic resource allows for a richer representation of World Literature, with a more balanced knowledge about Transnational writers (+2% of them are associated with at least one work). Furthermore, such data collections shows the impact of communities of readers on the diffusion of writers and their works.

| Source | Average rating | N. of works | N. of readers |
|---|---|---|---|
| Open Library | 3.91 (3.99) | 112,708 (6.0%) | 1.2$M$ (8.5%) |
| Goodreads | 3.86 (3.77) | 1.7$B$ (4.9%) | – |

**Table 3.** Number of readers interactions in Goodreads and Open Library. Interactions about Transnational writers are reported in parenthesis.

## 5 Visualization Platform

The World Literature Knowledge Graph is built to support advanced queries and is seamlessly integrated with SKATEBOARD, the Semantic Knowledge Advanced Tool for Extraction Browsing Organization Annotation Retrieval and Discovery, providing users with an intelligent and intuitive way to explore the vast world of literature. With the World Literature Knowledge Graph and SKATEBOARD interface, our goal is to enable users to uncover deep insights and connections within literary works and enhance their understanding of the literary world. The SKATEBOARD platform presented in this research builds upon the work of Bernasconi et al. [4] and represents an extension and updated version of their work to fit our specific context of use. The interface features two main views: "Author" and "Work". The navigation flow that starts with an initial search for a topic of interest. Once a relevant topic is found, the user can drag the resource onto the central board and explore its relationships with other objects and predicates, creating a visual representation of the connections. This feature is illustrated in Figure 4.

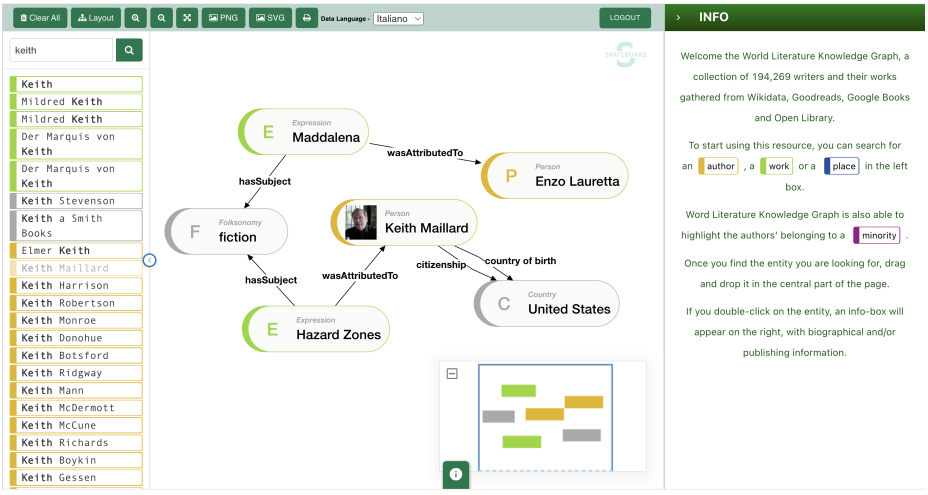



**Fig. 4.** A snapshot of the visualization platform. On the left, the search box; in the middle, the whiteboard where entities can be dragged; on the right, info pane about the selected entity.

By clicking on resources of type "Person" (as visible in fig. 5), the user can access information about an author, including both direct relationships such as published works and indirect relationships such as all the topics covered in their works, or a map of all the locations where their works were published. Clicking on resources of type "Expression" (as visible in fig. 6) displays information specific to a particular work, such as editions, languages, and readers ratings.

Literary searches may also start from different type of entities in the Knowledge Graph. It is possible to retrieve all writers by their country of birth or by their citizenship, as well as perform searches based on specific minorities (eg.: African Americans). The platform also allows navigations based on subjects: users can browse all works linked to a specific URB:FOLKSONOMY. The graph-based navigation encourages serendipitous discovery, allowing users to stumble upon unexpected connections and relationships.

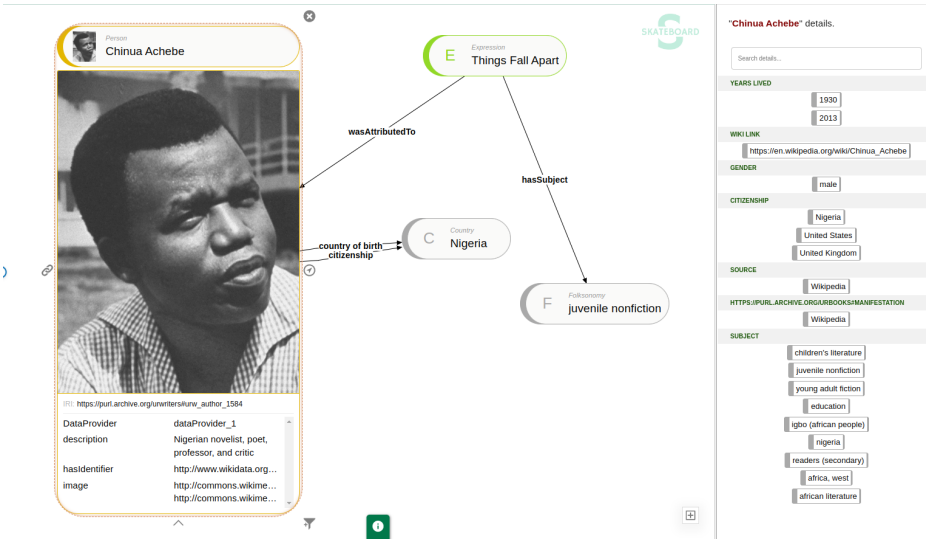



**Fig. 5.** Person view: on the left, the central area of the interface, where selected entities can be dragged for visualising their provenance and associated media and their relations with other entities according to the node-link paradigm (here, Chinua Achebe); on the right, the Info pane displaying the information about the entity (e.g., biographical dates, citizenship).

In summary, the visualization platform presented in this research offers an updated and customizable interface for exploring and visualizing relationships between topics, authors, and works, with potential applications in various research fields.

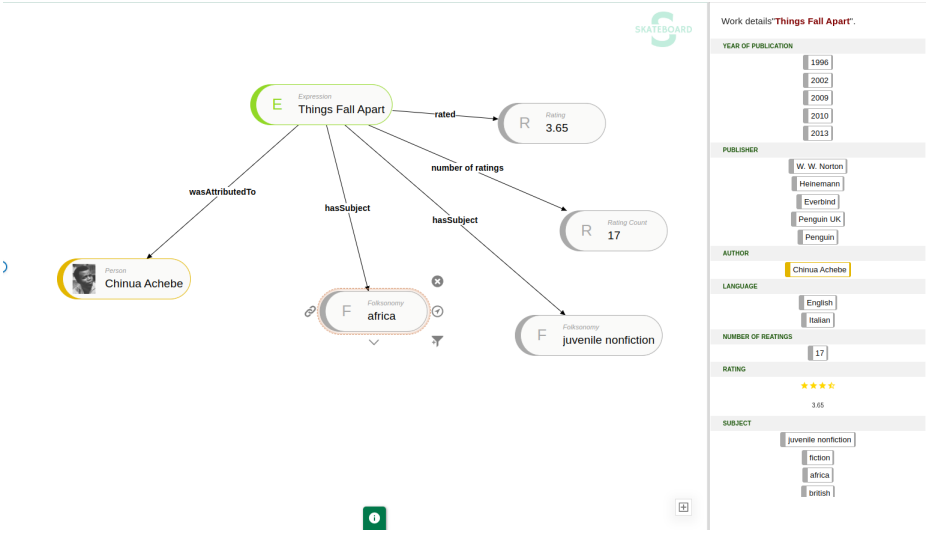



**Fig. 6.** Expression view: on the left, the central area of the interface where a work (top left, "Things Fall Apart") is connected with its author (Chinua Achebe, see Fig.5 ). On the right, the Info pane displaying the information about the work in tabular form (an Expression in FRBR terms), such as publisher, language, rating, etc.

## 6 Resource Evaluation

The current form of the WL-KG and its visualization platform are the result of a two-year interactive process of design and development carried out in constant interaction with domain experts. The contribution of domain experts to the process has been two-fold: on one side, they helped in defining the geopolitical and temporal boundaries of the resource, suggesting post-colonial studies as the conceptual reference framework for the study of underrepresentation; on the other side, they suggested that a graph-based visualization would be better suited to encourage exploration – and support their professional tasks – than the archival-based visual metaphors employed in the first prototype, for its capability to encourage the discovery of new authors through the connections displayed in the graphical interface.

For the evaluation of the WL-KG we organized a series of structured interviews with a group of potential targets of our resource, in line with the paradigm of user-centered design [41]: 4 teachers, 6 researchers in the humanities, and 3 professionals in the publishing industry. Each interview was articulated in two parts: the first part, targeted on the search of Transnational writers and works, was focused on the use of the platform; the second focused on the potential uses of the resource in the users' field of work and research.

**User Experience** After asking the users to search for at least one Transnational author and work of their choice, the user experience was investigated along three

dimensions: the usability of the platform, the completeness of the results, and the accuracy of the results.

Concerning the usability of the platform, most users experienced difficulties in navigating the WL-KG. First of all, they didn't realise that every element in the search area can be dragged into central whiteboard – according to the incremental paradigm that controls the interaction between the search area and the central whiteboard (as described in Section 5). Secondly, they failed to explore the information linked to the selected entity by expanding the relations between that entity and the other entities connected with it in the graph, which can be navigated in the whiteboard according the node-link paradigm (Section 5). Conversely, a minority of respondents who had already experience of Knowledge Graphs found the platform easy to use and appreciated the possibility of selecting the entities of interest by dragging them into the whiteboard, a function that they saw as a way to overcome the limitations of the standard navigation tools for graph-based representations. Based on these observations, we hypothesize that the difficulties in the use of the visualization platform reported in the interviews can be mainly attributed to the users' lack of experience with graph-based resources. For these users, the drag and drop selection of entities and the link-based navigation were not intuitive and can be improved by providing more guidance in the exploration (e.g., through tooltips, demo-mode navigation, etc.). This is in line with the comment made by some respondents who suggested to initialize the platform with an already loaded example. Concerning the entry of the search parameters, some users expressed their difficulty in finding a suitable author or work, motivating it with their limited knowledge of the domain. To bypass this difficulty, a user suggested to create a list of writers' names, indexed by country of birth, in a separated section of the site. We think that this suggestion is valuable, although it partly overlaps with the possibility of exploring the graph by starting from different types of entities (e.g., subjects, countries, topics), which is already available in the current version of the platform.

Concerning the completeness of the resource, a criticism derived from a misconception about its objectives shared by most respondents, who compared it with standard online archives, such as Wikipedia: the latter, being targeted at end users, include richer information about the entities in textual form, but are not suited for the development of applications that rely on the graph-based representations. This issue can be addressed by revising the description of the resource with a clearer definition of its intended usages. A more challenging request, then, emerged from the scholars in post-colonialism, who complained about some missing associations between works and subjects. It is the case of Andrea Levy's work 'The Long Song': although this book is about 'slavery', it is not linked to this subject in the KG, an issue derived from the lack of attribution of this subject within the digital sources from which data were gathered.

As for completeness, almost all respondents found the resource accurate, with a few errors that we could track from sources. For instance, 'Candide oder der Optimismus', namely the German translation of Voltaire's 'Candide', was attributed to Stephan Hermlin, its translator, due to an error propagated from

Goodreads. To address this issues, a functionality for signaling missing and wrong information will be added in a future version of the platform.

**Use Cases** The discussion of use cases was structured in two main parts: the comparison of the resource with the existing known archives and the collection of feedback about use cases and missing functionalities. Participants tended to rate the resource as useful for the discovery of new writers, but not useful for exploring new works. Such feedback reflects our data collection strategy, that was limited to the existing entities of the type writer on Wikidata and to the works that had received at least one reaction on the platforms where they are archived, aiming at relevance rather than completeness for what concerns works.

Interviews also showed that almost all respondents use general purpose archives like Google, Wikipedia, and Goodreads for the literary searches, showing a gap in the usage of knowledge bases designed for specific domains of application. The discovery of new literary facts has been pointed out as the major use case for all respondents. Interestingly, from the structured interviews with teachers, it emerged that the students themselves may be potential users of the platforms, since they could take advantage of subject-based search for supporting essay writing. Finally, it emerged the need of exposing in the knowledge base all the places where authors lived during their lives, in order to discover deeper connections between them.

## 7 Conclusion and Future Work

In this paper we presented the WL-KG, a knowledge base of writers and works designed for the discovery of literary facts from different parts of the world and exploring the underrepresentation of non-Western writers. The resource includes 194,346 writers and 971,120 works collected from Wikidata, Goodreads, and Open Library. The integration of knowledge from different sources had an impact on reducing the underrepresentation of Transnational writers, about whom there is more available information in Goodreads and Open Library than in Wikidata. Our resource also allows exploring how works are received by different communities of readers.

The WL-KG is publicly available through a graph-based visualization platform that simplify its usage by non-expert users. The resource and the visualization platform were evaluated by a group of professionals whose work may be supported by the KG. Their feedback shows that the platform may be useful especially to discover new writers from multiple kinds of entities: works, subjects, countries, minority groups. Respondents also highlighted the novelty of the platform compared to existing archives: the graph-based browsing experience, designed according the node-link paradigm, has been perceived as a valuable and alternative tool for exploring literary facts, even if its usability is not immediately intuitive, since graph-based resources are not widespread.

Future work will be devoted to improve the WL-KG with feedback emerged during the evaluation: we plan to increase the knowledge base with knowledge

from new communities of readers and thematic platforms; we also plan to release a new version of the visualization platform focused on improving its user experience for non-expert users. To do so, the platform will be adopted as a didactic tool in undergraduate and graduate courses that tackle the postcolonial aspects in World Literature. Finally, we plan to test a recommender system based on our knowledge graph, in order to test its impact in providing fairer recommendations.

## Acknowledgements

This work was partially supported by the PNRR projects "CHANGES: Cultural Heritage Active Innovation for Next-Gen Sustainable Society", CUP H53C22000860006, and "Fostering Open Science in Social Science Research (FOSSR)", CUP B83C22003950001.

## References

1. Adams, J., Brückner, H., Naslund, C.: Who counts as a notable sociologist on Wikipedia? Gender, race, and the "professor test". Socius **5**, 2378023118823946 (2019)
2. Benwell, B., Procter, J., Robinson, G.: Postcolonial audiences: Readers, viewers and reception. Routledge (2012)
3. Bernasconi, E., Ceriani, M., Mecella, M.: Linked data interfaces: a survey. In: Falcon, A., Ferilli, S., Bardi, A., Marchesin, S., Redavid, D. (eds.) Proceedings of the XIX Italian Research Conference on Digital Libraries - Information and Research Science connecting to Digital and Library Science (IRCDL2023). vol. 3365, p. 16. Central Europe (CEUR) Workshop Proceedings, Bari, Italy (23–24 February 2023)
4. Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T.: Design, realization, and user evaluation of the ARCA system for exploring a digital library. International Journal on Digital Libraries **24**(1), 1–22 (2023)
5. Bernasconi, E., Ceriani, M., Mecella, M., Morvillo, A.: Automatic knowledge extraction from a digital library and collaborative validation. In: Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G.M., Golub, K., Ferro, N., Poggi, A. (eds.) Linking Theory and Practice of Digital Libraries. pp. 480–484. Lecture Notes in Computer Science, Springer (2022)
6. Bolina, M.: Yewno Discover. Nordic Journal of Information Literacy in Higher Education **11**(1) (2019)
7. Boter, B., Rensen, M., Scott-Smith, G.: Unhinging the National Framework: Perspectives on Transnational Life Writing. Sidestone Press (2020)
8. Chawuthai, R., Takeda, H.: Rdf graph visualization by interpreting linked data as knowledge. In: JIST (2015)
9. Crenshaw, K.W.: On intersectionality: Essential writings. The New Press (2017)
10. Damrosch, D.: What is world literature?, vol. 5. Princeton University Press (2003)
11. Desimoni, F., Po, L.: Empirical evaluation of linked data visualization tools. Future Generation Computer Systems **112**, 258–282 (2020). https://doi.org/10.1016/j.future.2020.05.038

12. Fenlon, K., Kariuki, B., Welberry, A.: Researchspace: A platform for digital cultural heritage data management and linked data publication. Journal of Digital Humanities **6**(1) (2017)
13. Field, A., Park, C.Y., Lin, K.Z., Tsvetkov, Y.: Controlled analyses of social biases in Wikipedia bios. In: Proceedings of the ACM Web Conference 2022. pp. 2624–2635 (2022)
14. Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., Trilcke, P.: Programmable corpora: Introducing dracor, an infrastructure for the research on european drama. Digital Humanities **2019**, 5 (2019)
15. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with dolce. In: Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web: 13th International Conference, EKAW 2002 Sigüenza, Spain, October 1–4, 2002 Proceedings 13. pp. 166–181. Springer (2002)
16. Haag, F., Lohmann, S., Siek, S., Ertl, T.: Queryvowl: A visual query notation for linked data. In: Proceedings of ESWC 2015 Satellite Events. LNCS, vol. 9341, pp. 387–402. Springer (2015)
17. Heim, P., Lohmann, S., Stegemann, T.: Interactive relationship discovery via the semantic web. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30–June 3, 2010, Proceedings, Part I. Lecture Notes in Computer Science, vol. 6088, pp. 303–317. Springer (2010)
18. Hube, C., Fischer, F., Jäschke, R., Lauer, G., Thomsen, M.R.: World literature according to wikipedia: Introduction to a dbpedia-based framework. arXiv preprint arXiv:1701.00991 (2017)
19. Jauss, H.R., Benzinger, E.: Literary history as a challenge to literary theory. New literary history **2**(1), 7–37 (1970)
20. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. W3C Recommendation, World Wide Web Consortium, United States (2013)
21. Lohmann, S., Link, V., Marbach, E., Negru, S.: Webvowl: Web-based visualization of ontologies. In: EKAW (2014)
22. Moretti, F.: Conjectures on world literature. New left review **1**, 54 (2000)
23. Nakamura, L.: "Words with friends": socially networked reading on Goodreads. Pmla **128**(1), 238–243 (2013)
24. O'Donnell, D.P., Walter, K.L., Gil, A., Fraistat, N.: Only connect: the globalization of the digital humanities. A new companion to digital humanities pp. 493–510 (2015)
25. Pfeffer, M., Roth, M.: Japanese visual media graph: Providing researchers with data from enthusiast communities. In: International Conference on Dublin Core and Metadata Applications. pp. 136–141 (2019)
26. Po, L., Bikakis, A., Desimoni, F., Papastefanatos, G.: Linked Data Visualization: Techniques, Tools, and Big Data, vol. 10. Morgan & Claypool Publishers (2020)
27. Polley, S., Ghosh, S., Thiel, M., Kotzyba, M., Nürnberger, A.: Simfic: An explainable book search companion. In: 2020 IEEE International Conference on Human-Machine Systems (ICHMS). pp. 1–6. IEEE (2020)
28. Rajpurkar, S., Bhatt, D., Malhotra, P., Rajpurkar, M., Bhatt, M.: Book recommendation system. International Journal for Innovative Research in Science & Technology **1**(11), 314–316 (2015)
29. Ratcliff, J.W., Metzener, D., et al.: Pattern matching: The Gestalt approach. Dr. Dobb's Journal **13**(7), 46 (1988)

30. Schöch, C., Eder, M., Odebrecht, C., Kestemont, M., Primorac, A., Tonra, J., Poniž, K.M., Kanellopoulou, C.: Distant reading for european literary history. a cost action. Proceedings of DH2018 (2018)
31. Simpson, J., Brown, S.: From xml to rdf in the orlando project. In: 2013 International Conference on Culture and Computing. pp. 194–195. IEEE (2013)
32. Spivak, G.C.: Can the subaltern speak? Die philosophin **14**(27), 42–58 (2003)
33. Stranisci, M.A., Damiano, R., Mensa, E., Patti, V., Radicioni, D., Caselli, T.: WikiBio: a semantic resource for the intersectional analysis of biographical events. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12370–12384. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.acl-long.691
34. Stranisci, M.A., Patti, V., Damiano, R.: User-generated world literatures: a comparison between two social networks of readers. In: Proceedings of IRCDL 2023. vol. 3365, pp. 38–46 (2023)
35. Stranisci, M.A., Patti, V., Damiano, R., et al.: Representing the under-represented: A dataset of post-colonial, and migrant writers. Proc. of the 3rd Conference on Language, Data and Knowledge (LDK 2021), Open Access Series in Informatics **93**, 1–14 (2021)
36. Tillett, B.B.: Frbr and cataloging for the future. Cataloging & classification quarterly **39**(3-4), 197–205 (2005)
37. Troullinou, G., Kondylakis, H., Daskalaki, E., Plexousakis, D.: Rdf digest: Efficient summarization of rdf/s kbs. In: ESWC. Lecture Notes in Computer Science, vol. 9088, pp. 119–134. Springer (2015)
38. Van Remoortel, M., Birkholz, J.M., Alesina, M., Bezari, C., D'Eer, C., Forestier, E.: Women editors in europe. Journal of European Periodical Studies **6**(1), 1–6 (2021)
39. Viola, F., Roffia, L., Antoniazzi, F., D'Elia, A., Aguzzi, C., Salmon Cinotti, T.: Interactive 3d exploration of rdf graphs through semantic planes. Future Internet **10**(8) (2018)
40. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
41. Wood, L.E.: Semi-structured interviewing for user-centered design. interactions **4**(2), 48–61 (1997)