Representing the Under-Represented: a Dataset of Post-Colonial, and Migrant Writers

Marco Antonio Stranisci
Department of Computer Science, University of Turin, Italy

Viviana Patti
Department of Computer Science, University of Turin, Italy

Rossana Damiano
Department of Computer Science, University of Turin, Italy

Abstract

In today's media and in the Web of Data, non-Western people still suffer a lack of representation. In our work, we address this issue by presenting a pipeline for collecting and semantically encoding Wikipedia biographies of writers who are under-represented due to their non-Western origins, or their legal status in a country. The two main components of the ontology will be described, together with a framework for mapping textual biographies to their corresponding semantic representations. A description of the data set, and some examples of biographical texts conversion to the Ontology Classes, will be provided.

2012 ACM Subject Classification Information systems → Ontologies

Keywords and phrases Ontologies, Knowledge Graph, Language Resources, Migrations

Digital Object Identifier 10.4230/OASIcs.LDK.2021.7

Funding The work of Marco Antonio Stranisci and Viviana Patti is partially funded by the project "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call).

1 Introduction

Social media, and other User Generated Content platforms have given voice to an unprecedented number of people, while the Semantic Web offers encyclopedic knowledge about the world in an open, machine readable format. However, such technological transformation has not completely resulted in a more pluralist communicative environment, because the voices of people from non-Western countries are often unheard in crucial contexts. For instance, the involvement of minority journalists in mainstream newspapers is an open issue [23], as long as the integration of post-colonial perspectives within school textbooks [21]. This under-representation could be problematic since it precludes a full appreciation and understanding of diversity in our society.

The Under-Represented Writers (URW) project[1] aims at reducing this under-representation through a semantic modeling of authors whose biographies are characterized by belonging to a former colony country or being migrant. Its aim is twofold: encoding their lives in a non-stereotypical way, and providing a publicly available, semantically encoded knowledge source about them.

Modeling the biography of a writer who is potentially under-represented due to his ethnicity raises two main issues addressed in our work. (1) The interaction of the attributed ethnicity of a person, which is a subjective concept, with her/his legal status in the place where she/he lives. The project relies on an ontology to provide an explicit and objective

[1] The project is available at `https://w3id.org/UnderRepresentedWritersOntology/`

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 7; pp. 7:1–7:14
OASICS OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

representation of this interplay, further specialized to describe the citizenship laws of a particular set of countries. (2) The knowledge extracted from Linked Data sources in the form of RDF triples does not allow arranging the legal statuses of a person in a coherent whole during her/his lifetime, since it relies on a set of vocabularies that have not been designed to express this type of knowledge. However, representing the biography of a writer in terms of the relations with different countries along time is crucial because it allows interpreting her/his literary production in the light of the social context in which she/he was situated when she/he created them.

The Under-Represented Writers (URW) ontology was used to gather a collection of writers, and their biographies in English language from Wikidata, and DBpedia. The resulting Knowledge Graph expresses an ordered and systematic list of features about authors' birthplace and time of birth, together with their legal statuses along time, all the facts about their lives gathered from Wikipedia, and a mapping of verbs in their biographies according to the ERE ontology [3]. Writers' countries of birth are classified along three dimension: their status of former colony, their Human Development Index score, and their mobility score.

The paper is structured as follows: in Section 2, a review of the related work is introduced. Section 3 presents the ontology: the formalization of the interplay between ethnicity, and legal status is described in 3.1, while the representation of life events is explained in 3.2. Section 4 provides a description of the data gathering process (4.1), together with an overview of the Knowledge Graph (4.2). In Conclusion (Section 5), results and open issues are discussed.

2 Background and Related Works

The project described here relies on three main lines of research. In Section 2.1, literary and narratives theories that guided the ontology design process are presented. Then, an overview of the related work on semantic representation of biographies (Section 2.2), and event annotation models (Section 2.3) is provided.

2.1 Post-Colonial Literatures, and Post-Classical Narrative Theories

Post-Classical narrative theories [24] were born during the Eighties as an alternative to the semiotic approach to narratives. Instead of focusing only on texts, scholars started investigating the economic and political contexts in which cultural works are produced, thus highlighting the strong interconnection between the author and the cultural norms and values shaping her/his narratives ([18, 4]). This paradigm shift led to a wide set of theories, such as feminist narratology [15] and ethnic narratology [8].

Alongside the spreading of post-classical approaches to the study of narratives, the heterogeneous field of research labeled under the term "Post-Colonial studies" raised interest around the issue of under-representation of former colony country citizens' voices. "In the context of colonial production, the subaltern has no history and cannot speak" [33] due to an epistemic violence perpetrated by European countries through a systematic practice of silencing [7]. Non-dominant indigenous groups did not take part to the elaboration of their countries' official culture and history, while local élites were raised with a Western education. Besides the usage of violence, Europeans enacted a textual takeover [10] of non-Western countries by imposing their cultural traditions. During this process, colony citizens suffered a linguistic and physical displacement [1], since they lost control on their countries, and languages. Post-colonial literature is a heterogeneous cultural project aimed at reaching emancipation from the colonizer countries.

A complementary problem affects the reception of migrants narratives within the context of European countries. Recent studies showed that the fortune of novel from ethnic minorities is often related to an ethnographic interest in exoticism, and not to a need of a deep understanding of other cultures [12]. Readers, rather than being interested in the literary work itself, focus on how it conveys information about immigrant writers' ethnic identities. In response to this expectation, some of them deliver stereotypical representation of their ethnicity in their novels [22].

Our project acknowledges the relevance of political and social factors in studying the narratives produced by writers by explicitly modeling in the ontology the conditions leading to a lack of representation. More to the point, our attempt is to describe three types of biographical situations potentially correlated with under-representation: living in a former colony country, being a migrant, and belonging to an ethnic minority.

2.2 Biographic Ontologies

Two projects have tackled the issue of collecting and describing the biographies of writers, and under-represented people. The CWRC Ontology[2] [2, 31] has been developed to support the Orlando project[3]: a data set of 1,300 women British writers aimed at widening the study of feminist literary research. The ontology has an extensive taxonomy of classes describing the biography of a woman writer with the set of characteristics that determines her condition, such as ethnicity, political affiliation, reproductive history, and sexuality.
The Enslaved Ontology[4] [30] is a modular ontology aimed at mapping several databases about African slavery in a single Knowledge Graph[5] aligned with Wikidata [38]. Similarly to the Orlando project, the Enslaved Ontology models socio-cultural information about people in the data set, in order to reconstruct the social networks characterizing slavery. However, controlled vocabularies were chosen – rather than formalised concepts – to express detailed information about persons and events.

Our Ontology shares some similarities with these projects, as long it aims at representing a group of persons sharing a specific condition, such as belonging to an ethnic minority. However, the concept of "being under-represented" is challenging from the modeling perspective, because it has blurred boundaries and it can be very subjective. Furthermore, our project intentionally does not model ethnographic features, choosing instead to fully describe the interplay between a person and the places where she/he lives during her/his life.

Some proposals are specifically targeted at the representation of biographies. The Biography Ontology [14, 13] models biographical events as time-dependent knowledge by directly adding temporal arguments to a materialised triple. In the example below, the marriage between Tony Blair and Cherie Booth is first expressed, then its temporal boundaries are added, together with other optional information. Semantic conciseness characterizes such approach, which has a major drawback in the generalization of complex events determined by multiple factors.

```
    tony blair marriedTo cherie booth
‘‘1980-03-29’’xsd:date ‘‘2015-05-08’’^^xsd:date
    location London
```

[2] http://sparql.cwrc.ca/ontologies/cwrc.html
[3] http://www.artsrn.ualberta.ca/orlando/
[4] https://docs.enslaved.org/ontology/
[5] https://enslaved.org/

BKOnto [36] is built upon the Time[6] and the StoryLine[7] ontologies. Token-reified biographical events are arranged in StoryLine slots, and further decomposed to express more detailed spatial and temporal information about them. The BIO vocabulary[8] collects 34 types of life events which can be used to create a biographical timeline.

Our work is aimed at modeling time-dependent knowledge like The Biography Ontology [14], but relies on the Ontology Design Pattern approach [27, 11], since it encodes the status of a person resulting from a combination of *roles* she/he experiences in a given situation. For such reason, a rich account of modular and expressive legal statuses related to citizenship and discriminatory factors are necessary, rather than a closed taxonomy of biographical events.

The study of prosopography as a methodological tool for historical research is the object of the Factoid Prosopography Ontology (FPO) [26]. Leveraged by several projects targeted at the study of Middle Ages in Europe and Asia[9], FPO connects the representation of personal factoids (such as birth, death, acquisition of goods or social status) with the documentation about them (e.g., legal statements, seals, and other artifacts), so as to support a systematic investigation of biographies through documents. In our project, the availability of authoritative sources, stored in digital form, about the identity and status of the biographies under study makes the representation of documentary sources less relevant than it would be when dealing with ancient ages, bringing instead into focus the unambiguous definition of domain-specific notions such as citizenship and migration.

2.3 Event Encoding

Many approaches aimed at encoding and annotating events have been proposed in the last years. Despite the common representational goal, they vary significantly, since events can be formalized at different levels of granularity.

The Clinical Narrative Temporal Relation Ontology (CNTRO) [35] provides a representation of clinical events offering a comprehensive taxonomy of temporal representations mapped onto authoritative representations of time [34].

The Story Intention Graphs (SIG) [9] encodes the intentions of narrative agents by linking them to textual fragments. Its application to personal narratives has been proposed by [17]. An approach sharing a similar granularity is the one proposed by [5]. Here, events are frame-like structures of the type <Agent, Predicate, Theme, PP> whose affective polarity is annotated.

Finally, there are several works that analyze events at a word level. The ACE/ERE projects [6, 32] rely on the identification of the events through the annotation of "Trigger" words or multi-word expression. The TimeML annotation scheme [28] has been specifically designed for identifying temporal expressions in a text, and annotating the chronological relation between them. Finally, the Richer Event Description (RED) framework [25] simplifies the taxonomy of events proposed in TimeML, but adds information about the causal relationships over them.

Even if all these approaches contribute to an exhaustive representation of events in texts, a unifiying model that systematically links the syntactic and the semantic layers of an event is still missing. Our work tries to bridge together these two levels by mapping ERE's "Trigger" verbs of movement and Wordnet synsets, according to Semantic Web principles[19].

[6] https://www.w3.org/TR/owl-time/
[7] https://www.bbc.co.uk/ontologies/storyline
[8] https://vocab.org/bio/
[9] https://www.kcl.ac.uk/factoid-prosopography/projects

3 The Ontology of Under-Represented Writers

The Ontology of Under-represented Writers (URW) is an attempt to provide a formal and objective description of authors who potentially suffer a lack of prominence due to the context where they were born. The URW is aimed at highlighting the biographical events, and situations during which a writer may has been experienced the condition of being a subaltern [33].

With these goals in mind, we designed the URW Ontology to answer the following Competency Questions:

- What people, born in a former colony country, wrote at least one literary work while living in their birthplace?
- Which are the writers who experienced the condition of being migrant?
- Which second generation migrants or minorities wrote at least one literary work?

In the next sections, a description of how our semantic model fits with the first two Competency Questions is provided. The encoding of second generation migrants, and minorities and is not addressed in this work, due to the lack of information about ethnicity, rarely specified in Wikidata (Section 4).

3.1 Modeling the Interplay between Ethnicity and Legal Status of a Person

Since the post-colonial framework is irreducible to a unifying taxonomy (see Section 2.1), it is crucial to describe the concept of under-representation related to ethnicity without falling into arbitrary categorizations. Hence, we propose an agnostic model that operates through the intersection of two elements: the country where a person was born, and information about her/his family relationships.
The country of birth is the first feature to express under-representation. Instead of citizenship and ethnicity, which may change along time (the former), and be subjective (the latter), the fact of being born in a given place is an immutable, closer to objectiveness property. So, concerning the country of birth, we defined three indicators that may correlate with the under-representation of a writer:

- its status as a former colony;
- the mobility score of its passport[10];
- its Human Development Index[11], aimed at excluding rich former colonies from the collection, such as Singapore or Israel.

Finally, family relationships are used to determine whether a person is a second generation migrant or if she/he belongs to an ethnic minority in a given country. The interplay of these two features (country and family) helps to determine the condition of under-representation. However, this model can be adapted to other domains of knowledge in which a representation of the relation of a person, her/his birth country, and her/his family network is needed. In Figure 1, a graphical representation of Chimamanda Ngozi Adichie (a contemporary Nigerian writer) birth country is provided. In this case, the simultaneous membership of Nigeria to sets of different classes (URW:COUNTRYWITHLOWHDI, URW:COUNTRYWITHLOWMOBILITYSCORE, and URW:FORMERCOLONYCOUNTRY) is a signal of a *potential* lack of representation of this author.

[10] https://www.passportindex.org/

[11] http://hdr.undp.org/en/content/human-development-index-hdi

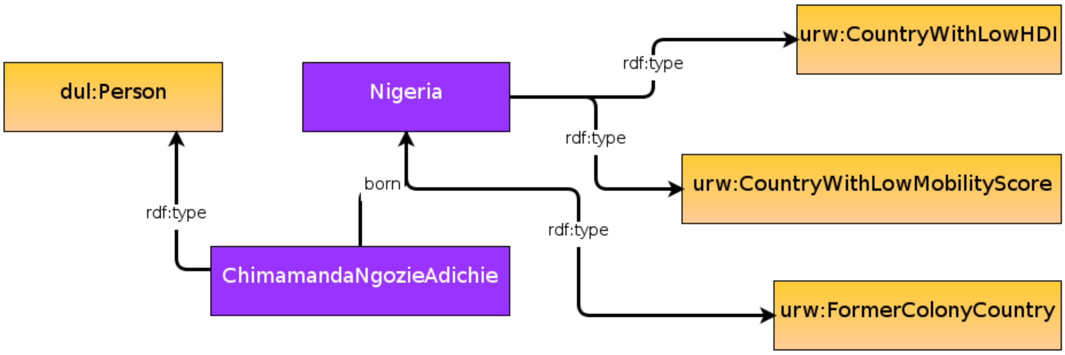


Figure 1 The representation of Chimamanda Ngozie Adichie place of birth, that lead to consider her an under-represented writer.
Within the paper, classes are represented with yellow boxes, while purple boxes identify individuals.

3.2 Modeling Biographies

As mentioned in Section 2.2, there are some proposed approaches for encoding a biography. For our purpose of modeling the life of an under-represented person, two kinds of situations need to be described: the process of migrating, and the status of a person in a given country. Both are embodied in a specific time interval, and this relation of time-dependency need to be formally expressed for two reasons: on one side, it is essential to order life events in a chronological fashion; on the other side, it allows drawing a link between a writer's life, and her/his cultural production. Our solution relies on the Ontology Design Pattern (ODP) framework, since it provides foundationally sound, re-usable building blocks for representing common patterns across ontologies, with advantages for design and interoperability. More specifically, we adopted the BasicPlanExecution ODP to describe a migration, since a migration represents the execution of a intentionally devised line of action, and the TimeIndexedPersonRole for modeling the legal status of a person, because the legal status of a person with respect to a country is typically non-rigid and can be modelled as a role.

The URW:MIGRATION class (see Figure 2) is subclass of the DUL:PLANEXECUTION class and, as such, **dul:isSettingFor** six elements: the action of URW:MIGRATING, which is an event, a DUL:PERSON, namely the agent who is migrating, and her/his URW:MIGRATIONROLE in the migration process. The URW:MIGRATION class is also a setting for the spatio-temporal coordinates of the migration: the DUL:TIMEINTERVAL along which it occurs, the URW:PLACEOFARRIVAL of the migration, and its URW:PLACEOFDEPARTURE. We modeled the reasons for a person to leave her/his country for another as a URW:MIGRATIONREASON that is part of a URW:MIGRATIONPLAN satisfied by the URW:MIGRATION situation. The URW:MIGRATIONREASON **dul:defines** the URW:MIGRATIONROLE of a person.

The URW:TIMEINDEXEDPERSONSTATUS class **dul:isSettingFor** a DUL:PERSON and her/his CONDITIONROLE. Furthermore, it **dul:hasComponent** a DUL:CLASSIFICATION, which **dul:isSettingFor** a DUL:TIMEINTERVAL, a DUL:PLACE, and a URW:CONDITION. The latter is primarily used to express the legal status of a person (eg: citizen, economic migrant, refugee), but it also can be used to describe other features that determines her/his condition. For instance, religion, sexual orientation, or social class. These additional aspects currently fall outside the scope of the ontology, so they are purposely left open to the integration with other semantic resources, such as ontologies (e.g. [2]) or controlled vocabularies (e.g. [30]).

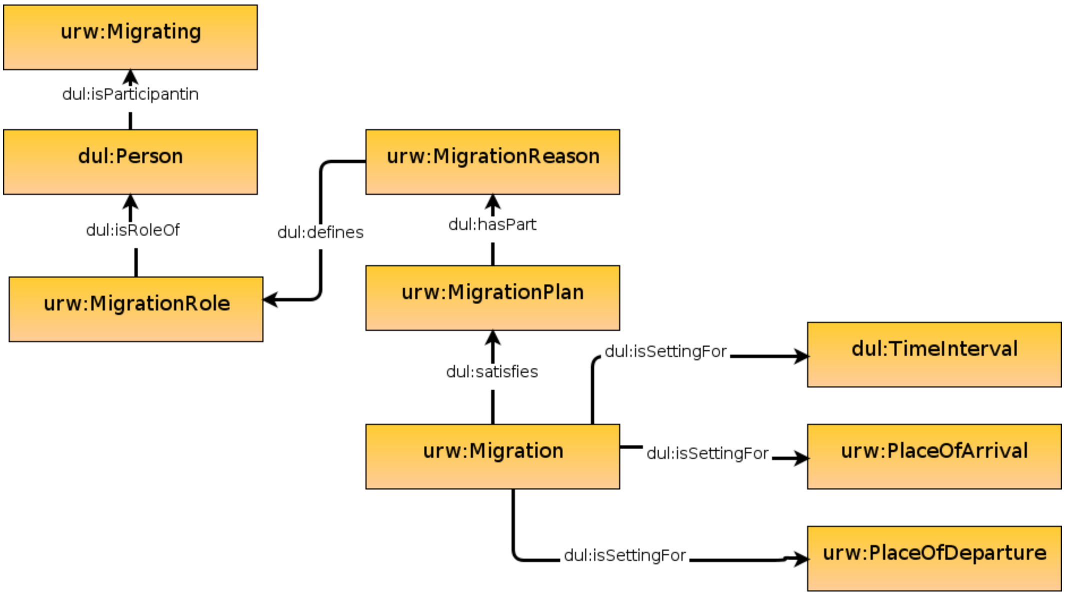


Figure 2 A graphical representation of the URW:MIGRATION class. The "urw" prefix stands for Under-Represented Writer Ontology; "dul" is the prefix of Dolce, the upper ontology which is the reference for foundational concepts in our work.

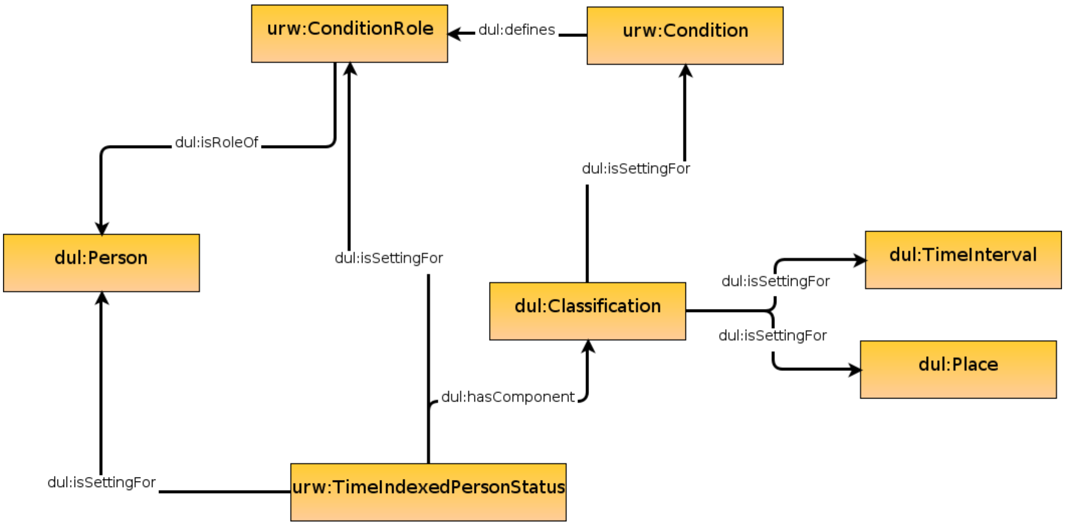


Figure 3 A graphical representation of the urw:TimeIndexedPersonStatus class.

3.3 Integration with Other Ontologies

In addition to the uw:TimeIndexedPersonStatus, and the uw:Migration, we partly integrated in the URW Ontology three other ontological resources. The above mentioned Named Authority List of countries maintained by the European Union[12]: an authoritative, comprehensive, and multilingual reference for country names. In our project, its main use is to standardize information about authors biographical places.

The PROV-O [16] Ontology is a standard to express the provenance information of a work. In this context, its use has two aims: making explicit the sources of information about writers biographies; pairing authors to their works.

The Ontolex-Lemon model [20] semantically enriches the event triggers defined within the ERE project [3], by mapping the morphological and syntactic properties of lexical entries to the semantic categories expresses by Classes. This is supposed to facilitate the process of converting the raw text of the authors' biographies to RDF triples, as described in Section 4.1.

4 The URW Knowledge Graph

In this section, we describe the first version of the URW Knowledge Graph, a collection of writers and biographies from Wikidata. After describing the data gathering process, we give an overview of the data set and provide some examples of the encoded entities.

4.1 Data Gathering

A preliminary identification of features about authors to be included in the Knowledge Graph led us to disregard ethnicity. Such information can introduce a bias in the collection because of the demographics of Wikipedia editors [37]. Moreover, the ethnic group of an author is available only in the 4.8% of the cases. The birthplace and time of birth appeared to be two widespread and objective features, instead. Therefore, data gathering has been devoted to obtain this information for each author. The data collection pipeline (see Figure 4) consisted of several steps:

- First, we collected through the Wikidata Query Service[13] all the instances (corresponding to the class WDT:P31 in Wikidata) of type human (WD:Q5), which has occupation (WDT:P106) of type novelist (WD:Q6625963), poet (WD:Q49757), or writer (WD:Q36180). We thus obtained 246,574 records.
- Then, we obtained the dates of birth (WDT:P569), by using the Pywikibot Python library[14]. In order to avoid duplicates or data misalignment, we only stored the year of birth of each writer, of available. The collection was reduced to 227,840 items.
- We then filtered only writers from a specific historical period to nowadays. We chose the Berlin Conference held in 1884, which formally started the scramble for Africa, an emblematic moment of the European textual takeover of non-Western world, directly related to the subsequent Decolonization and spreading of post-colonial literatures. 155,294 authors were born from 1884.
- Next, for each collected author, we queried her/his place of birth (WDT:P19). The information about birth place is very heterogeneous in Wikidata: it can be a country,

[12] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/\resource/dataset/country

[13] https://query.wikidata.org/

[14] The library, available at https://github.com/wikimedia/pywikibot, was also used to collect places of birth and their corresponding countries.

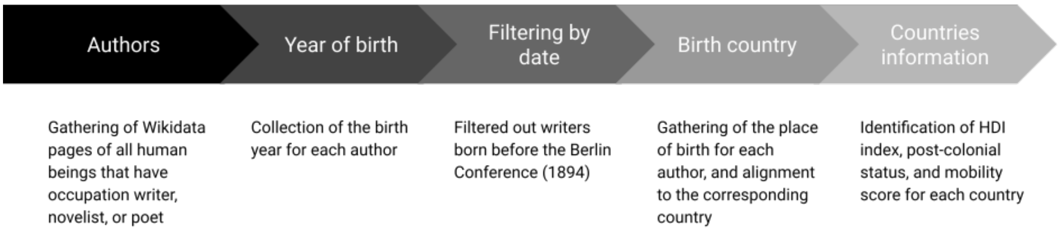

Authors
Year of birth
Filtering by date
Birth country
Countries information
Gathering of Wikidata pages of all human beings that have occupation writer, novelist, or poet
Collection of the birth year for each author
Filtered out writers born before the Berlin Conference (1894)
Gathering of the place of birth for each author, and alignment to the corresponding country
Identification of HDI index, post-colonial status, and mobility score for each country

Figure 4 The diagram representing the data gathering pipeline.

an administrative region, a city or even a district. In order to align all the birthplaces to a common format, we further queried the countries (WDT:P17) of all birthplaces. Throughout all this process, we used Europeana Eurovoc as an authoritative source to align all the geographical entities.

- Finally, we associated the Human Developed Index, the mobility score, and the eventual status of former colony to each country in order to group writers according to their level of under-representation. The resulting Knowledge Graph includes 127,141 authors.

4.2 KG Description

For each urw:Author of the Knowledge Graph the urw:CountryOfBirth, and urw:YearOfBirth are specified. Moreover, the urw:wikipediaText data property contains the reference text, expressed as a string. At a first glance (see Table 1), the URW Knowledge Graph shows an unbalanced distribution of writers across different continents[15]. More specifically, the data set seems to be Eurocentric, since the 64% of the authors were born in this continent. On the opposite side, African writers are the less represented amount of population taken into account. Finally, the number of individuals is dramatically reduced to 45,793 if we consider only the ones with a Wikipedia page in English. This drop mainly affects Europe and Latin America continents.

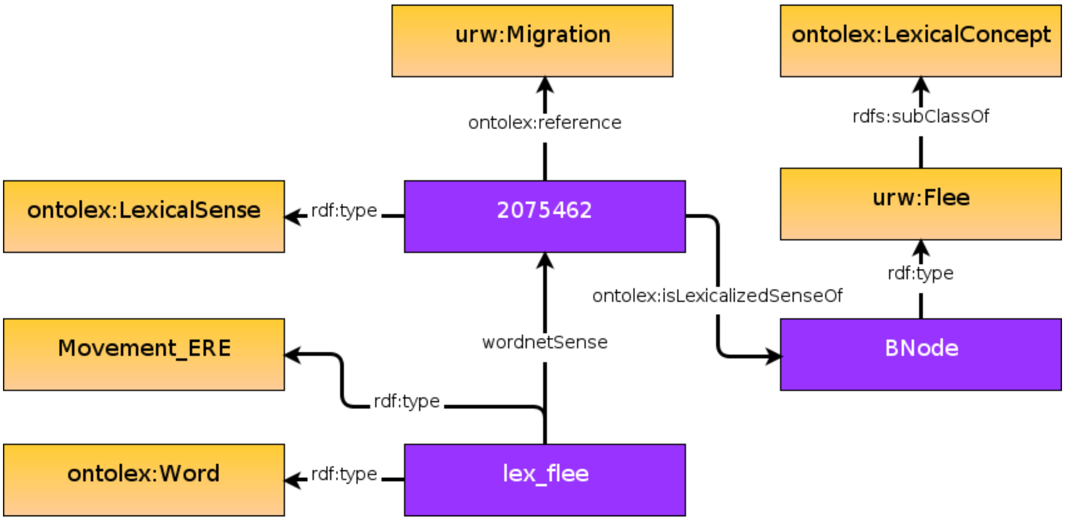
Table 1 The list of writers stored in the URW Knowledge Graph, divided by continent of birth.

Continent	Population	Writers on Wikidata	People per writer	Authors with English page
Africa	1,340,598.113	3,528	374,259.6	53.6%
Asia	4,641,054.786	14,993	309,548.1	45.9%
Europe	747,636.045	81,832	9,136.2	22.6%
Latin America	653,962.332	9,643	67,817.3	28.2%
North America	368,869.644	17,389	21,212.8	77.5%
Oceania	42,677.809	1,365	31,265.7	85.9%

4.3 Event Encoding Description and Examples

Wikipedia authors pages are concise texts providing a limited taxonomy of biographical facts: educational background, personal life events, movements, works, and awards. Such a homogeneous narrative style facilitates the extraction and encoding of migration-related

[15] Both the 6-continents model and the estimation of population by continent were taken from the UN Department of Economic and Social Affairs: `https://population.un.org/wpp/`



🟨 **Figure 5** The encoding of the Movement_ERE event trigger *flee* mapped to the corresponding Wordnet offset. The prefix "ontolex" refers to the OntolexLemon ontology adopted for the mapping.

patterns (see 3.2) from raw text using preexisting linguistic resources: REO Ontology [3], and Ontolex-Lemon [19, 20]. Our conversion process encompassed two passages: (1) The collection of all the "Trigger" verb of "Movement_ERE" events in the REO Ontology [3], and the identification their occurrences in each biography. (2) The mapping of Movement_ERE "Triggers" with Wordnet offsets in RDF triples according to the Lemon methodology [19, 20].

According to the Ontolex-Lemon specification, each ONTOLEX:LEXICALENTRY (a word, a multiword expression of an affix) has a corresponding set of ONTOLEX:LEXICALSENSE. In our case, every ONTOLEX:LEXICALSENSE is a Wordnet offset, namely the lexicalized sense of an ONTOLEX:CONCEPT. The latter represents the mental concept evoked by a lexical entry. Finally, the ONTOLEX:LEXICALSENSE has a semantic reference within the Ontology, in our case a URW:MIGRATION or a URW:TIMEINDEXEDPERSONSTATUS.
Figure 5 is a graphical representation of how the verb "to flee" is encoded in the ontology. The ontolex:lex_flee is both a ONTOLEX:WORD and an EREONTOLOGY:MOVEMENT_ERE "Trigger", namely a textual token expressing an event in which a person moves from a place to another. The ontolex:lex_flee has a ONTOLEX:LEXICALSENSE, which in our case is the Wordnet offset id number of the verb flee (2075462), related to the definition "run away quickly". The Wordnet offset is the lexicalized sense of the URW:FLEE concept evoked by the ontolex:lex_flee. Finally, the ONTOLEX:LEXICALSENSE has a **ontolex:reference** to the URW:MIGRATION class of the ontology (Section 3.2).

We provide two examples of manually encoded biographical events to illustrate the pipeline. The first, depicted in Figure 6, shows a conversion from raw text to urw:Migration Class (Section 3.2). Below, there is an excerpt of the Wikipedia page of Kim Thúy, a Vietnamese-born Canadian writer.

> At the age of ten, Thúy **left** Vietnam with her parents and two brothers, joining more than one million Vietnamese boat people **fleeing** the country's communist regime after the fall of Saigon in 1975. The Thúys **arrived** at a refugee camp in Malaysia, run by the United Nations High Commission for Refugees, where they spent four months.[16]

[16] https://en.wikipedia.org/wiki/Kim_Thúy

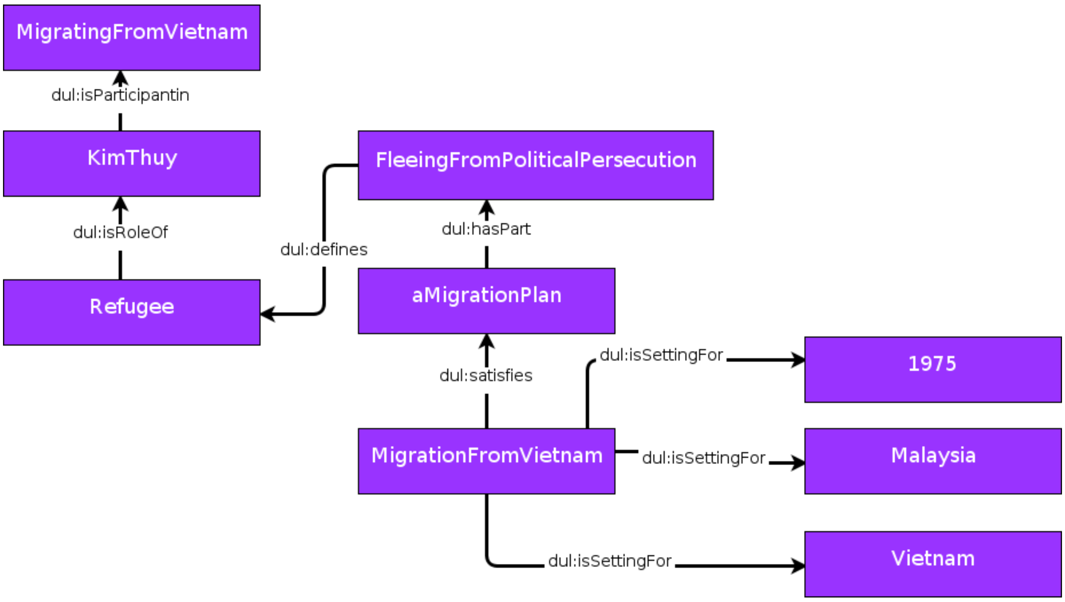


Figure 6 The encoding of a Kim Thúy childhood event as a uw:Migration.

Verbs in bold – *left*, *fleeing* (Figure 5), *arrived* – are "Triggers" of a movement event and identify a URW:MIGRATION situation and a URW:MIGRATING event, respectively labeled as urw:MigrationFromVietnam and urw:MigratingFromVietnam[17]. Vietnam, the URW: PLACEOFDEPARTURE, and Malaysia, the URW:PLACEOFARRIVAL, have been easily identified, since they are explicitly mentioned in the text. The individuation of the time when the URW:MIGRATION occurred was more difficult because a comparison between the expression "At the age of ten", and the URW:BIRTHYEAR of the writer was necessary to derive it by difference. Finally, textual references to "the fall of Saigon", and to "refugee camp" allowed identifying the URW:MIGRATIONREASON as fleeing from political persecution, and the URW:MIGRATIONROLE as a urw:Refugee.

It is worth mentioning that the occurrence of "run" in the example does not imply a movement. A more rigorous approach such as Lexico-Semantic Pattern [29] is needed to avoid false positives within an automatic conversion process.

The second example illustrates the text-to-urw conversion applied to Chimamanda Ngozi Adichie's biography (Figure 7). Again, the verb (*left*) is a trigger for the event of leaving a country for another, which allows identifying a URW:TIMEINDEXEDPERSONSTATUS instance, and the United States as the DUL:PLACE where the situation is experienced by the person. Similarly to the previous example, the beginning of the DUL:TIMEINTERVAL had to be inferred from the expression "At the age of 19". Several textual references to the concept of studying led to the individuation of the URW:CONDITION of being a urw:MigrantStudent. Finally, there are not any unambiguous verbal signals triggering the end of the experience of Adichie as a migrant student in the United States, as long as the reference to Yale University does not directly express the country which is the URW:PLACE of the URW:TIMEINDEXEDPERSONSTATUS. Again, a more sophisticated approach is needed, as it is pointed out in Section 5.

[17] This labeling was adopted to make the description clearer, since, in the Knowledge Graph, they are blank nodes

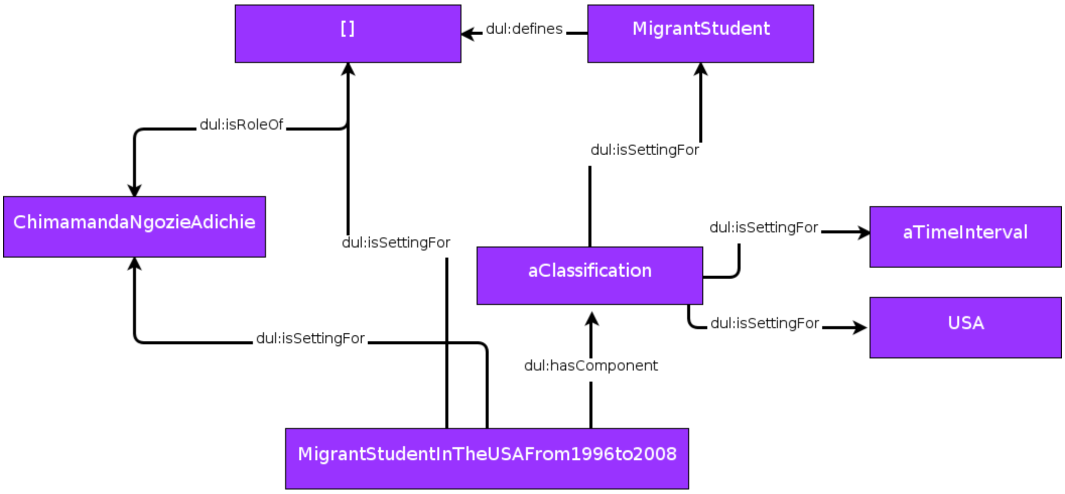


Figure 7 The encoding of the Adichie experience of migrant student in the United States according to the uw:TimeIndexedPersonStatus pattern.

> At the age of 19, Adichie **left** Nigeria for the United States to study communications and political science at Drexel University in Philadelphia.[...] In 2008, she received a Master of Arts degree in African studies from Yale University.[18]

5 Conclusion and Future Work

In this paper, we described the ongoing construction of a data set that relies on the URW ontology, a semantic model designed to encode the lives of migrant, post-colonial writers. After describing the ontology and the pipeline, we provided some examples of conversion from raw text biographies to URW:MIGRATION, and URW:TIMEINDEXEDPERSONSTATUS through the Ontolex-Lemon model. However, a systematic encoding has not been performed yet. This is a necessary step to gather, organize, and analyze narratives belonging to under-represented authors.

A first overview of the obtained Knowledge Graph shows that a lack of representation of non-Western authors is also present on Wikidata, together with the need to adopt a multilingual approach, since only the 36% of writers has an English Wikipedia page.

References

1 Bill Ashcroft, Gareth Griffiths, and Helen Tiffin. *The empire writes back: Theory and practice in post-colonial literatures.* Routledge, 2003.

2 Susan Brown, Patricia Clements, Isobel Grundy, Sharon Balazs, and Jeffrey Antoniuk. An introduction to the orlando project. *Tulsa Studies in Women's Literature*, 26(1):127–134, 2007.

3 Susan Windisch Brown, Claire Bonial, Leo Obrst, and Martha Palmer. The rich event ontology. In *Proceedings of the Events and Stories in the News Workshop*, pages 87–97, 2017.

4 L. E. Bruni. Cultural narrative identities and the entanglement of value systems. In *Differences, Similarities and Meanings: The Interplay of Differences and Similarities in Communication and Semiotics.* De Gruyter Mouton, In press.

[18] https://en.wikipedia.org/wiki/Chimamanda_Ngozi_Adichie

5 Haibo Ding, Tianyu Jiang, and Ellen Riloff. Why is an event affective? classifying affective events based on human needs. In *AAAI Workshops*, pages 8–15, 2018.

6 George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004. European Language Resources Association (ELRA). URL: `http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf`.

7 Kristie Dotson. Tracking epistemic violence, tracking practices of silencing. *Hypatia*, 26(2):236–257, 2011.

8 Laura Doyle and Laura Anne Doyle. *Bordering on the body: The racial matrix of modern fiction and culture*. Oxford University Press on Demand, 1994.

9 David K Elson. Detecting story analogies from annotations of time, action and agency. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative, Istanbul, Turkey*, pages 91–99, 2012.

10 Leela Gandhi. *Postcolonial theory: A critical introduction*. Columbia University Press, 2019.

11 Aldo Gangemi and Valentina Presutti. Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer, 2009.

12 Graham Huggan. *The postcolonial exotic: Marketing the margins*. Routledge, 2002.

13 Hans-Ulrich Krieger and Thierry Declerck. Tmo – the federated ontology of the trendminer project. In *LREC*, pages 4164–4171. Citeseer, 2014.

14 Hans-Ulrich Krieger and Thierry Declerck. An owl ontology for biographical knowledge. representing time-dependent factual knowledge. In *BD*, pages 101–110, 2015.

15 Susan S Lanser. Toward a feminist narratology. *Style*, pages 341–363, 1986.

16 Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. Technical report, World Wide Web Consortium, 2013. URL: `https://www.w3.org/TR/prov-o/`.

17 Stephanie M Lukin, Kevin Bowden, Casey Barackman, and Marilyn A Walker. Personabank: A corpus of personal narratives and their story intention graphs. *arXiv preprint*, 2017. `arXiv:1708.09082`.

18 Dan P McAdams. Narrative identity. In *Handbook of identity theory and research*, pages 99–115. Springer, 2011.

19 John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. Integrating wordnet and wiktionary with lemon. In *Linked Data in Linguistics*, pages 25–34. Springer, 2012.

20 John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21, 2017.

21 Pia Mikander et al. Westerners and others in finnish school textbooks. *University of Helsinki, Institute of Behavioural Sciences, Studies in Education*, 2016.

22 Magnus Nilsson. Swedish "immigrant literature" and the construction of ethnicity. *Tijdschrift voor skandinavistiek*, 31(1), 2010.

23 Katsuo A Nishikawa, Terri L Towner, Rosalee A Clawson, and Eric N Waltenburg. Interviewing the interviewers: Journalistic norms and racial diversity in the newsroom. *The Howard Journal of Communications*, 20(3):242–259, 2009.

24 Ansgar Nünning. Narratology or narratologies? taking stock of recent developments, critique and modest proposals for future usages of the term. *What Is Narratology? Questions and Answers Regarding the Status of a Theory*, pages 239–75, 2003.

25 Tim O'Gorman, Kristin Wright-Bettner, and Martha 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)*, pages 47–56, 2016.

26 Michele Pasin and John Bradley. Factoid-based prosopography and computer ontologies: towards an integrated approach. *Digital Scholarship in the Humanities*, 30(1):86–97, 2015.

27 Valentina Presutti and Aldo Gangemi. Content ontology design patterns as practical building blocks for web ontologies. In *International Conference on Conceptual Modeling*, pages 128–141. Springer, 2008.

28 James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34, 2003.

29 Lama Saeeda, Michal Med, Martin Ledvinka, Miroslav Blaško, and Petr Křemen. Entity linking and lexico-semantic patterns for ontology learning. In *European Semantic Web Conference*, pages 138–153. Springer, 2020.

30 Cogan Shimizu, Pascal Hitzler, Quinn Hirt, Dean Rehberger, Seila Gonzalez Estrecha, Catherine Foley, Alicia M Sheill, Walter Hawthorne, Jeff Mixter, Ethan Watrall, et al. The enslaved ontology: Peoples of the historic slave trade. *Journal of Web Semantics*, 63:100567, 2020.

31 John Simpson and Susan Brown. From xml to rdf in the orlando project. In *2013 International Conference on Culture and Computing*, pages 194–195. IEEE, 2013.

32 Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, 2015.

33 Gayatri Chakravorty Spivak. Can the subaltern speak? *Die Philosophin*, 14(27):42–58, 2003.

34 Cui Tao, Harold R Solbrig, and Christopher G Chute. Cntro 2.0: a harmonized semantic web ontology for temporal relation inferencing in clinical narratives. *AMIA summits on translational science proceedings*, 2011:64, 2011.

35 Cui Tao, Wei-Qi Wei, Harold R Solbrig, Guergana Savova, and Christopher G Chute. Cntro: a semantic web ontology for temporal relation inferencing in clinical narratives. In *AMIA annual symposium proceedings*, volume 2010, page 787. American Medical Informatics Association, 2010.

36 Jian-hua Yeh. Towards a biographic knowledge-based story ontology system. In *Proceedings of the 2018 International Conference on Intelligent Information Technology*, pages 33–38, 2018.

37 Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A Hidalgo. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data*, 3(1):1–16, 2016.

38 Lu Zhou, Cogan Shimizu, Pascal Hitzler, Alicia M Sheill, Seila Gonzalez Estrecha, Catherine Foley, Duncan Tarr, and Dean Rehberger. The enslaved dataset: A real-world complex ontology alignment benchmark using wikibase. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3197–3204, 2020.