Mapping Biographical events to ODPs through Lexico-Semantic Patterns*

Marco Antonio Stranisci[0000−0001−9337−7250]
Valerio Basile[0000−0001−8110−6832]
Rossana Damiano[0000−0001−9866−2843]
Viviana Patti[0000−0001−5991−370X]

Dipartimento di Informatica, University of Turin, C.so Svizzera 185, Italy
marcoantonio.stranisci@unito.it
valerio.basile@unito.it
rossana.damiano@unito.it
viviana.patti@unito.it

Abstract. In this paper we present a collection of semantically-encoded biographies of authors who were born in former colony countries from 1945. The data set relies on an ontology that represents the life of an author through the two key concepts of migration from birth place and legal status in a country, both modeled on two Ontology Design Patterns: Time Indexed Person Status and Basic Execution Plan. Together with the resource, we describe a pipeline to convert the textual biographies of the authors gathered from Wikipedia into the roles experienced by them in migrations. The pipeline includes modules for linguistic preprocessing and named entity recognition, and an entity linking step relying on Wikipedia and Wikidata APIs to link places and organizations to their respective countries. A set of lexico-semantic patterns based on verb classes from the Unified Verb Index has been developed in order to extract migration-related knowledge from unseen text biographies.

Keywords: Biography · Immigration · Pattern-based information extraction · ODP.

1 Introduction

Under-representation of non-Western people is an open issue with a long tradition [24]. Ethnic minorities suffer this condition in crucial sectors of society, such as schools [13] and media players [14]. Even collaborative projects seem to be affected by cultural [28] and gender biases. For instance, [28] observed that most Wikipedia contributors are European and male, and this may have an influence on the creation of contents on this platform [26].

Our work addresses this topic by providing structured knowledge about writers who suffer a lack of representation on Wikipedia due to their ethnic origin [25]. In this paper, we present a pipeline for the automatic extraction of

biographical events from Wikipedia through the adoption of Lexico-Semantic Patterns [4]; biographical events are semantically described by referring to the Ontology Design Patterns (ODP) framework [5]. The development of a mapping from raw-text biographies to semantic categories is a preliminary step to linking the literary production of under-represented writers to their lives.

The paper is structured as follows. In Section 2 we discuss the Linked Data projects that inspired our work, and review the state-of-the-art approaches to event extraction and encoding. In Section 3 we present the Ontology of Under-Represented Writers, describing how we encoded their biographies through recurrent semantic patterns, and how we modeled the interplay between the authors and their places of birth. In Section 4, we present the pipeline for the automatic extraction of biographical events through Lexico-Semantic patterns. Finally, in Section 5, we analyze, and evaluate results. A discussion about open issues and future work concludes the paper.

2 Related Work

In recent years, thanks to the availability of sources in a digital form, a new interest in the study of biographies has arisen is literary, cultural and historical studies. In particular, three existing Knowledge Graphs share many similarities with ours: the Orlando project[1], Enslaved[2], and WeChangEd[3]. The Orlando Project is a collection of biographies of $1,300$ British women writers; Enslaved is a data set of $509,783$ people of the historical slave trade developed from 8 preexisting archives; WeChangeEd is a collection of $1,800$ female editors born between 1710 and 1920, aligned with Wikidata. All these data sets rely on Semantic Web technologies [22,21,27], which are used to represent socio-demographic information about the individuals, such as ethnicity, family relationships, and social status.

The URW project has a similar perspective to these projects in terms of aiming to represent a group of persons sharing a specific condition. However, the concept of "being under-represented" is challenging to model, because it has blurred boundaries and it can be very subjective. Our project intentionally does not rely on a taxonomy of ethnicities, choosing instead to fully describe the interplay between a person and the places where they lived their life, in order to avoid a Western representation of non-Western writers biographies.

Several approaches aimed at encoding and annotating events have been proposed in the last years. Despite the common representational goal, these approaches vary significantly, since events can be formalized at different levels of granularity. The Biography Ontology [8], part of an ontology network within the TrendMiner project [7], models biographical events as time-dependent knowledge by directly adding temporal arguments to the materialised triples.

[1] http://www.artsrn.ualberta.ca/orlando/
[2] https://enslaved.org/
[3] https://www.wechanged.ugent.be/

Other works analyze events at a word level. The ACE/ERE projects [2][23] rely on the identification of the events through the use of a lexical 'Trigger'. The TimeML annotation scheme [17] has been specifically designed for identifying all the temporal expressions in a text, and annotating the chronological relation between them. The Richer Event Description (RED) framework [15] simplifies the taxonomy of events proposed in TimeML, but adds information about the causal relations over them.

Biographical event extraction from raw text is the subject of works relying on Wikipedia as a source of knowledge. The Pantheon 1.0 data set [28] is a collection of 11,341 biographies available in more than 25 languages in Wikipedia. Individuals in the data set have been categorized according their occupation by using a controlled vocabulary relying on Freebase. Information about the number of page views for each biography is provided as a way to measure its popularity.

Other projects have attempted to extract time and geographical information from biographical texts. Russo et al. [18] collected 782 biographies of people deported to Nazi concentration camps, extracting relevant dates and places of their lives. Then, all information has been arranged into a structured representation by using the TimeML framework [17]. The RAMBLE ON application [12] takes as input a biographical raw text, and automatically detects Motion frames [1] together with the georeferencing of each place mentioned in frames.

Our proposal aims at extracting geographical knowledge and life events jointly, to provide a semantic model for representing biographies. Unlike existing approaches, which are focused on detecting the lexical entries triggering an event [17], our work provides a mapping between the textual and the semantic level. Biographical patterns, encoded by adopting the ODP framework, are populated extracting semantic knowledge from raw text biographies.

3 A Semantic Model for Under-Represented Writers

The semantic model is designed with the purpose of providing a formal and objective description of authors who are potentially under-represented due to the context where they were born. In particular, it encodes biographical events and situations in which a DUL:PERSON is: (i) a writer and (i) has experienced the condition of being under-represented. In this way, a correlation between biographical events and literary production of under-represented authors can be drawn, and employed to gain insight on the motivations and themes reflected in their narratives. The main components of this semantic model are: the condition of being under-represented and the identification of objective criteria to classify countries which correlate with this condition.

Biographical patterns. According to our formalization, a writer who is under-represented is a person who published one or more literary works, and may have experienced the process of migrating, intended as the movement from a country to another, and the condition of living in a given country after leaving one's place of birth. Within the latter situation, the author's legal or professional

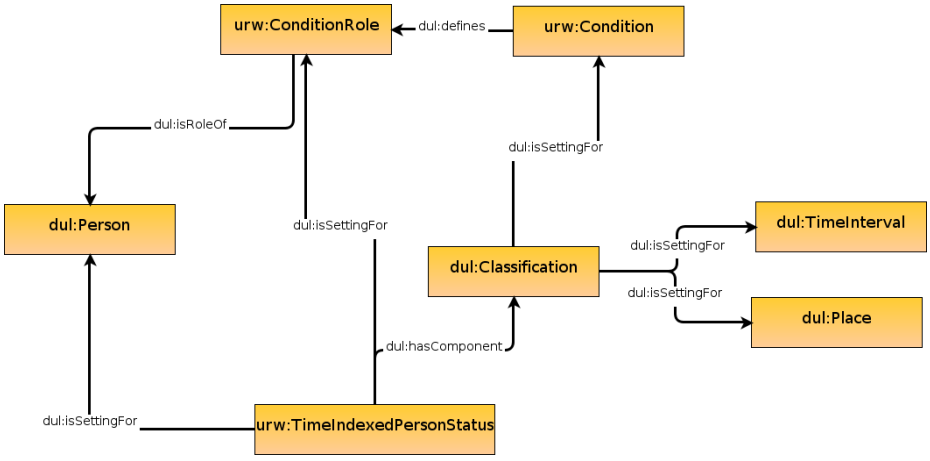

Fig. 1. A graphical representation of the urw:TimeIndexedPersonStatus pattern

status may be expressed. Our solution to encode these situations draws from the ODP framework, which provides foundationally sound, re-usable building blocks for representing common patterns across ontologies. More specifically, the URW:MIGRATION pattern refers to the BasicPlanExecution ODP [5], since a migration represents the execution of a intentionally devised line of action. The legal status of a person, URW:TIMEINDEXEDPERSONSTATUS (TIPS), relies on the TimeIndexedPersonRole ODP [16], since this condition is typically subject to change and can be modelled as time-bounded role. As can be seen in Figure 1 and 2, both Migration and TIPS describe situations that are the setting for an entity of the type DUL:PERSON, which refers to person according to the commonsense intuition, with a DUL:ROLE. A role in a TIPS is a URW:CONDITIONROLE, defined by one or more URW:CONDITIONS, such as being a foreign student, a worker, a refugee. Since multiple conditions could co-occur in defining a role, each of them has setting in a separate DUL:CLASSIFICATION situation. The URW:MIGRATION ROLE in the Migration pattern is defined by a URW:MIGRATIONREASON, namely the reason of the plan of migrating (e.g.: fleeing war, seeking for a job). Both situations are time-indexed, and take place in one or more specific URW:PLACE.

Integration with Existing Resources. In addition to the Migration and TIPS patterns, existing resources have been integrated in the semantic model: geographical resources for identifying the countries correlated with the lack of representation, and linguistic resources for mapping raw text biographical facts to the ontology. In fact, the TIPS and Migration patterns do not provide themselves a criterion to identify the under-representation, since they only portray the condition of living outside one's country. However, an author such as Italo Calvino, who was born in Italy and moved to France during his life should not be considered as under-represented, since his birthplace was a wealthy European

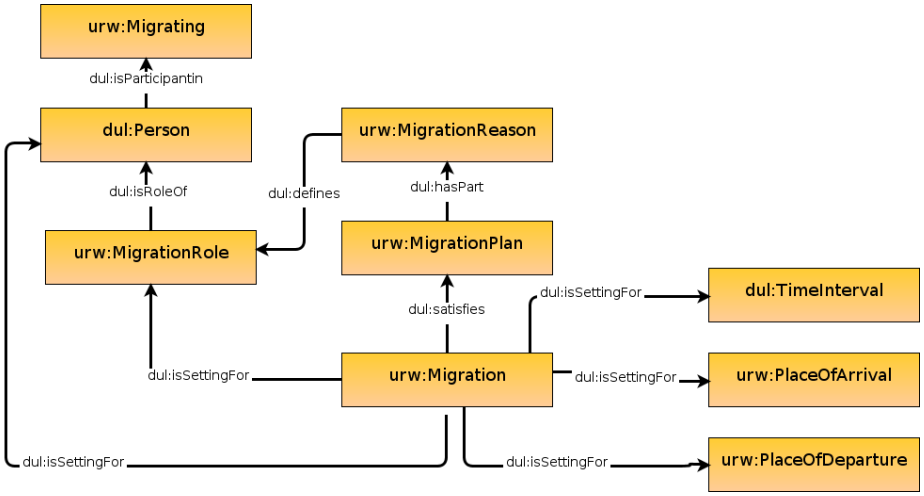

Fig. 2. A graphical representation of the URW:MIGRATION pattern.

country. Hence, three indicators have been encoded in the ontology to identify a country as under-represented:

- the country's colonial past;
- its Human Development Index (HDI)[4], a measure of the the global development of countries provided by the United Nations;
- its mobility score[5], namely the number of countries where a person could travel with the passport of the country.

In our formalization, an under-represented country must be a former colony, it must have a medium or lower HDI (below 0.8), and it must fall within the second half of the ranking of countries by mobility score. The Named Authority List of countries maintained by the European Union[6], an authoritative, comprehensive, and multilingual reference for country names, has been used to standardize and index all these sources of geographical knowledge.

Concerning the linguistic resources, we rely on the Ontolex-Lemon model, which [11] plays the function of mapping the morphological and syntactic properties of lexical entries to the semantic categories expressed by OWL classes. The use of this models facilitates the process of converting the raw text of the authors' biographies into RDF triples by maintaining the lexico-semantic information in the final representation, as described in Section 4.

Finally, the PROV-O Ontology [9] is a standard to express the provenance information of a work. In the context of our research, this model is used to identify

[4] http://hdr.undp.org/en/content/human-development-index-hdi
[5] https://www.passportindex.org/
[6] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/\resource/dataset/country

the LSPs as PROV:SOFTWAREAGENT, and the textual Wikipedia biographies as the source of knowledge from which biographical patterns have been derived.

4 From Ontology Patterns to Lexico-Semantic Patterns

Before collecting the biographies from Wikipedia, under-represented writers have been identified through the *occupation* Wikidata property (WDT:P106). Each person who worked as a writer, novelist, or poet has been collected and classified by retrieving the country of origin associated to her/his *birthplace* (WDT:P19). For each author, the biography in English language, if present, has been retrieved from Wikipedia. The total amount of collected person entities is 114, 675. Writers who were born from 1945 on, in any Asian or African under-represented country (see Section 3) have been chosen to highlight only on biographies of people who experienced or born after the Decolonization process.

Starting from this initial corpus, a pipeline to convert raw texts biographies in TIPS, and Migration classes based on Lexico-Semantic Patterns (LSP) [6,4] has been developed. LSPs are rules composed of semantic and syntactic elements related to classes and properties of an ontology. When a rule matches a string of text, the ontology is automatically populated with one or more RDF triples. An example of a Lexico-Semantic Pattern, created to extract geographical information from text, is the following [19]:

> The rule *$subject : Concept COMP RB? IN? $object : Concept* matches the phrase *Administrative territory of Prague is divided into localities* retrieving a mereological relation between *Prague*, and *localities* to be stored in an ontology.

Our pipeline is based on three steps: text parsing, LSP development, Information Extraction.

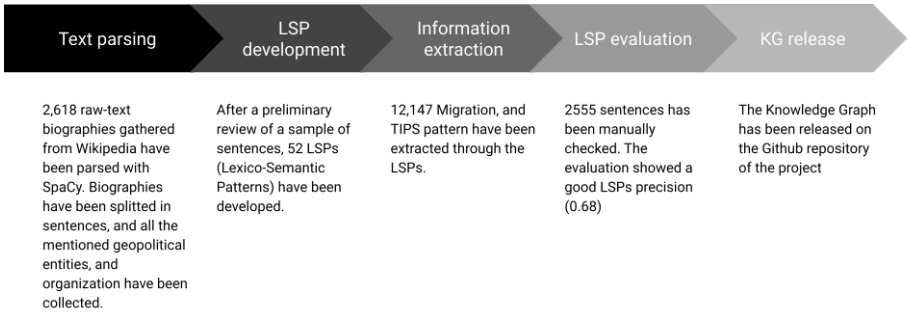

Fig. 3. The diagram representing the information extraction process in URW

Text parsing. Using the SpaCy library[7], each biography has been split in sentences, and only the ones containing at least one entity of the type Organization (ORG), or Geopolitical Entity (GPE) have been stored in JSON format, together with the name of the author, and her/his country and year of birth. Below, there is an example of an item, in JSON format, referring to the Nigerian writer, and radio presenter Dotun Adebayo:

```
{
    author: Dotun Adebayo,
    birthPlace: Nigeria,
    birthYear: 1960,
    places: [(Stationers' Company's Comprehensive School,
    ORG),(Stockholm University)],
    sentence: He then went on to Stationers' Company's Com-
    prehensive School in Hornsey, North London, followed by
    Stockholm University, where he studied Literature.
}
```

In parallel, each ORG and GPE has been linked with the respective country. All the strings identified as geopolitical entities or organizations by the SpaCy Named Entity Recognition module have been used as an input for search through the Wikipedia API. The first 10 results of the search have been subsequently analyzed, and, among them, the first candidate that holds the Wikidata property 'country' (WDT:P17) has been selected, if any. Only the 25,554 sentences containing an ORG or GPE belonging to different countries than the birth country of an author have been selected for the next step.

LSP development. After the sentences have been collected, a random subset of them has been analyzed in order to define LSP rules for encoding the biographic facts contained in the raw text into the two main patterns of the URW ontology: URW:TIMEINDEXEDPERSONCONDITION (TIPS), and URW:MIGRATION.

Given the structure of these patterns, three key elements have been identified as necessary in an input sentence to make it a candidate trigger: a **verb** expressing a change of place or a condition (e.g.: **fleeing** a country, **obtaining** a graduation), a **preposition**, and an **entity** of the type Organization (ORG) or Geo Political Entity (GPE) belonging to a different country from the place of birth. For instance (see Figure 4), the elements in bold face in the sentence "He [Dotun Adebayo] then **went** on **to Stationers' Company's Comprehensive School** in Hornsey, North London." match the pattern *(escape-51.1-1)(to—at—in)(GPE—ORG)*. So, from this rule, the following RDF triples are extracted:

```
[ a urw:Migration;
    dul:isSettingFor [
    a urw:MigrationRole;
    dul:isDefinedIn Study_Abroad
    dul:isRoleOf Dotun_Adebayo.
```

[7] https://spacy.io/

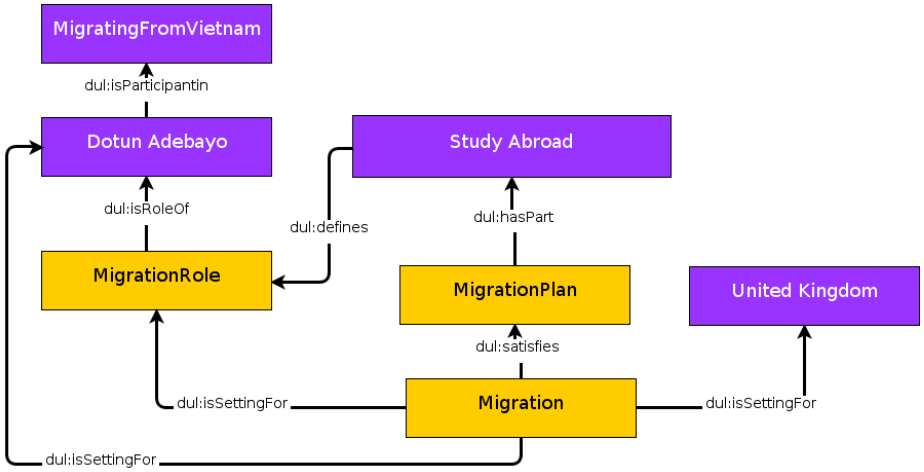

Fig. 4. A diagram representing the extraction of the Migration pattern related to Dotun Adebayo's biography.

```
    ]
    dul:isSettingFor England;
    dul:isSetting For Dotun_Adebayo;
    prov:wasDerivedFrom Wikidata;
    prov:wasAttributedTo [
    a LexicoSemanticPattern
    urw:hasPattern (escape-51.1-1)(to|at|in)(GPE|ORG)
    ]
]
```

The subsequent step in the definition of the LSPs has been the clustering of verbs through a mapping to general verb types, aimed at reducing the number of patterns and increasing their recall. To do so, we employed the Unified Verb Index[8], a repository resulting from the mapping of several lexical resources that provides syntactic and semantic frames of English verbs. In particular, we linked the verbs in our data to the VerbNet classes in Unified Verb Index (UVI) [20]. In the previous example, the relevant class for mapping movement verbs onto the Migration ontology patters is *escape-51.1-1-1*, which includes the following lemmas: *depart*, *disembark*, *escape*, *exit*, *flee*, *leave*, *vacate*.

As anticipated in Section 3, the mapping between LSPs and VerbNet classes is expressed in the ontology through the Ontolex-Lemon specification [10,11]. According to this model, each verb is an ONTOLEX:LEXICALENTRY with a corresponding set of ONTOLEX:LEXICALSENSEs (WordNet offsets [3]), which represent the lexicalized sense of the ONTOLEX:LEXICALCONCEPT, namely the VerbNet class. The ONTOLEX:LEXICALCONCEPT is the bridge between the lexical entries and the ontology classes. For instance, the ONTOLEX:LEXICALENTRY *lex_leave*

[8] https://uvi.colorado.edu/

has a corresponding ONTOLEX:LEXICALSENSE which is the *v#2009433* WordNet offset. The latter is one of the possible lexicalization of the *escape-51.1-1-1* VerbNet class, which is the ONTOLEX:CONCEPT.

Information Extraction. After the creation and refinement of the LSPs, 53 rules of the form:

VerbNet class $preposition GPE—ORG

have been formulated and applied to the annotated sentences.
The following is an example of how the same LSP matches sentences with different verbs and preposition, and encodes them as URW:TIMEINDEXEDPERSON STATUS:

LSP: obtain-13.5.2 from|for|at|by|in|as GPE|ORG

Ajunwa[9] ***received*** *her BA* ***at University of California****, Davis in 2003.*

He held a master's degree in Theatrical Directing which ***he obtained from the University of Sofia***

5 Analysis and evaluation of the results

From the life events encoding pipeline (Section 4) 12,147 sentences containing an instance of TIPS, Migration, or both have been obtained.[10]
Some preliminary statistics can help to assess the relevance of the data stored in the KG. The resulting Knowledge Graph includes 2,618 different authors' biographies, place of birth, and year of birth. 1,638 of these authors were born in Asia, 980 in Africa. In total, 39,167 RDF triples have been stored in the data set.

In order to test the precision of the LSP (Lexico-Syntactic Patterns), we manually evaluated a random sample of 2,555 sentences, which correspond to the 10% of sentences containing at least one GPE or ORG different from the author country of birth. Each sentence was labelled as expressing a 'TIPS—Migration' (48.5%) or 'None' (51.5%), then compared with the patterns.

Table 1 shows that LSPs performed with a precision of 0.68. The manual analysis of prediction errors revealed they had several causes. In some cases, the subject was not the author but another person (e.g., '**His father** left India

[9] https://en.wikipedia.org/wiki/Ifeoma_Ajunwa

[10] The first version of the data set is publicly available on the GitHub repository of the Under-Represented Writers project https://w3id.org/UnderRepresentedWritersOntology/. 10,569 sentences expressing at least one TIPS were detected, 3,549 with Migration patterns. In 1,971 cases both are present in the same sentences.

Table 1. Results of the evaluation of biographical patterns

Pattern	Precision
TIPS	0.665
Migration	0.805
TIPS and Migration	0.68

in early 1963 to study at Oxford University'). Another source of error is the presence of reported speech of the writer (e.g., 'Members of her **African audience have asserted** that Thiam does not understand why women may support FGM'). Finally, both the NER and the entity linking pipeline seem to introduce false positives (e.g., in the phrase 'Shatrughan Sinha, has also spoken in Kumar's favour on **Twitter**', Twitter is marked as an organization in the United States). It is important to mention an imbalance in the performance of the two biographical patterns: Migration situations are retrieved with a precision of 0.805, in line with recent findings from the literature [19], while precision for TIPS is 0.665. This difference is probably due to the nature of the latter pattern, which is highly heterogeneous and needs a deeper analysis to specialize it into specific patterns for different status types. In order to investigate the low performance of

Table 2. Categorization of status types under TIPS pattern

Status type	Occurrences (%)
Occupation	39.2
Publications	17.8
Education	12
Awards	8.5
Social causes	8
Other	14.5

the TIPS LSP, we conducted a closer analysis of the situations encompassed by the TIPS pattern. The results (Table 2) show that the type of status described in the sentences that matched this pattern is varied: it can refer to occupation (39.2% of the manually evaluated cases), publications (17.8%), education (12%), awards (8.5%), or involvement in social causes (8%). Since these situation types are highly consistent with the URW domain, this preliminary categorization suggests that more specific rules are needed to encode this information together with a deeper specification of TIPS within the ontology, and that this ability to discriminate will improve the performance.

6 Conclusions and Future Work

In this paper we presented a pipeline to extract life events of writers born in an Asian or African Former Colony Countries from 1945 onwards from Wikipedia biographies through Lexico-Semantic Patterns. At the present stage, the data set includes 12,147 biographical events about 2,618 authors.

A manual evaluation of a sample of the results showed a good precision of Lexico-Semantic Patterns. However, some rules need to be further specialized in order to extract a taxonomy of TIPS-related conditions. Despite these limitation, it is important to underline that a pipeline based on a small set of rules has produced a relatively large corpus, from which holistic knowledge about life's narratives can be extracted, and generalized to other types of biographies. Future works must take into account the chronological arrangement of Migration and TIPS patterns within a whole biography, and generalize Lexico-Semantic Patterns to other categories of under-represented people – ethnic minorities and second generation migrants, people with other occupations – which can be collected in the URW Knowledge Graph.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley Framenet project. In: 36th Annual Meeting of the ACL and 17th Int. Conf. on Computational Linguistics, Volume 1. pp. 86–90 (1998)
2. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In: Proceedings of the 4th Int. Conf. on Language Resources and Evaluation (LREC'04). ELRA, Lisbon, Portugal (2004)
3. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, MA (1998)
4. Frasincar, F., Borsje, J., Levering, L.: A semantic web-based approach for building personalized news services. International Journal of E-Business Research (IJEBR) **5**(3), 35–53 (2009)
5. Gangemi, A., Presutti, V.: Ontology design patterns. In: Handbook on ontologies, pp. 221–243. Springer (2009)
6. IJntema, W., Sangers, J., Hogenboom, F., Frasincar, F.: A lexico-semantic pattern language for learning ontology instances from text. Journal of Web Semantics **15**, 37–50 (2012)
7. Krieger, H.U., Declerck, T.: Tmo—the Federated Ontology of the TrendMiner Project. In: LREC. pp. 4164–4171 (2014)
8. Krieger, H.U., Declerck, T.: An OWL ontology for biographical knowledge. representing time-dependent factual knowledge. In: BD. pp. 101–110 (2015)
9. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The PROV ontology. Tech. rep., World Wide Web Consortium (2013), https://www.w3.org/TR/prov-o/
10. McCrae, J., Montiel-Ponsoda, E., Cimiano, P.: Integrating WordNet and Wiktionary with Lemon. In: Linked Data in Linguistics, pp. 25–34. Springer (2012)
11. McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P.: The Ontolex-Lemon model: development and applications. In: eLex 2017. pp. 19–21 (2017)

12. Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., Lepri, B.: RAMBLE ON: Tracing movements of popular historical figures. In: Software Demonstrations of the 15th Conf. of EACL. pp. 77–80 (2017)
13. Mikander, P., et al.: Westerners and others in Finnish school textbooks. University of Helsinki, Institute of Behavioural Sciences, Studies in Education (2016)
14. Nishikawa, K.A., Towner, T.L., Clawson, R.A., Waltenburg, E.N.: Interviewing the interviewers: Journalistic norms and racial diversity in the newsroom. The Howard Journal of Communications **20**(3), 242–259 (2009)
15. O'Gorman, T., Wright-Bettner, K., Palmer, M.: Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In: Proc. of the 2nd Workshop on Computing News Storylines (CNS 2016). pp. 47–56 (2016)
16. Presutti, V., Gangemi, A.: Content ontology design patterns as practical building blocks for web ontologies. In: Int. Conference on Conceptual Modeling. pp. 128–141. Springer (2008)
17. Pustejovsky, J., Castano, J.M., Ingria, R., Sauri, R., Gaizauskas, R.J., Setzer, A., Katz, G., Radev, D.R.: TimeML: Robust specification of event and temporal expressions in text. New directions in question answering **3**, 28–34 (2003)
18. Russo, I., Caselli, T., Monachini, M.: Extracting and Visualising Biographical Events from Wikipedia. In: BD. pp. 111–115 (2015)
19. Saeeda, L., Med, M., Ledvinka, M., Blaško, M., Křemen, P.: Entity linking and lexico-semantic patterns for ontology learning. In: Harth, A., Kirrane, S., Ngonga Ngomo, A.C., Paulheim, H., Rula, A., Gentile, A.L., Haase, P., Cochez, M. (eds.) The Semantic Web. pp. 138–153. Springer, Cham (2020)
20. Schuler, K.K.: VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, University of Pennsylvania (2006)
21. Shimizu, C., Hitzler, P., Hirt, Q., Rehberger, D., Estrecha, S.G., Foley, C., Sheill, A.M., Hawthorne, W., Mixter, J., Watrall, E., et al.: The Enslaved ontology: Peoples of the historic slave trade. Journal of Web Semantics **63**, 100567 (2020)
22. Simpson, J., Brown, S.: From XML to RDF in the Orlando Project. In: 2013 Int. Conf. on Culture and Computing. pp. 194–195. IEEE (2013)
23. Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., Ma, X.: From light to rich ERE: Annotation of Entities, Relations, and Events. In: Proc. of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. pp. 89–98 (2015)
24. Spivak, G.C.: Can the subaltern speak? Die Philosophin **14**(27), 42–58 (2003)
25. Stranisci, M.A., Patti, V., Damiano, R.: Representing the Under-Represented: a Dataset of Post-Colonial, and Migrant Writers. In: 3rd Conference on Language, Data and Knowledge (LDK 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik (2021)
26. Sun, J., Peng, N.: Men are elected, women are married: Events gender bias on Wikipedia. In: Proc. of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Vol. 2)). ACL (2021)
27. Van Remoortel, M., Birkholz, J.M., Alesina, M., Bezari, C., D'Eer, C., Forestier, E.: Women editors in europe. Journal of European Periodical Studies **6**(1), 1–6 (2021)
28. Yu, A.Z., Ronen, S., Hu, K., Lu, T., Hidalgo, C.A.: Pantheon 1.0, a manually verified dataset of globally famous biographies. Scientific data **3**(1), 1–16 (2016)