

WikiBias: a Framework to Explore Bias in the Wikimedia Ecosystem through Semantic Modeling and Biographical Event Extraction

Marco Antonio Stranisci^{a,*}, Mirko Lai^b, Viviana Patti^a and Rossana Damiano^a

^a *Department of Computer Science, University of Turin, Italy*

^b *Department of Computer Science, Università del Piemonte Orientale, Italy*

E-mails: marcoantonio.stranisci@unito.it, mirko.lai@uniupo.it, viviana.patti@unito.it, rossana.damiano@unito.it

Abstract. The presence of bias in Wikimedia projects has a potential impact in the lack of fairness of Artificial Intelligence technologies trained on this source of knowledge. However, research on this topic is fragmented, failing to address the complexity of this phenomenon that impacts against minorities in different ways. In this paper we present WikiBias: a framework for exploring bias in the Wikimedia ecosystem through biographical event extraction. WikiBias is designed to jointly study underrepresentation and *representational* bias, providing a multi-dimensional overview of the sources of harms against people vulnerable to discrimination. We test WikiBias on the case study of writers in Wikidata and Wikipedia, given the crucial role of literature in the definition of identity and otherness in our society. We adopt an intersectional perspective, considering the joint impact of writers' gender and origin in the perpetration of bias against them. Our results show that biographical event extraction can be effective in reducing the underrepresentation of writers with a non-Western origin but at the same time it might induce *representational* bias, especially against women. This knowledge augmentation has a dramatic impact in increasing the connections of writers with other people in Wikidata, potentially facilitating the discovery of underrepresented writers in the augmented knowledge base.

Keywords: Semantic Modeling, Biographical Event Detection, Bias Detection, Wikipedia, Wikidata

1. Introduction

With thousands of contributors, the Wikimedia ecosystem is the largest example of participatory knowledge creation, widely consumed by people all over the world. Content produced by the Wikimedia community not only plays a crucial role in how people access knowledge, but also in the Artificial Intelligence (AI) sector. A large number of AI-based technologies are developed by leveraging the knowledge stored in Wikimedia projects. Wikipedia pages have been extensively used to train the first generation of Language Models (LM) [28], to filter out poor-quality documents from pretraining [19] and to benchmark LMs across several QA tasks [57]. Similarly, Wikidata is often used as a knowledge base for Recommendation Systems [101], Information Retrieval [21], and Natural Language Generation [4].

*Corresponding author. E-mail: marcoantonio.stranisci@unito.it.

1 However, the adoption of Wikipedia and Wikidata in such a broad range of AI tasks and technologies 1
2 is not without flaws. It has been demonstrated that these projects are systematically affected by different 2
3 types of bias. For instance, editors' demographic is skewed towards Western white men [49]; Wikipedia 3
4 pages about women are affected by stereotypical gender roles [95]; many groups vulnerable to discrimi- 4
5 nation are highly underrepresented in these projects [2]. Nonetheless, research works that jointly study 5
6 multiple forms of bias in Wikimedia projects are still missing [12]. This hinders the thorough analysis of 6
7 the factors that induce bias in these sources of knowledge. 7

8 To fill this gap we present WikiBias: a framework for exploring the presence of biases in the Wikimedia 8
9 ecosystem by combining Semantic Web (SW) and Information Extraction (IE) technologies. The framework 9
10 is characterized by two components: *i.* the People in the Media Ontology (PiM-O), which models knowledge 10
11 from Wikidata to enable the discovery of underrepresented groups in this knowledge base. We adopt an 11
12 intersectional perspective [23] to analyze the interaction of gender and origin in the perpetration of biases 12
13 against minorities. *ii.* A biographical event extraction pipeline that automatically extracts biographical 13
14 triples from English Wikipedia pages. We investigate to what extent biographical extraction helps reduce 14
15 the underrepresentation of minorities in Wikidata. At the same time, we analyze whether this approach 15
16 induces *representational* harms against these groups. 16

17 As a case study of WikiBias we explore the representation of writers in Wikidata and Wikipedia. 17
18 Wikidata and Wikipedia are an unprecedented point of access to World Literature [24, 50], enabling the 18
19 discovery of authors and works from different parts of the world. However, the literary system emerging 19
20 from these sources of knowledge is biased towards men and people of Western origin [92]. Through the 20
21 implementation of our framework, we systematically detect these biases and identify strategies to mitigate 21
22 them. Specifically, WikiBias addresses the following research questions: 22

23 **RQ1. Is it Possible to Reduce Underrepresentation through Biographical Event Extraction?** Starting from 23
24 a set of 105,951 writers, we extract biographical triples from all the English Wikidata pages and compare 24
25 this extracted knowledge with the existing available information in Wikidata. Results show a dramatic 25
26 increase of triples about writers. Despite the augmentation is beneficial for people that have a non-Western 26
27 origin, it reduces the relative distribution of women in the dataset. 27

28 **RQ2. Does Biographical Event Extraction Induce Representational Bias?** We compare the distribution 28
29 of biographical events extracted from Wikipedia pages to analyze whether they portray stereotypical 29
30 representation of certain socio-demographic groups. Results show that women are more affected by *rep-* 30
31 *resentational* bias since they are mostly associated with events related to parenthood and marriage. The 31
32 intersection of gender and origin mitigates this bias, though. 32

33 **RQ 3. Does Biographical Triple Extraction Facilitate the Discovery of Minorities?** We perform a Network 33
34 Analysis (NA) on the writers gathered from Wikidata before and after the knowledge augmentation to 34
35 investigate whether minorities were better connected with the rest of the network. Results show that the 35
36 augmentation generally contributes to create a more connected network of writers and increases the links 36
37 between members of different socio-demographic groups. 37

38 The paper is organized as follows. In Section 2 we present previous work on bias in Wikimedia projects 38
39 and on biographical event detection. In Section 3 we describe our semantic modeling approach that led to 39
40 the creation of the People in the Media Ontology and to the People in the Media Knowledge Graph. In 40
41 Section 4 we present our biographical triple extraction pipeline. Section 5 is devoted to analyze the impact 41
42 of our framework on the presence of biases in Wikidata and Wikipedia. 42
43

44 2. Related Work 44

45
46
47
48 Research on bias is broad, inter-sectorial, and encompasses several types of bias [96]. According to Hovy 48
49 and Prabhumoye [47], bias can be propagated by models behavior [11], emerge from datasets [32], or be 49
50 the result of the design of the research [81]. On another perspective, Barocas et al. [9] and Blodgett et al. 50
51 [12] classify bias types on the basis of harm they produce: *representational* bias are about the association 51

1 of categories of people to stereotypical feature, *allocative* bias refers to the unequal distribution of oppor- 1
2 tunities among social groups. In this section, we survey existing work on bias in datasets, distinguishing 2
3 between *representational* and *allocative* harms, in order to position our framework at the intersection 3
4 between these two phenomena. 4

5 Works on *representational* bias analyze how people are represented in Wikipedia biographies and how 5
6 specific socio-demographic groups are discriminated in this representation. Sun and Peng [95] performed an 6
7 event extraction task on 10,412 biographies, showing that women’s Wikipedia pages contain more personal 7
8 events than men’s, while the latter biographies are more focused on events related to their career. Lucy 8
9 et al. [59] used Wikipedia occupation pages as a knowledge base to create person embeddings that are used 9
10 to investigate how professions are characterized in biographical pages. Stranisci et al. [91] trained a system 10
11 for biographical event extraction and leveraged it to explore the presence of bias in Wikipedia’s writers’ 11
12 biographies along the axis of gender and origin. Gaut and Sun [40] developed the WikiGenderBias corpus 12
13 to explore the presence of biases in Relation Extraction (RE). The corpus is a selection of biographies with 13
14 metadata derived from DBpedia [5] and results show that RE classifiers are better at extracting relations 14
15 involving men than women. Stranisci et al. [93] performed a similar experiments on a corpus of 92 relation 15
16 types extracted from Wikidata [102] and the REBEL corpus [21], discovering that RE classifiers trained 16
17 on Wikipedia systematically overlook women and non-Western people in the extraction of career-related 17
18 triplets. Outside the NLP community, Field et al. [34] provided a method for systematically extracting and 18
19 comparing biographies of different groups of people (e.g., by gender, ethnicity, sexual orientation) using 19
20 Wikipedia categories to generate samples. They further provided a multi-dimensional index (including 20
21 measures such as the average length of biographies and the number of languages for each biography) for 21
22 performing this comparison, finding that differences exist between biographies based on these groups. 22

23 Works on *allocative* bias are mainly devoted to explore the quantitative underrepresentation of specific 23
24 categories of people in the Wikimedia ecosystem. [105] provided a thorough study of famous people on 24
25 Wikipedia, showing that women and non-Western people are less likely to appear in a relevant number 25
26 (25) of Wikipedia language editions. Among the 100 most popular biographies, then, only 3 are about 26
27 women, and 8 about non-Western people. [2] studied sociologists’ Wikipedia pages, finding that non-white 27
28 male and female sociologists are more prone to under-representation. [83] performed a comparison between 28
29 the number of software developers, engineers, and scientists on Wikidata and the real-world population, 29
30 showing that people from Europe and North America are over-represented. Stranisci et al. [92] discovered 30
31 that writers with non-Western origin are highly underrepresented in Wikidata but aligning this archive with 31
32 knowledge from other communities of readers reduces this underrepresentation. Weathington and Brubaker 32
33 [104] provided an ethnographic study on the representation of queer people in Wikidata leveraging the 33
34 properties ‘sexual orientation’ (P91) and ‘gender’ (P21). From the analysis both *representational* and 34
35 *allocative* bias emerge. Das et al. [27] systematically studied how underrepresentation of people in Wikidata 35
36 propagates to systems for Link Prediction based on KG-embeddings. 36

37 Our paper introduces a novel framework that jointly analyzes *representational* and *allocative* bias in 37
38 Wikipedia and Wikidata adopting a multidisciplinary and inter-sectorial approach. The framework lever- 38
39 ages a semantic model for the representation of people, and an Information Extraction pipeline for the 39
40 extraction of biographical triplets. The framework adopts an intersectional approach [23] that enables the 40
41 identification of finer-grained forms of discrimination in the Wikimedia ecosystem and in any other digital 41
42 archives by observing how the intersection of socio-demographic features (e.g., gender and race) amplifies 42
43 the stereotypical representation of individuals or reduces their visibility. Not only that, the framework 43
44 enables the exploration of potential strategies to mitigate bias against vulnerable groups to discrimination 44
45 in Wikipedia and Wikidata. 45

46 47 48 **3. The People in the Media Knowledge Graph** 48

49 50 In this section we present the People in the Media Knowledge Graph (PiM KG), a knowledge base that 50
51 supports the analysis of bias against women and Transnational writers in Wikidata. In Section 3.1 we 51

1 present our definition of underrepresentation that relies on existing post-colonial studies. In Section 3.2
 2 we present the People in the Media Ontology (PiM-O), which has been used as a semantic model to create
 3 our KG. Section 3.3 describes the creation of the KG and reports some statistics about writers.

4 3.1. Defining Underrepresentation 5

6
 7 The social and cultural underrepresentation of people who are not born in Europe or in North America
 8 has been analyzed by different perspectives that gave place to a number of definitions and taxonomies
 9 of such underrepresentation. Post-Colonial studies [79] emphasize the cultural take over of African and
 10 Asian countries by colonizers achieved through the education of local elites according to Western principles
 11 [86]. This definition led to disagreement between scholars on the status of historical figures [25] related to
 12 the Decolonization process such as Gandhi, who studied law in England. The distinction between Global
 13 North and Global South, introduced by the Willy Brandt Commission [15], focuses on economic dispari-
 14 ties between developed and under-developed countries. Such a distinction, still adopted by international
 15 bodies¹, does not reflect the economic changes have occurred in the last decades. For instance, China is
 16 considered a Global-South country despite being one of the World leading economies. Coined within the
 17 World Literature framework [29], the term Transnational refers to people who “operated outside their own
 18 nation’s boundaries, or negotiated with them” [14]. Although this definition does not directly engage with
 19 social inequalities, its emphasis on the circulation of people and their works in a globalized world [103]
 20 makes it relevant for the objectives of our research.

21 Our modeling of underrepresentation is built upon the three definitions that we presented above and
 22 tied to the relation between writers and their countries of birth. We classified countries along three axes.
 23 **The historical axis** follows Post-Colonial studies and distinguishes between countries that have been for-
 24 mer colonies and countries that have not. **The economic axis**, which takes inspiration from the Global
 25 North/Global South distinction, ranks countries on the basis of their Human Development Index (HDI)².
 26 We consider Global North countries all those that score an HDI greater or equal than 0.8 [87] and Global
 27 South countries the others. **The mobility axis**, inspired by World Literature studies, ranks countries based
 28 on the mobility score of their passport³, which measures the number of countries that a person with a
 29 given passport can visit without applying for a VISA.

30 The classification of writers derives from the taxonomy of countries: we adopt the terms **Transnational**
 31 and **Western-educated** to identify the two groups and we consider a writer as Transnational if they meet
 32 any of the following conditions:

- 33 1. They were born in a country that has been a former colony and has an HDI below 0.8 or a mobility
 34 score lower than the average.
- 35 2. They belong to an ethnic minority in a Western country.
- 36 3. They were not born before the beginning of the Spanish-American war (1808) if their birth country
 37 is located in Latin America and Caribbean, or before the beginning of the Decolonization process
 38 (1917) if their birth country is located in Africa or Asia. This criterion removes from the group of
 39 Transnational writers those who were born in a former colony from European parents.

40 3.2. Encoding Biographical Knowledge 41

42
 43 Providing a formal representation of biographical knowledge is a key aspect to better understand how
 44 people are represented in media and social media. Since most of the information that can be found online
 45 about people is unstructured, the definition of a semantic model for the storage, alignment, and comparison
 46 of biographical knowledge obtained through Information Extraction (IE) methods is crucial.
 47

48 ¹https://www.ilo.org/sites/default/files/wcmsp5/groups/public/%40dgreports/%40exrel/documents/publication/wcms_907091.pdf

49 ²<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI>

50 ³<https://www.passportindex.org/>

We designed the People in The Media Ontology (PiM-O), an overarching ontology designed to represent the multifaceted way in which people can be represented. Based on the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [39] and on the Ontology Design Pattern (ODP) framework, PiM-O describes people’s biographical events as situations in which people participate with specific roles. Specifically, the biographical situation (PIM:BIOGRAPHICALSITUATION) is a subclass of a DUL:SITUATION that requires at least the presence of one person (PROV:PERSON) associated with at least one role (DUL:ROLE) that represents the social condition of the person within the situation itself. The same ODP enables the specification of additional elements, such as spatio-temporal information or the presence of other entities. For instance, the Wikidata triple that encodes the marriage between Chinua Achebe and Christie Ahinwe Okoli-Achebe in 1961, ‘wd:Q155845 wdt:spouse wd:Q109321135’, is represented in PiM as follows:

```
pim:marriage a pim:BiographicalSituation;
dul:isSettingFor wd:Q155845, wd:Q109321135, pim:husband, pim:wife, pim:1961 .
wd:Q155845 a prov:Person; rdfs:label ‘Chinua Achebe’; dul:hasRole pim:husband .
wd:Q109321135 a prov:Person; rdfs:label ‘Christie A. Okoli’; dul:hasRole pim:wife .
pim:husband a dul:Role .
pim:wife a dul:Role .
pim:1961 a time:TimeInterval .
```

Since the ontology represents people as entities holding specific roles in a given situation, it is suited to encode complex interactions and divergent representations of people emerging from language. The following sentence extracted from Chimamanda Ngozi Adichie’s Wikipedia biography⁴ contains two roles (migrant and student) associated to the same person in the same biographical situation: “[She] **moved** to the United States, to **study** communications at Drexel University in Philadelphia, Pennsylvania.”. The encoding of the sentence can be observed in the following KG snapshot:

```
pim:migration a pim:BiographicalSituation;
dul:isSettingFor wd:Q230141 .
wd:Q230141 dul:hasRole pim:migrant, pim:student ;
rdfs:label Chimamanda Ngozi Adichie .
```

3.3. Gathering Knowledge from Wikidata

Equipped with the semantic model, we were able to build a knowledge base of people derived from Wikidata: the People in the Media Knowledge Graph (PiM-KG). PiM-KG is a collection of 10.8 million entities of type human (WD:Q5) gathered from a Wikidata dump⁵. Each entity has been encoded according to our ontology and information about their country of birth, citizenship, gender, ethnic minority, and ethnic minority have been used to categorize them as Transnational or Western-educated. Figure 1 shows an example of SPARQL query from the PiM-KG endpoint⁶.

Given the scope of our research, we identified a subset of the KG consisting of writers, by selecting all people linked to the occupation of writer (WD:Q36180) (or one of its subclasses), resulting in a set of 273,130 entities. For our biographical triple extraction experiment we kept only writers with an English Wikipedia page, obtaining a set of 105,951 entities, 93,175 of which also have information on gender. As it can be observed in Figure 2, the collection is highly skewed towards Western-educated people and men. Among the collected writers, 79,943 are categorized as Western-educated and 13,232 as Transnational; 67,237 are men, 25,934 are women, and 301 are non-binary.

⁴https://en.wikipedia.org/wiki/Chimamanda_Ngozi_Adichie

⁵<https://academictorrents.com/details/7bee8ece634c55ab4ed7da5a56dd81578729ed2b>

⁶<https://kgccc.di.unito.it/sparql/wikibias>

```

1 PREFIX pim: <http://purl.archive.org/domain/people-in-the-media
2 #>
3 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4 SELECT ?authorName ?situation ?country
5 WHERE {
6   ?situation a pim:BiographicalSituation ;
7   pim:isSettingFor ?author .
8   ?author pim:countryOfBirth ?country .
9
10  FILTER (?country=pim:NGA)
11 }
12 LIMIT 10
13

```

Figure 1. An example of SPARQL query from the PiM-KG endpoint.

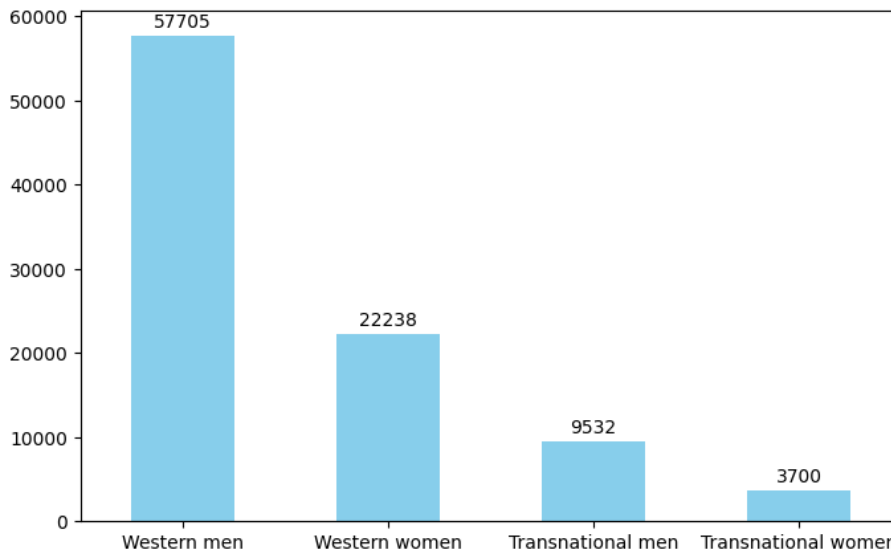


Figure 2. The distribution of writers in Wikidata broken down by socio-demographic groups

4. Biographical Triple Extraction

In this section we present the Information Extraction (IE) approach that we implemented to explore biases against Transnational writers. The approach combines Biographical Event Detection, Named Entity Recognition and Entity Linking (EL) to extract triples about writers from Wikipedia and encode them in the PiM-KG.

In line with with our research aims, which focus on connections between people, we limited the scope of our approach to the Wikidata triples characterized by the joint presence of people as subject and object of the triple, and thus connected through relations of the type: child (P40), family (P53), influenced by (P737), parent (P8810), partner in business or sport (P1327), relative (P1038), sibling (P3373), spouse (P26), student (P802), unmarried partner (P451). Given a writer’s Wikipedia page, for each sentence we extracted a biographical triple every time we jointly detected a biographical event (or state) and a mention of a person who is included in PiM KG. Note that the latter did not need to be of type writer.

In this section we first describe our event detection approach (Section 4.1), then the EL step (Section 4.2).

4.1. Event Detection

Although the meaning of Biographical Event Detection is clear, a stable formalization of this concept for computational purposes has not yet been established. In the context of this work, we define Biographical Event Detection as an entity-based detection task: given a biography, an IE model must be able to identify all the events related to the entity which is the target of the biography. To perform this task, we adopt the methodology elaborated by Stranisci et al. [91], who developed an *ad-hoc* resource for biographical event detection, WikiBio. We then combined WikiBio with the following corpora for event detection to train a system for this task:

- TimeBank [73], a corpus of events and temporal relationships annotated according to the TimeML scheme [74];
- OntoNotes [48], a multi-layered corpus annotated for Named Entity Recognition and Semantic Role Labeling [54];
- LitBank, a corpus of literary documents annotated with the TimeML scheme;
- NewsReader, a corpus of annotated events and named entities for the task of entity-centric event detection across documents.

WikiBio is a corpus composed of 1,691 annotated sentences from 20 Wikipedia biographies of African and African-American writers. The annotation was made at a token level, an approach which is widely adopted in existing resources and is well-suited for extracting fine-grained knowledge about biographical events. The annotation of events in WikiBio is mainly based on TimeML [74] and RED [71] guidelines, according to which the term ‘event’ must be understood “a cover term for situations that happen or occur”. Therefore, we annotated as events all the tokens that express an event or a state, without providing distinction between them. Example 1 shows an example in which the events ‘married’ and ‘known’ have been annotated as relevant.

1. He **married** Wendy Bruce, whom he had **known** since they were teenagers

Given the relatively small size of WikiBio, we composed a number of training datasets in which the corpus was combined with other resources annotated for biographical event detection. In total, we created 16 different training sets of 4,492 items by selecting sentences from the corpora mentioned above (OntoNotes, Newsreader, LitBank, and TimeBank). For each corpus, we provided a training set with and without adding texts from WikiBio. In addition, we created three miscellaneous training sets containing different miscellanea of existing corpora with and without texts from WikiBio. The 1,691 sentences containing events annotated in our corpus were split into three sets of equal size that were used for development (563), testing (564), and training (564).

4.1.1. Evaluation of event detection.

The upper part of Table 1 shows the results of the model trained on each of the 16 training sets after five epochs, while in the lower part it is possible to observe the results of the models trained on the four training sets that led to a better performance. As it can be noticed, the latter all achieved an F-score on the test set above 0.85. In particular, the training set composed of sentences from Timebank and WikiBio achieved the best result (0.859), but it is worth mentioning that the model trained on a miscellany of all corpora and WikiBio shows the lowest drop in performance between the F-scores obtained on the training and test sets. This may signal a higher generalization of model’s prediction to other types of texts.

4.2. Entity Linking

We implemented an Entity Linking (EL) approach to retrieve all the events connecting each writer to other named entity. This enabled the extraction of biographical triples aimed at increasing the number of connections between the target entities and other entities included in PiM.

Training Dev Test (5 EPOCHS)	F-Score_train	F-Score_dev	F-Score_test
WikiBio WikiBio WikiBio	0.479	0.479	0.479
Litbank WikiBio WikiBio	0.847	0.640	0.622
Litbank + WikiBio WikiBio WikiBio	0.835	0.814	0.813
Misc_01 WikiBio WikiBio	0.885	0.863	0.801
Misc_01 + WikiBio WikiBio WikiBio	0.871	0.831	0.827
Misc_02 WikiBio WikiBio	0.866	0.816	0.819
Misc_02 + WikiBio WikiBio WikiBio	0.861	0.837	0.832
Misc_03 WikiBio WikiBio	0.850	0.811	0.817
Misc_03 + WikiBio WikiBio WikiBio	0.844	0.839	0.831
Onto WikiBio WikiBio	0.950	0.800	0.790
Onto + WikiBio WikiBio WikiBio	0.936	0.873	0.809
Onto_mod WikiBio WikiBio	0.997	0.823	0.814
Onto_mod + WikiBio WikiBio WikiBio	0.888	0.869	0.829
Timebank WikiBio WikiBio	0.89	0.801	0.790
Timebank + WikiBio WikiBio WikiBio	0.865	0.856	0.821
NewsReader WikiBio WikiBio	0.453	0.479	0.479
NewsReader + WikiBio WikiBio WikiBio	0.467	0.479	0.479
Training Dev Test (15 EPOCHS)	F-Score_train	F-Score_dev	F-Score_test
Misc_01 + WikiBio WikiBio WikiBio	0.890	0.852	0.853
Misc_02 + WikiBio WikiBio WikiBio	0.900	0.855	0.856
Misc_03 + WikiBio WikiBio WikiBio	0.896	0.859	0.855
Timebank + WikiBio WikiBio WikiBio	0.919	0.850	0.859

Table 1

Results of the event detection experiments. The 10 training sets are based on samples of documents selected from each of the four corpora used, with and without the addition of data from WikiBio. Additionally, we created three miscellaneous corpora: Misc_01 (all 4 resources); Misc_02 (OntoNotes, Timebank, and Litbank); Misc_03 (OntoNotes, and Timebank). The bottom part of the Table focuses on the 4 models that achieved the best F-scores after five epochs.

To adopt a computationally efficient approach for large-scale data, we designed a pipeline that combines neural network methods for named entity detection with the heuristics validated in our previous work [92] to link them to the dataset. The pipeline is organized in three steps.

1. To detect all the named entities of the type PERSON in all the sentences including at least one biographical event (Section 4.1) we used the largest SpaCy model.⁷
2. To retrieve the titles of the Wikipedia pages about the entities detected in the previous step we queried Wikipedia APIs⁸ for all the retrieved named entities, then filtered out all the named entities that were highly dissimilar from Wikipedia titles. Our heuristic [92] was based on gestalt pattern similarity [75], which determines the similarity of two strings by computing the ratio between the matching characters multiplied by 2 and the total number of characters of both strings. The result is a score between 0 and 1. We computed the gestalt pattern similarity for each candidate and filtered out all candidates with a score below 0.8. An example of named entity that has not been kept for the linking is ‘Alisson’, which is associated to ‘Alisson Becker’ with a score of 0.66, while ‘Kitaro’, which is associated to the entity ‘Kitarō’ with a score of 0.83, was kept for the EL step.
3. We retrieved all the Wikidata IDs corresponding to the Wikipedia titles by querying again the APIs, and kept only the entities with a Wikidata ID that is present in our knowledge base.

⁷<https://spacy.io/>, en_core_web_sm

⁸<https://www.mediawiki.org/wiki/API:Search>

As an example of extracted biographical triples, we consider again Example 2:

2. ‘Sehnaoui currently **lives** and **works** in Beirut with husband Marwan, President of the Lebanese Order of Malta and sons **Salim Sehnaoui** and Khalil Sehnaoui’

In the first step of the pipeline we identified two biographical events (‘lives’ and ‘works’). In the second step, we identified ‘Salim Sehnaoui’ as a named entity that is included in PiM-KG.

The resulting biographical pattern is the following:

```
pim:Lives a pim:BiographicalSituation;
dul:isSettingFor wd:Q21932336 , wd:Q22121200;
prov:wasDerivedFrom pim:Wikipedia;
dc:identifier ‘https://en.wikipedia.org/wiki/Mouna_Bassili_Sehnaoui’ .
wd:Q21932336 rdfs:label Mouna Bassili Sehnaoui .
wd:Q22121200 rdfs:label Salim Sehnaoui .
```

For the extraction of biographical triples, we applied the triple extraction pipeline to the Wikipedia pages of 105,951 writers. From these, we identified all biographical relations between the page subject and other entities of type person included in the PiM Knowledge Graph. A biographical triple was extracted from a sentence only when both a biographical event and a person were simultaneously detected. As a result, we obtained a set of 583,578 biographical events. Among these, 156,052 (26.74%) represent relations between writers, while 427,526 (73.26%) link writers to non-writer entities. This confirms that the majority of biographical relations involve individuals outside the subset of writers. The average distribution among writers reveals a mean of 8.05 relations per author, with a median of 4 and a standard deviation of 16.19, indicating a highly skewed distribution where a minority of individuals are connected to a substantially larger number of others. This pattern is consistent with findings in social network analysis, particularly in scale-free networks [8] and scientific collaboration networks [70], where degree distributions often follow a power-law and a small number of highly connected nodes dominate the structure.

4.2.1. Evaluation of Entity Linking

In order to assess the EL pipeline, we performed a manual evaluation of the extracted biographical triples. We sampled 1,000 sentences containing at least one biographical event and the mention of a named entity who is not the target of the biography. To select the sample we chose 10 sentences for each of the 100 most frequent events extracted in our pipeline, which correspond to 10.92% of the total number of triples extracted. We split the sample into three subsets and assigned each to one member of our team.

The evaluation method we adopted consists of two steps: *i*. Identify whether the extracted knowledge is correct; *ii*. Select the spans of events and mentions that are relevant to this identification within the sentence.

As can be observed in Table 2, 84.5% of the extracted triples have been evaluated as correct by annotators. The good accuracy is also demonstrated by the moderate variation in the evaluation performed by annotators which ranges between 81.2% of Annotator_2 and 90.45% of Annotator_3.

From an analysis of the examples marked as incorrect, some patterns emerge. The first recurrent error is caused by named entities disambiguation. In several cases, the subject of a biography is mentioned differently from its label on Wikidata, which can lead to linking the mention to a different entity. This is the case for the following sentence, extracted from Iris Dexter’s biography:⁹ “While in the role of war correspondent, her pen-name journalist **Margot Parker** introduced readers to another reporter”. A second type of error is the lack of a direct relation between the writer and another person in the same sentence. This occurs when the writer is not mentioned or is referred to indirectly in a sentence, as illustrated by the following example from the biography of Himanshi Shelat¹⁰, where instead of the writer another entity is

⁹https://en.wikipedia.org/wiki/Iris_Dexter

¹⁰https://en.wikipedia.org/wiki/Himanshi_Shelat

Annotator	n. of sentences	% of correct sentences
1	250	82%
2	400	81.2%
3	400	90.4%
Total	1,050	.845

Table 2

Results of the manual evaluation of the Entity Linking pipeline

mentioned: “Vinod Meghani died on 15 February 2009”. Finally, there are some cases in which two entities are mentioned in the same sentence (e.g., Orville and Wilbour Wright) but only one of them is recognized.

A qualitative analysis of results shows that a great number of triples extracted through our pipeline do not align with Wikidata’s taxonomy. For instance, in the following sentence extracted from Phan Bội Châu biography¹¹ “Liang **introduced** Phan to many prominent politicians, including **Ōkuma Shigenobu**” the link between the writer and the other person is the event ‘introduced’, which has no correspondence among existing Wikidata properties. Similarly, our approach extracted a property of the type ‘interview’ that could link Wendy Williams¹² to Whitney Houston in “Media outlets have described Williams’s 2003 **interview with Whitney Houston** as her most infamous”.

The systematic extraction of relations that are not present in Wikidata’s taxonomy can be useful to identify and reduce the knowledge gaps that affect this semantic model, providing a more complete representation of people in Wikidata and increasing the potential links between them (see Section 5.3).

5. Analysis of Results

In this section we present our analysis of bias in the Wikimedia ecosystem. In Section 5.1, we describe the impact of our framework on reducing *allocative* bias concerning Transnational writers in Wikidata; Section 5.2 addresses the presence of representational bias that our IE pipeline may introduce; Section 5.3 analyzes the effect of our pipeline on connecting Transnational writers with other people in the KG.

5.1. Is it Possible to Reduce Underrepresentation through Biographical Event Extraction?

The first part of our analysis focuses on the impact of our biographical triple extraction pipeline (described in Section 4) on the underrepresentation of Transnational writers. Specifically, we want to assess whether this augmentation results in a knowledge base that is more balanced towards Transnational people.

Table 3 shows the increase of connections between writers and other people and its effects on all the four socio-demographic groups: Western-Educated men, Transnational men, Western women, and Transnational women. At a general level, the augmentation is highly significant with a 40-fold increase of biographical situations in the PiM-KG that shift from 62,936 to 2,552,225. The effects of this increase on Transnational writers, however, are less significant. The number of biographical situations involving Transnational men increases by 1.9%, while the increase is only of 0.3% for Transnational women. Western men obtain the highest increase with a relative increase of 2.6%. Western Women are negatively affected by the augmentation, losing 5.1% of representation in triples extracted from English Wikipedia.

We performed an additional analysis by observing the impact of augmentation on the average number of triples associated with writers and on the number of writers who appear in at least one triple. The results in Table 4 show a dramatic increase in both metrics. The average number of triples *per* writers increases from 2.2 to 29.7 while the percentage of writers appearing in at least one triple reaches 80% from an initial 26%. Since only triples connecting two persons have been considered in this experiment, this result can be

¹¹https://it.wikipedia.org/wiki/Phan_Bòeĩ_Chõâu

¹²https://en.wikipedia.org/wiki/Wendy_Williams

	Wikidata	Biographical Triple Extraction	Δ
all	62,936	2,552,225	–
Western-Educated Men	65.1%	67.7%	+2.6%
Transnational Men	6.3%	8.2%	+1.9%
Western Women	25.6%	20.5%	–5.1%
Transnational Women	2.7%	3.0%	+0.3%

Table 3

The number of events before and after the augmentation

	Wikidata	Biographical Triple Extraction
all	2.2 (26%)	29.7 (80%)
Western-Educated Men	2.3 (27%)	33.12 (81%)
Transnational Men	2.1 (17%)	25.0 (73%)
Western Women	2.1 (30%)	23.8 (83%)
Transnational Women	1.9 (19%)	22.0 (74%)

Table 4

The average number of events *per* group on Wikidata and after the augmentation. The percentage of writers with at least one biographical event is between parenthesis

interpreted as a general increase of connections between writers and other people in the KG. The impact of the augmentation appears to be significant for all socio-demographic groups without any statistically significant differences: The percentage of Transnational men with at least one biographical triple increases by 56%; for Transnational women, the increase is 55%; for Western men, 54%; and for Western women, 53%. When the average number of triples is considered, the highest increase is observed on Western men, which shifts from 2.3 events *per* person to 33.12. The lowest increase is observed in Transnational women who are associated with 22.0 triples *per* person starting from an average of 1.9. The gender axis appears to be the one that benefits the less from the augmentation: the average number of triples involving men reaches 32.15, triples related to women 23.67, showing a delta of 8.48 triples. The delta along the origin axis is 6.37, with an average of 30.51 triples about Western people against 22.14 about Transnational people. The augmentation significantly increases the average number of biographical relationships. Although this effect is consistent across sociodemographic groups, Western men still experience the greatest increase in average events; gender differences show that men benefit slightly more from the increase than women.

5.2. Does Biographical Event Extraction Induce Representational Bias?

In Section 5.1, we observed the positive impact of biographical triples augmentation on reducing the underrepresentation of vulnerable groups to discrimination. In this section, we evaluate whether this augmentation induces *representational* bias against specific groups of writers. Specifically, based on related work presented in Section 2, we explore the presence of stereotypical associations between socio-demographic groups and extracted triples.

In Table 5 the five most frequently occurring biographical predicates associated to each socio-demographic group are reported. Results are broken down by source: in the first row there are the top-5 predicates gathered from Wikidata; in the second, the top-5 predicates extracted from Wikipedia pages. Considering the limited number of Wikidata properties that connect two persons in Wikidata, four properties are shared by all groups and highlight relationships from the private life: ‘child’, ‘spouse’, ‘sibling’, and ‘relative’. The only difference that emerges concerns the property ‘influenced by’, which is among the top-5 triples in all socio-demographic groups except from Western women. The latter is the only group that is exclusively associated with private-life triples. The analysis of the top-5 biographical triples derived from English Wikipedia shows a higher focus on career-related events (‘wrote’, ‘directed’, ‘published’,

	W. men	T. men	W. women	T. Women
Wikidata	child, spouse, sibling, influenced by, relative	child, spouse, sibling, influenced by, relative	spouse, child, sibling, relative, unmarried partner	spouse, child, sibling, influenced by, relative
Expanded	wrote, published, born, married, worked	wrote, published, directed, worked, born	married, wrote, published, born, met	published, married, wrote, born, received

Table 5

The top-5 biographical triples associated to each socio-demographic group gathered from Wikidata and after the triples augmentation process.

‘worked’, ‘received’) than on private-life ones (‘married’, ‘born’, ‘met’). The gender axis seems to be the more vulnerable to *representational* bias. Western women are associated to 3 private-life events out of 5 and the most frequently occurring one is ‘married’. On the opposite, Transnational men is the only group that is not associated to the event ‘married’ among the top-5 and is associated with only 1 private-life event out of 5. After the augmentation with data from Wikipedia, a shift toward career-related events is observed, but gender disparities persist: Western women remain predominantly associated with private-life events, while Transnational men are linked almost exclusively to career-related events.

To study the nature of the *representational* bias across group induced by the augmentations, we compared the distributions of extracted triples between socio-demographic groups, with the goal of identifying the events that are more significant to a group if compared to another one. To do so, we adopted a pairwise comparison approach [37, 91] and used the Jensen-Shannon Divergence [62] (JSD) as a metric to compare the distributions of the events. Given two distributions P and Q, JSD is defined as follows:

$$\text{JSD}(P \parallel Q) = \frac{1}{2}D_{KL}(P \parallel M) + \frac{1}{2}D_{KL}(Q \parallel M) \quad (1)$$

where

$$M = \frac{1}{2}(P + Q) \quad (2)$$

and $D_{KL}(P \parallel M)$ is the Kullback-Leibler divergence defined as:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (3)$$

In our setting, we choose pairs of socio-demographic groups (e.g., women *versus* men) and for each pair we compute the JSD of the probability distributions of the extracted biographical events. Given an extracted biographical event, we first compute its relative distribution r_i and the distribution of all the other events $1 - r_i$. For instance, the event ‘married’ is represented as the distribution $p_w = (0.017, 0.983)$ for women and $p_m = (0.008, 0.992)$ for men. The closest the JSD score is to 1 or -1 , the highest is the divergence between the distributions. The polarity of the score conveys information about the group with which the event is more significantly associated. In the example, the JSD is 0.028 shows that the event ‘married’ is more significantly associated with women.

Table 6 shows the pairwise comparisons between all socio-demographic groups. Coherently with previous analysis and the existing literature [95], *representational* bias mainly affects women. In the comparison based only on gender, 7 events out of 10 are about parenthood or marriage and 2 about private life (‘moved’, ‘lived’). The only career-related event associated to women is ‘published’. The intersection of gender and

Pair	Group 1	Group 2
Transnational <i>vs</i> Western	starring, debut, sung, translated, award, is, started, career, comeback, awarded	increasing, closure, pursue, listen, discussing, echoed, acknowledge, convincing, app, advocates
Women <i>vs</i> Men	married, born, marriage, mother, daughter, moved, published, lived, wife, divorced	posting, impressed, assumed, veto, pushed, send, pursued, opposed, resisted, places
T. Women <i>vs</i> T. Men	losing, daughter, concert, mother, married, reach, explores, contributor, defeating, coming	negotiations, considers, influential, became, declared, place, charges, engaged, used, provided
T. Women <i>vs</i> W. Men	translated, daughter, mother, award, awarded, received, participated, is, degree, awards	respect, founded, controversial, accolades, collaboration, identified, spotted, serve, expansion, inviting
T. Women <i>vs</i> W. Women	defeated, reach, defeating, translated, award, concert, lost, participated, losing, starring	secure, seek, nominated, described, fell, recovered, event, came, followed, speaking
W. Women <i>vs</i> W. Men	married, born, marriage, mother, moved, published, lived, daughter, wife, met	respect, founded, controversial, premiere, banned, accolades, collaboration, identified, informed, spotted
W. Women <i>vs</i> T. Men	married, born, moved, lived, died, met, marriage, appeared, divorce, published	negotiations, considers, statement, decided, issued, receive, held, selected, be, replaced
T. Men <i>vs</i> W. Men	starring, sung, debut, career, directed, started, conferred, award, composed, translated	respect, founded, controversial, premiere, banned, accolades, collaboration, identified, spotted, serve

Table 6

The pairwise comparison of biographical events between different socio-demographic groups. The 10 most significant events for each pair have been computed through the Jensen-Shannon Divergence (JSD).

origin mitigates *representational* biases. The presence of events related to parenthood and marriage is mitigated by a higher number of career-related events, both in comparison with Transnational men and Western men (e.g., ‘loosing’, ‘concert’, ‘contributor’, ‘translated’, ‘awarded’). However, this mitigation can be considered as the side-effect of the high underrepresentation of Transnational women writers in Wikidata, who represent the 4.3% of the total. This might suggest that only highly recognized people belonging to this socio-demographic group are included in Wikidata and Wikipedia, and for such a reason more career-related triples have been extracted about them.

The Transnational axis appears to be less relevant in the perpetration of representational bias. No private-life events show a strong association to a group when people of different origin and same gender are compared. It is worth mentioning that, when compared to Western people, both Transnational men and women are associated with events related to other professions (e.g., ‘concert’, ‘starred’, ‘sung’, ‘composed’), suggesting that writers belonging to these socio-demographic groups are represented as more polyhedral than their Western counterparts.

5.3. Does Biographical Triple Extraction Facilitate the Discovery of Minorities?

Our third analysis of bias focuses on the connections of Transnational writers with other people in the PiM-KG. We hypothesize that augmenting the KG through biographical triple extraction increases the number of links between Transnational writers and entities, facilitating their potential discovery in downstream applications such as recommendation systems [10]. To do so, we chose to represent the relationships between individuals using a graph structure. The graph consists of nodes representing individuals and edges representing biographical relationships between them. For example, an edge between two individuals exists if they are married or linked by any other relation type. The graphs are undirected simple graphs: if more than one links between the same pair of entities, their relationship is represented by a

single edge. Therefore, the network consists of individuals, including writers, and each edge connects two individuals, whose at least one is a writer. To compare the impact of augmentation on links, we considered three distinct graphs:

- *Wikidata*: a knowledge graph based on Wikidata relationships
- *Expanded*: a knowledge graph derived from biographical triples extracted from Wikipedia pages
- *Expanded + Wikidata*: a merged graph combining the two

Largest Giant Component. Our first analysis focuses on the Giant Connected Component (GCC), defined as the largest subgraph in which all nodes are reachable from one another. A larger GCC means a more interconnected network of writers. We aim to assess the extent to which biographical event extraction expands the Giant Connected Component (GCC) of the network, thereby increasing the proportion of nodes connected within the knowledge base. Table 7 presents the key metrics for the three graphs. As can be observed, the table shows a dramatic increase of the GCC that shifts from 13.49% in the graph *Wikidata* to 68.53% in the graph *Expanded + Wikidata*, which combines Wikidata and the triples extracted from Wikipedia biographies. This indicates that the graph *Expanded* introduces new connections between writers and new individuals, preventing individual writers or small groups of them from being segregated (isolated from the GCC, which consequently increases). In fact, the presence of these added individuals allows writers to be connected through them, even in the absence of a direct event-based relationship with one another. Moreover, in both networks *Expanded* and *Expanded + Wikidata*, the proportion of Transnational writers increases. This suggests that Wikipedia extraction is more likely to uncover new connections for Transnational writers than for Western ones. For what concerns women, their proportion decreases in network *Expanded* (-0.78%) but increases in network *Expanded + Wikidata* (+0.47%). This indicates that while the additional connections introduced by Wikipedia extraction reduce their relative presence, merging the two networks restores and even increases their representation. Furthermore, it is important to highlight that the percentage of writes decreases in the *Expanded* graph due to the fact that 70.99% of the newly added nodes are not writers but individuals who are linked to writers through the extracted biographical events.

Metric	Wikidata	Expanded	Expanded + Wikidata
Number of nodes	72,056	291,619	323,877
Number of edges	61,641	583,578	629,434
Size of the giant connected component (%)	13.49%	89.80%	89.46%
Percentage of writers in the GCC (%)	35.31%	29.01%	26.58%
Percentage of transnational writers in the GCC (%)	6.38%	11.95%	13.07%
Percentage of women writers in the GCC (%)	24.03%	23.25%	24.5%

Table 7

Statistics about LGC components and percentage by sociodemographic

Density. Our second analysis focuses on the connections of different socio-demographic groups in the network. Our aim is to investigate whether the extraction of triples augments the connections between groups along the axes of gender and origin. The key metric for this analysis is density, defined as the ratio between the number of observed edges and the number of possible edges within or between groups. We compute intra-group densities (e.g., connections among Transnational women) and inter-group densities (e.g., connections between Western men and Transnational women) before and after augmentation. An increase in density indicates a higher level of connectivity and suggests a richer representation of the group's relations in the knowledge graph. This allows us to assess whether the augmentation process introduces more balanced or more skewed relational patterns across groups.

In Table 8 it is possible to observe the density of connections within a each group (intra-group) and between group pairs (inter-group) where a higher density means a lower segregation within or between

groups. Each column reports the density for each graph. Focusing on the origin axis, it emerges that the augmentation generally reduces the segregation and Transnational and Western people are more connected after the augmentation. The opposite happens for the gender-axis where the density between men and women decreases from $1.77e-05$ to $1.37e-05$, while the density within groups increases. This result is coherent with the quantitative analysis shown in Section 5.1 in which it emerges that the triple augmentation has a positive effect on Transnational people and a negative effect on Women.

Measure	Wikidata	Expanded	Expanded+Wikidata
Intra-group Transnational	0.0001801	9.51e-05	9.63e-05
Intra-group Western	3.34e-05	4.37e-05	4.40e-05
Inter-group Transnational vs Western	6.77e-06	7.63e-06	7.68e-06
Intra-group Women	2.82e-05	3.91e-05	3.97e-05
Intra-group Men	2.71e-05	4.84e-05	4.80e-05
Inter-group Women vs Men	1.77e-05	1.37e-05	1.40e-05

Table 8

Network density measures for different graphs.

Closeness. In the last part of our analysis we study whether the biographical triple extraction has a positive impact on augmenting the connections of entities in the graph. The key metric of this analysis is *Closeness Centrality*, which indicates how ‘close’ a node is to all other nodes in the network. This score is obtained by computing the inverse of the sum of the shortest path (that is, the minimum number of steps or edges needed to reach another node) distances connecting each node to all others. In this network, the writers with the highest Closeness Centrality are the ones who can reach all other writers through the shortest average distance [36].

In Table 9 we reported the top-10 ranked writers according to their centrality for each socio-demographic group for all the three observed networks. As expected, Western men score the highest centrality in all the networks. Transnational men writers are the ones who benefit the most from the augmentation. The first-ranked Transnational writer for closeness in ‘wikidata’ is the 165th overall, while the first one in ‘wikidata + expanded’ is ranked 42nd. In parallel, both Western and Transnational women are negatively affected by the augmentation: except from the first two top-ranked Transnational women, who have a higher closeness on ‘wikidata + expanded’ than ‘wikidata’, all the other top-10 women writers lose centrality after the augmentation.

Qualitative Analysis. A qualitative analysis of the results shows taxonomic issues related to the categorization of writers. The modeling of Transnational people described in Section 3 is not free from false positives, such as the children of diplomats in Global South countries (e.g., Italo Calvino, Sonia Orwell). Additional socio-demographic features are needed to identify these cases. However, Wikidata alone does not offer the needed properties to do this kind of additional filtering: more nuanced taxonomies are needed to overcome this issue.

A second issue emerging from the qualitative analysis regards the status of people who are considered writers in Wikidata. Many influential nodes in the network are associated with the profession of ‘writer’ in Wikidata even if they are not recognized as such by the public opinion. It is the case of politicians (Ronald Reagan), television celebrities (Oprah Winfrey), and models (Gisele Bündchen) who are the authors of a book but they are not properly writers. This is induced by the participatory nature of Wikidata that leads to biases introduced by contributors: as long as there are no strict guidelines to follow for adding records in this knowledge base and for ranking them as preferred, it is natural to have this noise. Identifying methods to clean the dataset is an open issue that must be further investigated.

Rank	Transnational Women		Western Women		Transnational Men		Western Men	
	Writer	Overall Rank	Writer	Overall Rank	Writer	Overall Rank	Writer	Overall Rank
Wikidata								
1	María Kodama	194	Rosalie Mackenzie Poe	3	Italo Calvino	165	Edgar Allan Poe	1
2	Sonia Orwell	223	Virginia Eliza Clemm Poe	4	Gabriel García Márquez	219	Charles Dickens	2
3	Gisele Bündchen	533	Catherine Dickens	10	Louis Althusser	274	Friedrich Nietzsche	5
4	Zoya Akhtar	564	Mary Angela Dickens	28	Rodrigo García	292	Guy de Maupassant	7
5	Padma Lakshmi	623	Eleanor Marx	30	Jacques Derrida	337	Émile Zola	8
6	Honey Irani	650	Laura Marx	31	Salman Rushdie	425	Charles Dickens, Jr.	9
7	Sujata Nahar	696	Augusta, Lady Gregory	38	Farhan Akhtar	563	African Spir	12
8	Nandana Sen	852	Olivia Shakespear	39	Javed Akhtar	566	Lev Shestov	13
9	Margarita Zorrera	880	Helena Blavatsky	40	Zuhair Al-Karmi	604	William Butler Yeats	14
10	Shira Geffen	1028	Louise Colet	42	Paulo Freire	659	Stefan Zweig	15
Expanded								
1	Indira Gandhi	159	Hillary Clinton	4	John McCain	37	Bill Clinton	1
2	Gisele Bündchen	1270	Pauline Kael	34	Salman Rushdie	42	Ronald Reagan	2
3	Joanna Lumley	1491	Oprah Winfrey	35	Saddam Hussein	50	Roger Ebert	3
4	Nadine Gordimer	1658	Jane Fonda	45	Nelson Mandela	86	Steven Spielberg	5
5	Deepa Mehta	1673	Lauren Bacall	68	Muammar Gaddafi	220	Mark Twain	7
6	Ayaan Hirsi Ali	1707	Barbra Streisand	69	Narendra Modi	238	Ernest Hemingway	8
7	María Kodama	1890	Gwyneth Paltrow	101	Mario Vargas Llosa	299	T. S. Eliot	9
8	Marlene van Niekerk	1903	Shirley MacLaine	108	Louis C.K.	401	Martin Luther King Jr.	10
9	Margaret Busby	1983	Vanessa Redgrave	117	Derek Walcott	522	Martin Scorsese	11
10	Juman Malouf	2227	Toni Morrison	142	Jacques Derrida	541	William Faulkner	12
Expanded + Wikidata								
1	Indira Gandhi	171	Hillary Clinton	4	Salman Rushdie	42	Bill Clinton	1
2	Gisele Bündchen	1357	Pauline Kael	41	John McCain	44	Ronald Reagan	2
3	Sonia Orwell	1382	Oprah Winfrey	46	Saddam Hussein	58	Roger Ebert	3
4	Joanna Lumley	1672	Jane Fonda	52	Nelson Mandela	67	Steven Spielberg	5
5	Nadine Gordimer	1705	George Eliot	55	Narendra Modi	227	Ernest Hemingway	6
6	María Kodama	1765	Lauren Bacall	72	Muammar Gaddafi	243	Mark Twain	7
7	Deepa Mehta	1777	Barbra Streisand	83	Mario Vargas Llosa	249	T. S. Eliot	8
8	Ayaan Hirsi Ali	1840	Gwyneth Paltrow	121	Jacques Derrida	368	William Faulkner	10
9	Padma Lakshmi	1858	Shirley MacLaine	131	Louis C.K.	449	Martin Luther King Jr.	11
10	Margaret Busby	1963	Eleanor Roosevelt	133	Derek Walcott	495	Martin Scorsese	12

Table 9

Top 10 writers by Closeness Centrality for graphs Wikidata, Expanded, and Wikidata + Expanded, categorized by group, with their overall ranking position.

6. Conclusion and Future Work

In this paper we presented WikiBias: a framework for the exploration of bias in the Wikimedia Ecosystem based on semantic modeling and biographical event extraction. The framework, designed for the joint analysis of underrepresentation and representational bias in datasets, has been tested on the case study of writers in Wikidata and in English Wikipedia. We analyzed the presence of bias against socio-demographic groups characterized by the intersection of gender and origin and the impact of biographical event detection on their representation.

Results show that the integration of Semantic Modeling and Information Extraction techniques has an effect in mitigating the lack of representation of entities belonging to vulnerable groups. Before the augmentation, Transnational people are affected by high underrepresentation in Wikidata: they are a small minority that is not connected with other nodes in the knowledge base. Augmenting the triples about them through biographical event extraction, these authors are more connected to the rest of the network: a higher number of Transnational writers are associated with at least one person in the KG and the average number of triples about them significantly increases. The augmentation, however, increases the relative underrepresentation of Western women, showing a negative impact on the gender axis. Not surprisingly, the analysis of the events extracted from English Wikipedia shows the presence of gender

1 stereotypes in Wikipedia pages. Women are more frequently associated with events related to parenthood
 2 and marriage than men. This form of misrepresentation is mitigated when gender and origin are intersected:
 3 Transnational women writers are less associated to private-life events than their Western counterparts.
 4 However, this might be an effect on the strong underrepresentation of this group, which represents only
 5 the 3.9% of the writers in the graph. The evaluation of writers' graph through Network Analysis shows
 6 that the augmentation might be a key factor to increase the discovery of entities belonging to minorities.
 7 The comparison of networks before and after the augmentation shows a dramatic increase of Transnational
 8 writers that are connected to the rest of the network, shifting from 6.83% to 13.07%. The analysis also
 9 shows an increase of connections between socio-demographic groups both on the gender and the origin
 10 axis, demonstrating that biographical event extraction might facilitate the discovery of writers belonging
 11 to minorities. From a modeling perspective, PiM-O enabled the identification of knowledge gaps that affect
 12 Wikidata's taxonomy. Developing an ODP that does not provide restrictions on event types supported
 13 the encoding of a wider set of relations that can possibly connect two individuals in the PiM KG. For
 14 instance, properties like 'introduce' and 'interview', which are absent from Wikidata, are highly frequent
 15 among the extracted triples. This might open to a potential revising of the existing Wikidata's taxonomy
 16 with the aim of improving the automatic integration between structured knowledge and textual data.

17 Future work will explore different research directions. The extraction of biographical events from different
 18 sources than English Wikipedia might have a role in reducing *representational* bias against women, while
 19 expanding the extraction of triples to other types of named entities (e.g., locations, organizations) can
 20 lead to richer representation of writers and tighten their connections with other nodes in the network. The
 21 implementation of our biographical event extraction pipeline to other languages can support a culturally-
 22 aware analysis of bias in Wikipedia, showing the presence of representational biases that are related
 23 to language-specific versions of Wikipedia. The adoption of the augmented KG in AI applications like
 24 Recommendation Systems and Language Models can provide a significant contribution to the study of the
 25 impact of datasets on the emergence of bias in AI technologies.

26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51

- References
- [1] I.R. Abubakar, K.M. Maniruzzaman, U.L. Dano, F.S. AlShihri, M.S. AlShammari, S.M.S. Ahmed, W.A.G. Al-Gehlani and T.I. Alrawaf, Environmental sustainability impacts of solid waste management practices in the global South, *International journal of environmental research and public health* **19**(19) (2022), 12717.
 - [2] J. Adams, H. Brückner and C. Naslund, Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test", *Socius* **5** (2019), 2378023118823946.
 - [3] R. Agerri, I. Aldabe, Z. Beloki, E. Laparra, M.L. de Lacalle, G. Rigau, A. Soroa, A. Fokkens, R. Izquierdo, M. van Erp et al., Event detection, version 2 deliverable 4.2. 2, *Deliverable, NewsReader Project* (2014).
 - [4] G. Amaral, O. Rodrigues and E. Simperl, WDV: A broad data verbalisation dataset built from Wikidata, in: *International Semantic Web Conference*, Springer, 2022, pp. 556–574.
 - [5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, Dbpedia: A nucleus for a web of open data, in: *international semantic web conference*, Springer, 2007, pp. 722–735.
 - [6] C.F. Baker, C.J. Fillmore and J.B. Lowe, The berkeley framenet project, in: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.
 - [7] D. Bamman, B. O'Connor and N.A. Smith, Learning latent personas of film characters, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 352–361.
 - [8] A.-L. Barabási, R. Albert and H. Jeong, Mean-field theory for scale-free random networks, *Physica A: Statistical Mechanics and its Applications* **272**(1) (1999), 173–187. doi:[https://doi.org/10.1016/S0378-4371\(99\)00291-5](https://doi.org/10.1016/S0378-4371(99)00291-5). <https://www.sciencedirect.com/science/article/pii/S0378437199002915>.
 - [9] S. Barocas, K. Crawford, A. Shapiro and H. Wallach, The problem with bias: Allocative versus representational harms in machine learning, in: *9th Annual conference of the special interest group for computing, information and society*, New York, NY, 2017, p. 1.
 - [10] P. Basile, C. Musto, M. de Gemmis, P. Lops, F. Narducci and G. Semeraro, Content-based recommender systems+DBpedia knowledge= semantics-aware recommender systems, in: *Semantic Web Evaluation Challenge: SemWebEval 2014 at ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, Springer, 2014, pp. 163–169.
 - [11] E.M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

- [12] S.L. Blodgett, S. Barocas, H. Daumé III and H. Wallach, Language (Technology) is Power: A Critical Survey of “Bias” in NLP, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5454–5476.
- [13] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama and A.T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, *Advances in neural information processing systems* **29** (2016).
- [14] B. Boter, M. Rensen and G. Scott-Smith, *Unhinging the National Framework: Perspectives on Transnational Life Writing*, Sidestone Press, 2020.
- [15] W. Brandt, The north-south dialogue: The issue is survival, *Challenge* **25**(4) (1982), 22–30.
- [16] J. Brooke, A. Hammond and G. Hirst, GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus, in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 2015, pp. 42–47.
- [17] S. Brown, Scaling Up Collaboration Online: Toward a Collaboratory for Research on Canadian Writing, *International Journal of Canadian Studies* **48** (2014), 233–251.
- [18] S.W. Brown, C. Bonial, L. Obrst and M. Palmer, The rich event ontology, in: *Proceedings of the Events and Stories in the News Workshop*, 2017, pp. 87–97.
- [19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language models are few-shot learners, *Advances in neural information processing systems* **33** (2020), 1877–1901.
- [20] R. Bunescu and M. Paşca, Using Encyclopedic Knowledge for Named entity Disambiguation, in: *11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Trento, Italy, 2006, pp. 9–16. <https://aclanthology.org/E06-1002>.
- [21] P.-L.H. Cabot and R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2370–2381.
- [22] B. Collier and J. Bear, Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions, in: *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 383–392.
- [23] K.W. Crenshaw, Mapping the margins: Intersectionality, identity politics, and violence against women of color, in: *The public nature of private violence*, Routledge, 2013, pp. 93–118.
- [24] D. Damrosch, *What is world literature?*, Vol. 5, Princeton University Press, 2003.
- [25] D. Damrosch and G.C. Spivak, Comparative literature/world literature: A discussion with Gayatri Chakravorty Spivak and David Damrosch, *Comparative Literature Studies* **48**(4) (2011), 455–485.
- [26] D. Das, D. Chen, A.F. Martins, N. Schneider and N.A. Smith, Frame-semantic parsing, *Computational linguistics* **40**(1) (2014), 9–56.
- [27] P. Das, S.K. Karnam, A.B. Soni and A. Mukherjee, Social Biases in Knowledge Representations of Wikidata separates Global North from Global South, in: *Proceedings of the 17th ACM Web Science Conference 2025*, 2025, pp. 12–21.
- [28] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [29] T. d’Haen, D. Damrosch, D. Kadir et al., *The Routledge companion to world literature*, Routledge London, 2012.
- [30] F. Dib, S. Lindberg and P. Nugues, Extraction of Career Profiles from Wikipedia., in: *BD*, 2015, pp. 33–38.
- [31] G.R. Doddington, A. Mitchell, M.A. Przybocki, L.A. Ramshaw, S.M. Strassel and R.M. Weischedel, The automatic content extraction (ace) program-tasks, data, and evaluation., in: *Lrec*, Vol. 2, Lisbon, 2004, pp. 837–840.
- [32] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell and M. Gardner, Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 1286–1305.
- [33] K. Dotson, Tracking epistemic violence, tracking practices of silencing, *Hypatia* **26**(2) (2011), 236–257.
- [34] A. Field, C.Y. Park, K.Z. Lin and Y. Tsvetkov, Controlled analyses of social biases in wikipedia bios, in: *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2624–2635.
- [35] A. Fokkens, S. Ter Braake, N. Ockeloën, P. Vossen, S. Legêne, G. Schreiber and V. de Boer, Biographynet: Extracting relations between people and events, *arXiv preprint arXiv:1801.07073* (2018).
- [36] L.C. Freeman, Centrality in social networks conceptual clarification, *Social Networks* **1**(3) (1978), 215–239. doi:[https://doi.org/10.1016/0378-8733\(78\)90021-7](https://doi.org/10.1016/0378-8733(78)90021-7).
- [37] R.J. Gallagher, M.R. Frank, L. Mitchell, A.J. Schwartz, A.J. Reagan, C.M. Danforth and P.S. Dodds, Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts, *EPJ Data Science* **10**(1) (2021), 4.
- [38] A. Gangemi and V. Presutti, Ontology design patterns, in: *Handbook on ontologies*, Springer, 2009, pp. 221–243.
- [39] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari and L. Schneider, Sweetening ontologies with DOLCE, in: *International conference on knowledge engineering and knowledge management*, Springer, 2002, pp. 166–181.
- [40] A. Gaut and T. Sun, Towards Understanding Gender Bias in Relation Extraction, *Association for Computational Linguistics (ACL 2019)* (2020).
- [41] S. Gehman, S. Gururangan, M. Sap, Y. Choi and N.A. Smith, RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2020* (2020).
- [42] A. Gil and É. Ortega, Global outlooks in digital humanities: Multilingual practices and minimal computing, in: *Doing digital humanities*, Routledge, 2016, pp. 58–70.

- [43] S. Giorgi, V. Zavarella, H. Tanev, N. Stefanovitch, S. Hwang, H. Hettiarachchi, T. Ranasinghe, V. Kalyan, P. Tan, S. Tan, M. Andrews, T. Hu, N. Stoehr, F.I. Re, D. Vegh, D. Atzenhofer, B. Curtis and A. Hürriyetoglu, Discovering Black Lives Matter Events in the United States: Shared Task 3, CASE 2021, in: *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, Association for Computational Linguistics, Online, 2021, pp. 218–227. doi:10.18653/v1/2021.case-1.27. <https://aclanthology.org/2021.case-1.27>.
- [44] L.A. Goodman, Snowball sampling, *The annals of mathematical statistics* (1961), 148–170.
- [45] N. Hare, The battle for Black studies, *The Black Scholar* 3(9) (1972), 32–47.
- [46] N. Heist and H. Paulheim, Uncovering the semantics of Wikipedia categories, in: *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, Springer, 2019, pp. 219–236.
- [47] D. Hovy and S. Prabhunoye, Five sources of bias in natural language processing, *Language and linguistics compass* 15(8) (2021), e12432.
- [48] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw and R. Weischedel, OntoNotes: the 90% solution, in: *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, 2006, pp. 57–60.
- [49] C. Hube, Bias in wikipedia, in: *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017b, pp. 717–721.
- [50] C. Hube, F. Fischer, R. Jäschke, G. Lauer and M.R. Thomsen, World literature according to wikipedia: Introduction to a DBpedia-based framework, *arXiv preprint arXiv:1701.00991* (2017a).
- [51] B. Hui, L. Zhang, X. Zhou, X. Wen and Y. Nian, Personalized recommendation system based on knowledge embedding and historical behavior, *Applied Intelligence* (2022), 1–13.
- [52] W. IJntema, J. Sangers, F. Hogenboom and F. Frasincar, A lexico-semantic pattern language for learning ontology instances from text, *Journal of Web Semantics* 15 (2012), 37–50.
- [53] M. Joshi, D. Chen, Y. Liu, D.S. Weld, L. Zettlemoyer and O. Levy, Spanbert: Improving pre-training by representing and predicting spans, *Transactions of the association for computational linguistics* 8 (2020), 64–77.
- [54] P.R. Kingsbury and M. Palmer, From TreeBank to PropBank., in: *LREC*, 2002, pp. 1989–1993.
- [55] K. Kipper, A. Korhonen, N. Ryant and M. Palmer, A large-scale classification of English verbs, *Language Resources and Evaluation* 42 (2008), 21–40.
- [56] H.-U. Krieger and T. Declerck, An OWL Ontology for Biographical Knowledge. Representing Time-Dependent Factual Knowledge., in: *BD*, 2015, pp. 101–110.
- [57] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee et al., Natural questions: a benchmark for question answering research, *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.
- [58] P. Leskinen, E.A. Hyvönen and J.A. Tuominen, Analyzing and visualizing prosopographical linked data based on biographies, in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, CEUR Workshop Proceedings, 2018.
- [59] L. Lucy, D. Tadimeti and D. Bamman, Discovering Differences in the Representation of People using Contextualized Semantic Axes, in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.
- [60] P. Manghi, Miriam Baglionil, Andrea Mannocci1, Gina Pavone1, Michele De Bonis1 and (2023), 47–59.
- [61] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard and D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.
- [62] M.L. Menéndez, J.A. Pardo, L. Pardo and M.d.C. Pardo, The jensen-shannon divergence, *Journal of the Franklin Institute* 334(2) (1997), 307–318.
- [63] S. Menini, R. Sprugnoli, G. Moretti, E. Bignotti, S. Tonelli and B. Lepri, RAMBLE ON: Tracing movements of popular historical figures, in: *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 77–80.
- [64] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young and R. Grishman, The NomBank project: An interim report, in: *Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004*, 2004, pp. 24–31.
- [65] S. Milgram et al., The small world problem, *Psychology today* 2(1) (1967), 60–67.
- [66] A.-L. Minard, M. Speranza, R. Urizar, B. Altuna, M. Van Erp, A. Schoen and C. Van Son, MEANTIME, the News-Reader multilingual event and time corpus, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 2016, pp. 4417–4422.
- [67] M. Mitchell, G. Attanasio, I. Baldini, M. Cliniciu, J. Clive, P. Delobelle, M. Dey, S. Hamilton, T. Dill, J. Doughman et al., SHADES: Towards a multilingual assessment of stereotypes in large language models, in: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 11995–12041.
- [68] M. Nadeem, A. Bethke and S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5356–5371.
- [69] L. Nakamura, “Words with friends”: socially networked reading on Goodreads, *Pmla* 128(1) (2013), 238–243.

- [70] M.E.J. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, *Phys. Rev. E* **64** (2001), 016132. doi:10.1103/PhysRevE.64.016132.
- [71] T. O’Gorman, K. Wright-Bettner and M. Palmer, Richer event description: Integrating event coreference with temporal, causal and bridging annotation, in: *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, 2016, pp. 47–56.
- [72] A. Plum, M. Zampieri, C. Orasan, E. Wandl-Vogt and R. Mitkov, Large-scale data harvesting for biographical data (2019).
- [73] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro et al., The timebank corpus, in: *Corpus linguistics*, Vol. 2003, Lancaster, UK., 2003a, p. 40.
- [74] J. Pustejovsky, J.M. Castano, R. Ingria, R. Sauri, R.J. Gaizauskas, A. Setzer, G. Katz and D.R. Radev, TimeML: Robust specification of event and temporal expressions in text., *New directions in question answering* **3** (2003b), 28–34.
- [75] J.W. Ratcliff and D.E. Metzener, Pattern-matching-the gestalt approach, *Dr Dobbs Journal* **13**(7) (1988b), 46.
- [76] J.W. Ratcliff, D. Metzener et al., Pattern matching: The Gestalt approach, *Dr. Dobb’s Journal* **13**(7) (1988a), 46.
- [77] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger and T. Bogaard, Building event-centric knowledge graphs from news, *Journal of Web Semantics* **37** (2016), 132–151.
- [78] I. Russo, T. Caselli and M. Monachini, Extracting and Visualising Biographical Events from Wikipedia., in: *BD*, 2015, pp. 111–115.
- [79] E.W. Said, Orientalism, in: *Social theory re-wired*, Routledge, 2023, pp. 362–374.
- [80] V. Sanh, L. Debut, J. Chaumond and T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* (2019).
- [81] M. Sap, D. Card, S. Gabriel, Y. Choi and N.A. Smith, The risk of racial bias in hate speech detection, in: *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 1668–1678.
- [82] K.K. Schuler, *VerbNet: A broad-coverage, comprehensive verb lexicon*, University of Pennsylvania, 2005.
- [83] Z. Shaik, F. Ilievski and F. Morstatter, Analyzing race and citizenship bias in Wikidata, in: *2021 IEEE 18th international conference on mobile Ad Hoc and smart systems (MASS)*, IEEE, 2021, pp. 665–666.
- [84] D. Shaver and M.A. Shaver, Books and digital technology: A new industry model, in: *Special Issue on the Changing World of Publishing*, Routledge, 2020, pp. 71–86.
- [85] M. Sims, J.H. Park and D. Bamman, Literary Event Detection, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3623–3634. doi:10.18653/v1/P19-1353. <https://aclanthology.org/P19-1353>.
- [86] G.C. Spivak, Can the Subaltern Speak?, in: *Colonial discourse and post-colonial theory*, Routledge, 2015, pp. 66–111.
- [87] M.A. Stranisci, V. Patti and R. Damiano, Representing the under-represented: A dataset of post-colonial, and migrant writers, in: *3rd Conference on Language, Data and Knowledge, LDK 2021*, Vol. 93, Schloss Dagstuhl-Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, 2021a, pp. 1–14.
- [88] M.A. Stranisci, V. Basile, R. Damiano and V. Patti, Mapping Biographical events to ODPs through Lexico-Semantic Patterns?, in: *12th Workshop on Ontology Design and Patterns, WOP 2021*, Vol. 3011, CEUR-WS, 2021b, pp. 1–12.
- [89] M.A. Stranisci, E. Mensa, R. Damiano, D. Radicioni and O. Diakite, Guidelines and a Corpus for Extracting Biographical Events, in: *Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation within LREC2022*, 2022a, pp. 20–26.
- [90] M.A. Stranisci, G. Spillo, C. Musto, V. Patti and R. Damiano, The URW-KG: a Resource for Tackling the Underrepresentation of non-Western Writers, *arXiv preprint arXiv:2212.13104* (2022b).
- [91] M.A. Stranisci, R. Damiano, E. Mensa, V. Patti, D. Radicioni and T. Caselli, 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)*, Association for Computational Linguistics, Toronto, Canada, 2023a, pp. 12370–12384. doi:10.18653/v1/2023.acl-long.691. <https://aclanthology.org/2023.acl-long.691>.
- [92] M.A. Stranisci, E. Bernasconi, V. Patti, S. Ferilli, M. Ceriani and R. Damiano, The World Literature Knowledge Graph, in: *The Semantic Web–ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings*, 2023b.
- [93] M.A. Stranisci, E. Bassignana, P.-L. Cabot, R. Navigli et al., Dissecting Biases in Relation Extraction: A Cross-Dataset Analysis on People’s Gender and Origin, in: *GeBNLP 2024-5th Workshop on Gender Bias in Natural Language Processing, Proceedings of the Workshop*, Association for Computational Linguistics (ACL), 2024, pp. 190–202.
- [94] B. Stroube, Literary freedom: Project gutenber, *XRDS: Crossroads, The ACM Magazine for Students* **10**(1) (2003), 3–3.
- [95] J. Sun and N. Peng, Men are elected, women are married: Events gender bias on wikipedia, *arXiv preprint arXiv:2106.01601* (2021).
- [96] H. Suresh and J.V. Guttag, A framework for understanding unintended consequences of machine learning, *arXiv preprint arXiv:1901.10002* **2**(8) (2019), 73.
- [97] S. Tedeschi and R. Navigli, Multinerd: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation), in: *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 801–812.

- 1 [98] S. Tedeschi, F. Friedrich, P. Schramowski, K. Kersting, R. Navigli, H. Nguyen and B. Li, ALERT: A Comprehensive 1
2 Benchmark for Assessing Large Language Models' Safety through Red Teaming, *CoRR* (2024). 2
- 3 [99] B. Tillett, What is FRBR? A conceptual model for the bibliographic universe, *The Australian Library Journal* 54(1) 3
4 (2005), 24–30. 4
- 5 [100] J.A. Tuominen, E.A. Hyvönen and P. Leskinen, Bio CRM: A data model for representing biographical data for prosopo- 5
6 graphical research, in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, 6
7 CEUR Workshop Proceedings, 2018. 7
- 8 [101] M.M. Voit and H. Paulheim, Bias in Knowledge Graphs—An Empirical Study with Movie Recommendation and Different 8
9 Language Editions of DBpedia, in: *3rd Conference on Language, Data and Knowledge*, 2021. 9
- 10 [102] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* 57(10) 10
11 (2014), 78–85. 11
- 12 [103] R.L. Walkowitz, The location of literature: The transnational book and the migrant writer, *Contemporary Literature* 12
13 47(4) (2006), 527–545. 13
- 14 [104] K. Weathington and J.R. Brubaker, Queer identities, normative databases: Challenges to capturing queerness on 14
15 Wikidata, *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1) (2023), 1–26. 15
- 16 [105] A.Z. Yu, S. Ronen, K. Hu, T. Lu and C.A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous 16
17 biographies, *Scientific data* 3(1) (2016), 1–16. 17
- 18 [106] A. Zeldes, The GUM corpus: Creating multilayer resources in the classroom, *Language Resources and Evaluation* 51(3) 18
19 (2017), 581–612. 19
- 20 [107] J. Zhao, T. Wang, M. Yatskar, V. Ordonez and K.-W. Chang, Gender Bias in Coreference Resolution: Evaluation and 20
21 Debiasing Methods, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for 21
22 Computational Linguistics: Human Language Technologies*, Vol. 2, 2018. 22
- 23 [108] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba and S. Fidler, Aligning books and movies: 23
24 Towards story-like visual explanations by watching movies and reading books, in: *Proceedings of the IEEE international 24
25 conference on computer vision*, 2015, pp. 19–27. 25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51