Wikipedia Cultural Diversity Dataset: helping editors to enrich cross-language coverage

David Laniado

Eurecat - Technology Center of Catalonia

Marc Miquel-Ribé

Universitat Pompeu Fabra, Barcelona

Motivation

- Most Wikipedia studies are based on the English version, despite 301 editions exist.
- Lack of content correspondence between languages, due to cultural and contextual factors.
- Multilingualism activities mostly happen with incursions to the English language edition made by a minority of very participative editors.

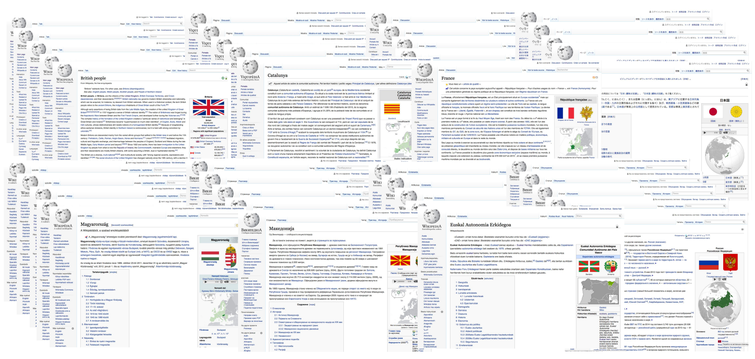
Problem: Wikipedia language editions do not reflect enough the world's cultural diversity.

Solution: Wikipedia Cultural Diversity Observatory (WCDO) http://wcdo.wmflabs.org
"a joint space for **researchers** and **activists** to study and address **knowledge gaps,** and increase cultural diversity in contents."

Dataset to draw a cartography of cultural diversity, and to develop tools to bridge the culture gap

To identify the **Cultural Context Content (CCC)** for each language, i.e. the articles related to the editors' cultural contexts (traditions, language, politics, biographies, places, events, etc.):

1. associate each language to the territories where it is spoken officially or where is native
2. identify articles that relate to each territory

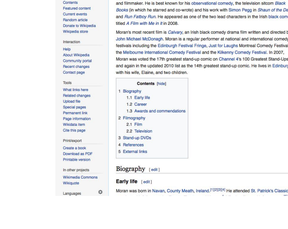


Language-territory mapping

Language-Territories Mapping Database

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	territoryname	territorynameNative	QitemTerritory	languagename	Wiki	demor	demon	ISO3166	ISO31662	region	country	indi	lan	officia	nu
2	Afar	Qafar	Q193494	Afar	aa			ET	ET-AF	yes	Ethiopia	yes	2	regional	0
3	Somali	Somali	Q202800	Afar	aa			ET	ET-SO	yes	Ethiopia	yes	2	regional	0
4	Amhara	Amhara	Q203009	Afar	aa			ET	ET-AM	yes	Ethiopia	yes	2	regional	0
5	Ali Sabieh	Ali Sabieh	Q821008	Afar	aa			DJ	DJ-AS	yes	Djibouti	yes	5	no	0
6	Arta	Arta	Q705941	Afar	aa			DJ	DJ-AR	yes	Djibouti	yes	5	no	0
7	Obock	Obock	Q844929	Afar	aa			DJ	DJ-OB	yes	Djibouti	yes	5	no	0
8	Dikhil	Dikhil	Q283979	Afar	aa			DJ	DJ-DI	yes	Djibouti	yes	5	no	0
9	Debubawi K'eyih	Debubawi K'eyih	Q27728	Afar	aa			ER	ER-DU	yes	Eritrea	yes	5	no	0
10	Semenawi K'eyi B	Semenawi K'eyi Bahri	Q27910	Afar	aa			ER	ER-SK	yes	Eritrea	yes			
11	Abkhazia	Аҧсны	Q23334	Abkhaz	ab	Abkhaz		GE	GE-AB	yes	Georgia	yes	2	regional	1
12	Aceh	Acèh	Q1823	Aceh	ace			ID	ID-AC	yes	Indonesia	yes	6	no	0
13	Sumatera Utara	Sumatra Barôh	Q2140	Aceh	ace			ID	ID-SU	yes	Indonesia	yes	6	no	0
14	Republic of Adyge	Адыгэй	Q3734	Adyghe	ady			RU	RU-AD	yes	Russian Federation	yes	2	regional	1
15	Krasnodar Krai	Краснодар край	Q3680	Adyghe	ady			RU	RU-KDA	yes	Russian Federation	yes	2	regional	1
16	Karachay-Cherkes	Къэрэщае-Черкес	Q5328	Adyghe	ady			RU	RU-KC	yes	Russian Federation	yes	2	regional	1
17	South Africa	Suid-Afrika	Q258	Afrikaans	af	South Afri	Suid-Afrika	ZA		no	South Africa	yes	1	national	1
18	Central	Sentraal distrik	Q57525	Afrikaans	af			BW	BW-CE	yes	Botswana	yes	5	no	1
19	Ghanzi	Ghanzi	Q57571	Afrikaans	af			BW	BW-GH	yes	Botswana	yes	5	no	1
20	Kgalagadi	Kgalagadi	Q57581	Afrikaans	af			BW	BW-KG	yes	Botswana	yes	5	no	1
21	Kgatleng	Kgatleng	Q57593	Afrikaans	af			BW	BW-KL	yes	Botswana	yes	5	no	1
22	Southern	Suid distrik	Q57609	Afrikaans	af			BW	BW-SO	yes	Botswana	yes	5	no	1
23	Botswana	Botswana	Q963	Afrikaans	af	Motswana;Botswan		BW		no	Botswana	yes	5	no	1
24	Ghana	Ghana	Q117	Akan	ak	Ghanaian		GH		no	Ghana	yes	3	no	1
25	Switzerland	Schweiz	Q39	German, Swiss	als	Swiss		CH		no	Switzerland	yes	5	no	0
26	Vorarlberg	Vorarlberg	Q38981	German, Swiss	als			AT	AT-8	yes	Austria	yes	5	no	0
27	Champagne-Arde	Champagne-Ardenne	Q14103	German, Swiss	als			FR	FR-G	yes	France	yes	6	no	0
28	Lorraine	Lothringen	Q1137	German, Swiss	als			FR	FR-M	yes	France	yes	6	no	0
29	Alsace	Elsass	Q1142	German, Swiss	als			FR	FR-A	yes	France	yes	6	no	0
30	Baden-Württemb	Baden-Württemberg	Q985	German, Swiss	als			DE	DE-BW	yes	Germany	yes	5	no	0

For each line (territory): Wikidata Language Qitem, Language name, Language name in Native language, ISO 639 code, associated territories at country level (ISO 3166 code, English name, Native language name, demonym, Qitem) or at first subdivision (ISO 3166-2 code, English name, Native language name, demonym, Qitem) according to the information generated by Ethnologue.



Examples:
Czech CCC only relates to concepts from Czech Republic.
Italian CCC includes articles related to Italy, San Marino, Vaticano, Canton Ticino, Istria among others.

Article features and classification

Different retrieval strategies to extract content from each language edition and label it as Cultural Context Content (CCC), according to article features:

1. **Geolocation coordinates**
2. Specific **keywords in article titles** (language name, territory name, and demonym)
3. Specific keywords in categories containing the article (in an iterative **category graph crawling**)

4. Wikidata Items that relate to groups of properties such as: Language, Location, Country, Part of, In relation with, …

5. Links to other articles Proportion of incoming and outgoing links connecting to CCC articles

and some negative features:

6. Geolocation in other territories
7. Wikidata properties associated to other territories
8. Percentage of Inlinks/Outlinks to geolocated articles in other territories

To obtain the final selection, we use Machine Learning:

Classifier: Random forest classifier with negative sampling (as we did not have a representative set of negative items, the classifier was trained to distinguish positive from random articles)

Training data: Groundtruth of articles having features that strongly and reliably associate them to the language's cultural context (e.g. geolocation, keywords in the article title, strong WikiData properties like "country of birth")

Testing data: All article having at least some weak features associating them to the language's cultural context

MANUAL ASSESSMENT

Language (ISO Code)	Articles	CCC %	FP %	FN %	F1
ca	584,760	17.1%	2	4	0.98
de	2,195,308	33.7%	1	2	0.99
en	5,676,573	44.2%	5	5	0.95
fa	629,125	21.9%	6	1	0.94
gn	715	19.9%	3	6	0.97
ja	1,110,617	51.0%	1	4	0.99
ms	306,055	22.1%	1	0	0.99
ru	1,481,560	32.2%	0	3	0.99
sw	42,422	19.0%	7	5	0.93
zu	1,111	14.2%	1	3	0.99

Manual assessment of the results for 10 diverse language editions

200 articles from each language edition (100 classified as positive and 100 as negative by the algorithm)

Precision: between 93% and 100%
Recall: between 94% and 100%

Dataset

A record for each article from each language: overall, **49,427,733 articles from 300 language editions**
For each record:

- all the features describing the relation with the corresponding language and territories
- additional features, such as length or number of edits
- final classification of the article (whether it belongs to CCC or not)

Record Example: Parmigiano Reggiano from the Italian Wikipedia

Feature	value
qitem	Q155922
page_title	Parmigiano_Reggiano
date_created	20040913
geocoordinates	
iso3166	
iso31662	
ccc_binary	1
main_territory	Q38 (Italy)
num_retrieval_strategies	5
language_weak_wd	
affiliation_wd	
has_part_wd	
num_inlinks_from_CCC	122
num_outlinks_to_CCC	206
percent_inlinks_from_CCC	0.865
percent_outlinks_to_CCC	0.278
other_ccc_country_wd	
other_ccc_location_wd	
num_inlinks_from_geolocated_abroad	3
num_outlinks_to_geolocated_abroad	9

Feature	value
country_wd	P495:Q38 (country of origin: Italy)
location_wd	P1071: Q1263: Q38; P1071: Q16228: Q38 (location of final assembly: Emilia-Romagna: Italy; location of final assembly: Province of Parma)
language_strong_wd	
created_by_wd	
part_of_wd	
keyword_title	
category_crawling_territories	Q38;Q652 (Italy;Italian)
category_crawling_level	1
percent_inlinks_from_geolocated_abroad	0.0213
percent_outlinks_to_geolocated_abroad	0.0122
num_inlinks	141
num_outlinks	739
num_bytes	13815
num_references	16
num_edits	471
num_editors	268
num_discussions	16
num_pageviews	639
num_wdproperty	16
num_interwiki	59
featured_article	

Applications

The possible uses of the dataset are many, we highlight three:

- Wikipedia Culture Gap assessment and improvement
- Academic research in the Digital Humanities field
- User-generated Content based technologies

Conclusion

- Wikimedia Foundation's horizon for 2030 is to "counteract structural inequalities to ensure a just representation of knowledge and people in the Wikimedia movement"
- With this dataset presented we expect to remove some of the main impediments to both recognize and foster cultural diversity in Wikipedia.
- The dataset is available for the 301 Wikipedias, and contains a fine-grained categorization of each article's relationships towards their nearby geographical and cultural entities.

References

Miquel-Ribé, M., & Laniado, D. (2018). Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. *Frontiers in Physics*, 5, 12. Open Access.

Miquel-Ribé, M., & Laniado, D. (2019). Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions. *Proceedings of ICWSM '19*.