



Deriving World Literatures from Wikidata and Other Communities of Readers

Berlin (Oct. 10, 2023 - Oct. 11, 2023)

Who am I?



- PhD student in Computer Science
- member of the [CCC research group](#)
- Humanities teacher (on leave)
- Activist



Research motivation

Studying and reducing the underrepresentation of non-Western writers on Wikidata and Wikipedia



a DH resource, but not a DH research



Not a critique to Wikimedia contents, but to their acritic usage in NLP



Structure of the presentation

1. Knowledge encoding
2. Building the World Literature Knowledge Graph (WL-KG)
3. Extracting biographical information from unstructured knowledge

1. Knowledge Encoding

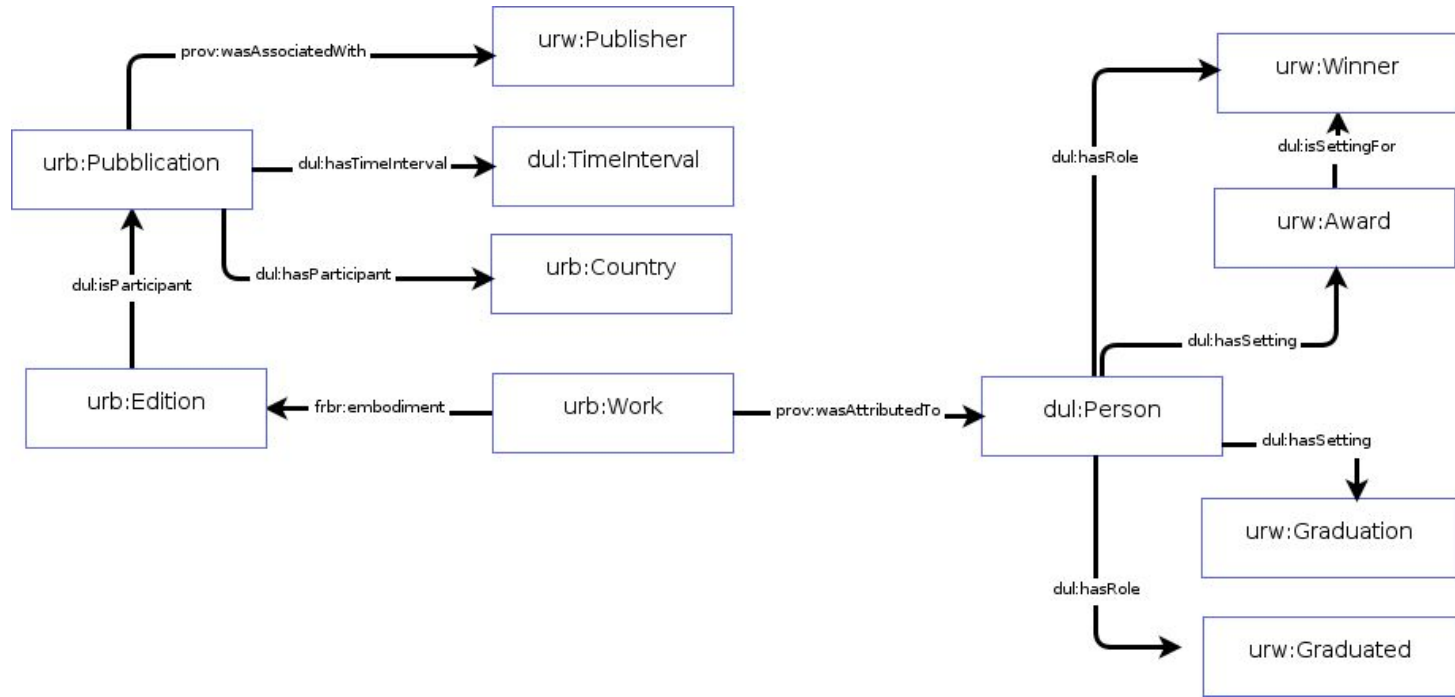


Classifying the underrepresentation

Transnational writers are:

1. People who belong to ethnic minorities in Western countries;
2. People who were born in former colonies from 1808 (Latin America) or 1917 (Africa and Asia)

Modeling writers lives and their works



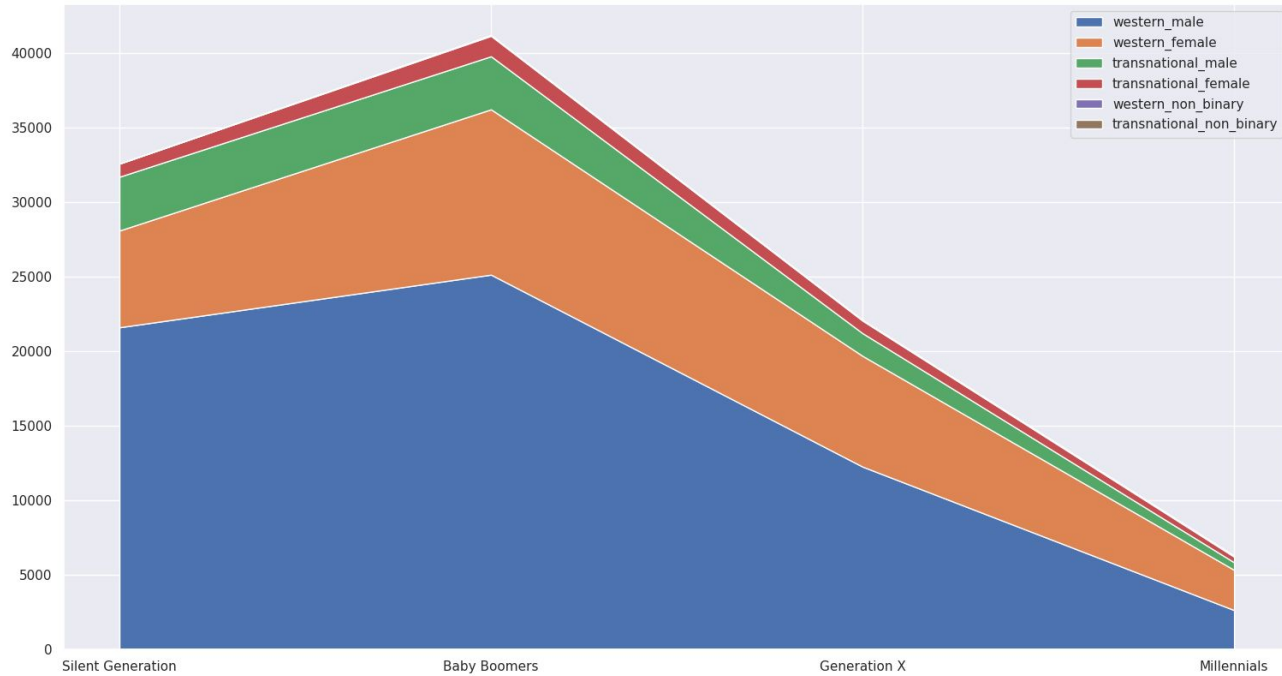
2. Building the WL-KG



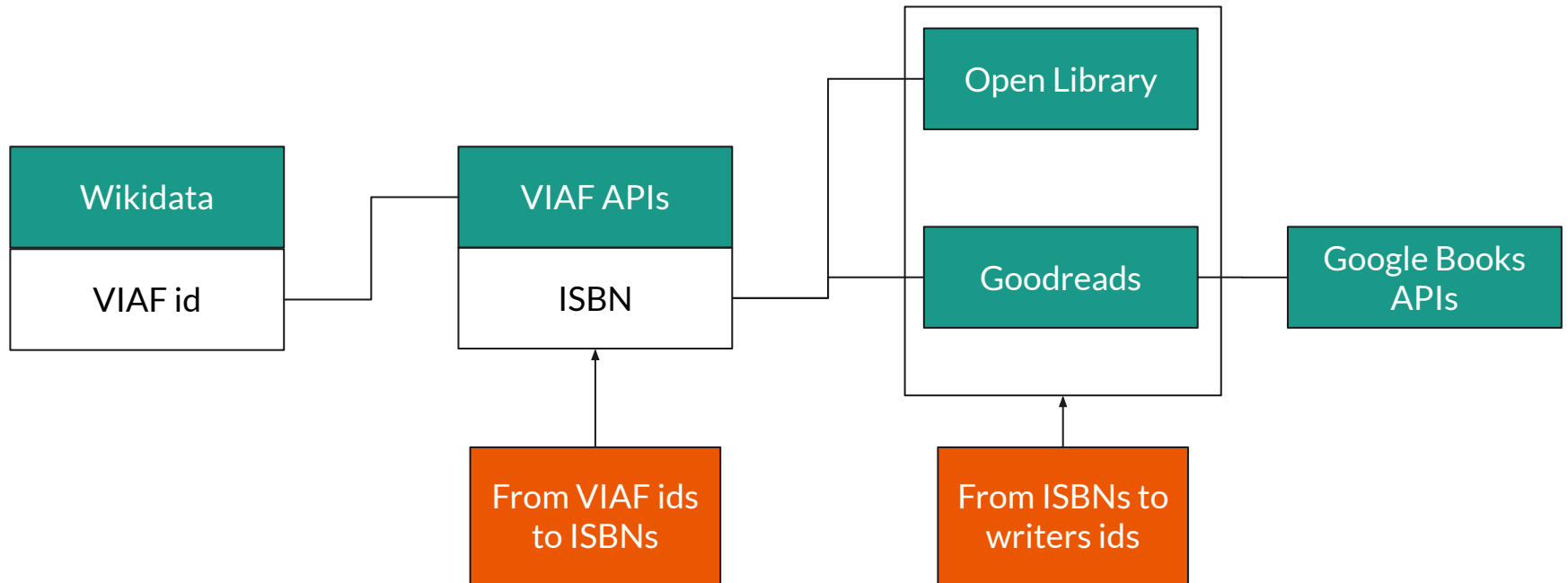
Data gathering

1. All people with occupation 'writer' (Q36180), 'poet' (Q49757), or 'novelist' (Q6625963)
2. Their country of birth, inferred from 'place of birth' (P19)
3. Their year of birth, inferred from 'date of birth' (P569)
4. Their works (P50)

Evaluating underrepresentation



Mapping with other communities of readers





Mapping results

Source	Writers with at least 1 work (% Transnational)	N. of works
Wikidata	22,515 (8.6%)	117,798
Open Library	24,370 (12.4%)	226,108
Goodreads	60,201 (11.0%)	627,214
Total	71,443 (10.6%)	971,120

Table 2. Number of works for each platform

Data visualization platform

The screenshot displays a web-based data visualization platform. At the top, a navigation bar includes buttons for 'Clear All', 'Layout', search, 'PNG', 'SVG', and a 'Data Language' dropdown set to 'Italiano'. A 'LOGOUT' button is on the right. A search bar on the left contains 'chinua achebe'. Below it, a list of search results is shown, with 'Chinua Achebe' selected. The main workspace features a semantic network with two nodes: a 'Person' node for 'Chinua Achebe' and an 'Expression' node for 'Things Fall Apart: An Adapted Cla...'. They are connected by a relationship labeled 'wasAttributedTo'. The right sidebar, titled 'INFO', provides work details for the selected expression, including publication year (2000), author (Chinua Achebe), language (English), number of ratings (46), rating (3.98), subject (igbo (african people)), and type (Expression).

Clear All Layout Search PNG SVG Data Language Italiano LOGOUT

chinua achebe

- Chinua Achebe
- Conversations with Chinua Achebe
- Chinua Achebe Reading Anthills of the Savannah and Arrow of God

Person Chinua Achebe ← wasAttributedTo → Expression Things Fall Apart: An Adapted Cla...

SKATEBOARD

INFO

Work details "Things Fall Apart: An Adapted Classic".

YEAR OF PUBLICATION 2000

AUTHOR Chinua Achebe

LANGUAGE English

NUMBER OF RATINGS 46

RATING 3.98

SUBJECT igbo (african people)

HTTP://WWW.W3.ORG/2002/07/OWL#THING Goodreads Reception

TYPE Expression

3. Extracting biographical events



Objective

Augmenting the amount of structured knowledge about Transnational writers from raw text biographies



The biographical event extraction pipeline

1. Coreference resolution of the entity-target of the biography
2. Event detection
3. Mapping to Wikibase data scheme
4. Named Entity Recognition of the triple's object
5. Entity Linking



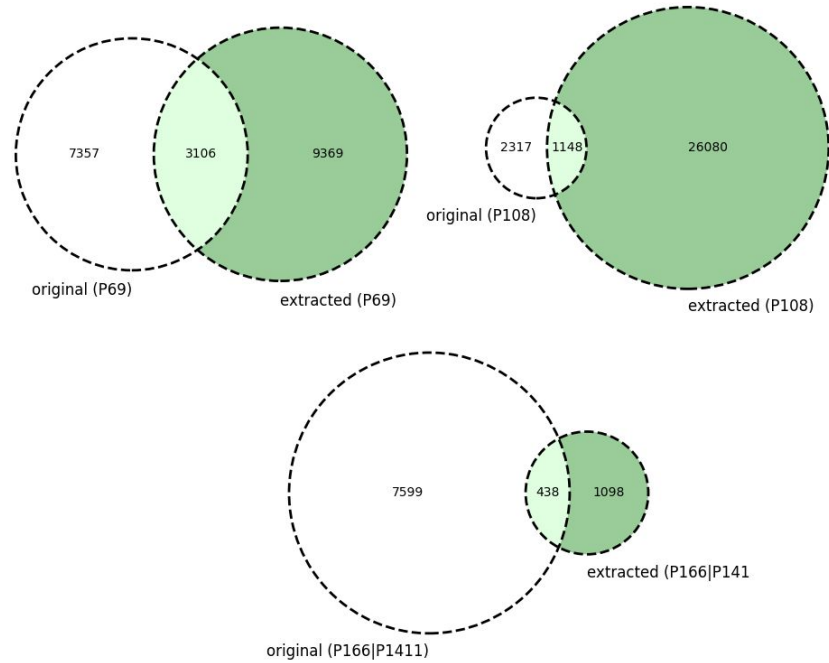
The pipeline, in practice

He served as the **Herald**'s **editor** during the 1951–52 school year.

- **He** → 'Chinua Achebe' (Q155845)
- **editor** → 'employer' (P108)
- **Herald** → The Herald (Q7739400)
- **Q155845** **P108** **Q7739400**



Results





Future Work

1. Revise the semantic model
2. Enrich the KG with knowledge from other sources
3. Improve the visualization platform

This project has received funding from the European Union's Horizon Europe research and innovation programme via the NGI Search grant agreement 101069364 and linked with the Next Generation Internet Initiative.



Bibliography

- Stranisci, M. A., Patti, V., & Damiano, R. (2021). Representing the under-represented: A dataset of post-colonial, and migrant writers. *OPEN ACCESS SERIES IN INFORMATICS*, 93, 1-14
- Stranisci, M. A., Basile, V., Damiano, R., & Patti, V. (2021). Mapping Biographical events to ODPs through Lexico-Semantic Patterns?. In *CEUR WORKSHOP PROCEEDINGS* (Vol. 3011, pp. 1-12). CEUR-WS
- Stranisci, et al (2022). Guidelines and a Corpus for Extracting Biographical Events. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*
- Stranisci, M. A., Patti, V., & Damiano, R. (2023). User-Generated World Literatures: a Comparison between Two Social Networks of Readers
- Stranisci, et al (2023). WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
- Stranisci, M. A., Bernasconi, E., Patti, V., Ferilli, S., Ceriani, M., & Damiano, R. (2023). The World Literature Knowledge Graph. *arXiv preprint arXiv:2307.16659*

Thanks for your attention

mail: marcoantonio.stranisci@unito.it

twitter: [@marcostranisci](https://twitter.com/marcostranisci)

linkedin: [in/marcostranisci](https://www.linkedin.com/in/marcostranisci)