Wikipedia Cultural Diversity Observatory (WCDO)

[https://meta.wikimedia.org/wiki/WCDO]

Dr. Marc Miquel

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, **Catalonia**

Amical Wikimedia (Catalan Wikipedia)

July 18th 2018 **Cape Town, South Africa**

”Knowledge equity: As a social movement, we will focus our efforts on the knowledge and communities that have been left out by structures of power and privilege. We will welcome people from every background to build strong and diverse communities. **We will break down the social, political, and technical barriers preventing people from accessing and contributing to free knowledge**."

2030 Strategic direction, Wikimedia Foundation

https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction

I. The Problem

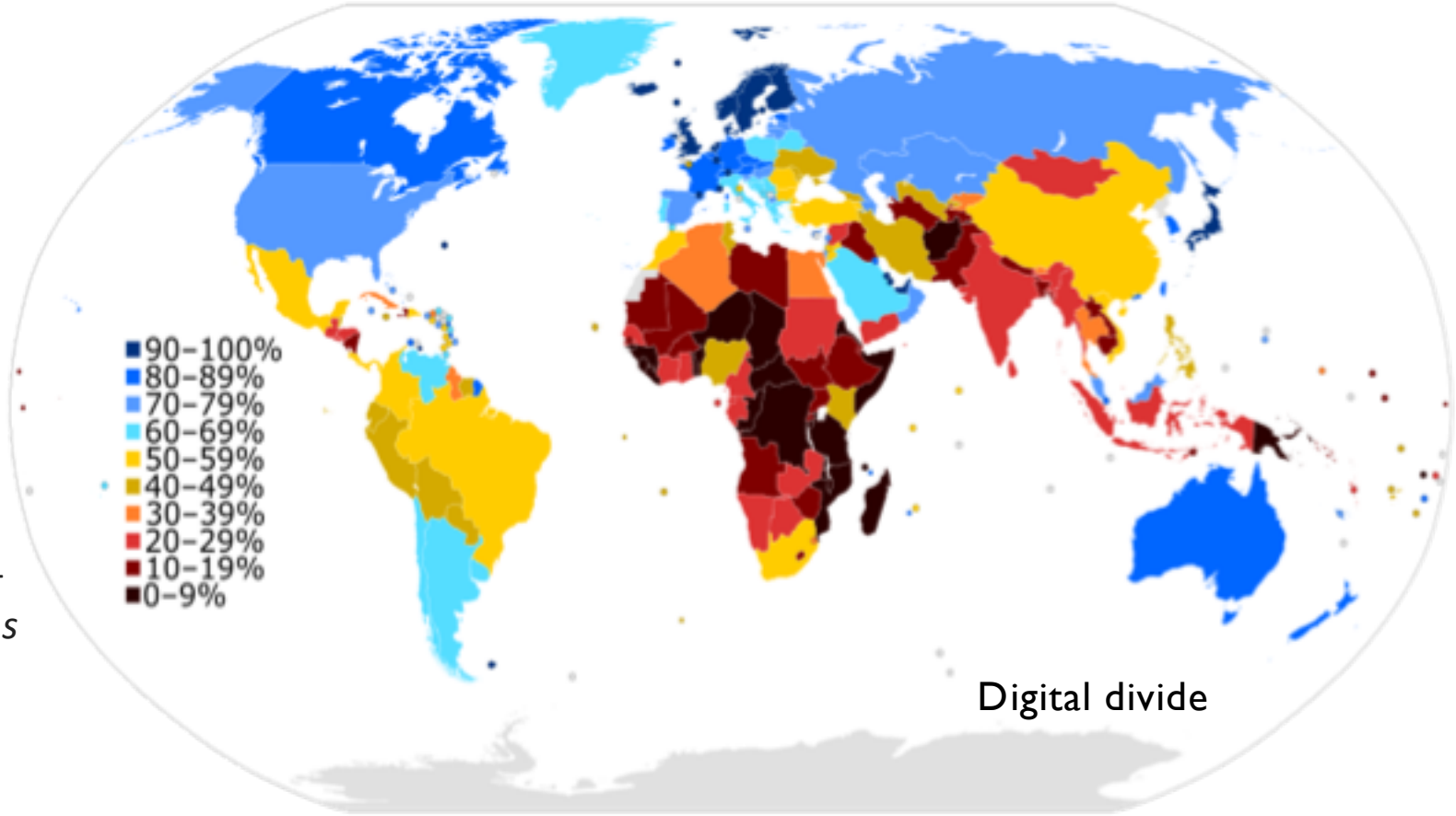
Wikipedia project does not reflect well enough the world's cultural diversity.

Some voices are missing or underrepresented →

- **First, because that many articles that should describe the world's cultural diversity do not exist because not everyone has a Wikipedia, or cannot contribute to it.**

We know that this is due to many factors such as the digital divide, language reputation, among others.

Van Dijk, Z. (2009). Wikipedia and lesser-resourced languages. *Language Problems and Language Planning*, *33*(3), 234-250.



Digital divide

- **Second, because there exists some *language gaps*:**

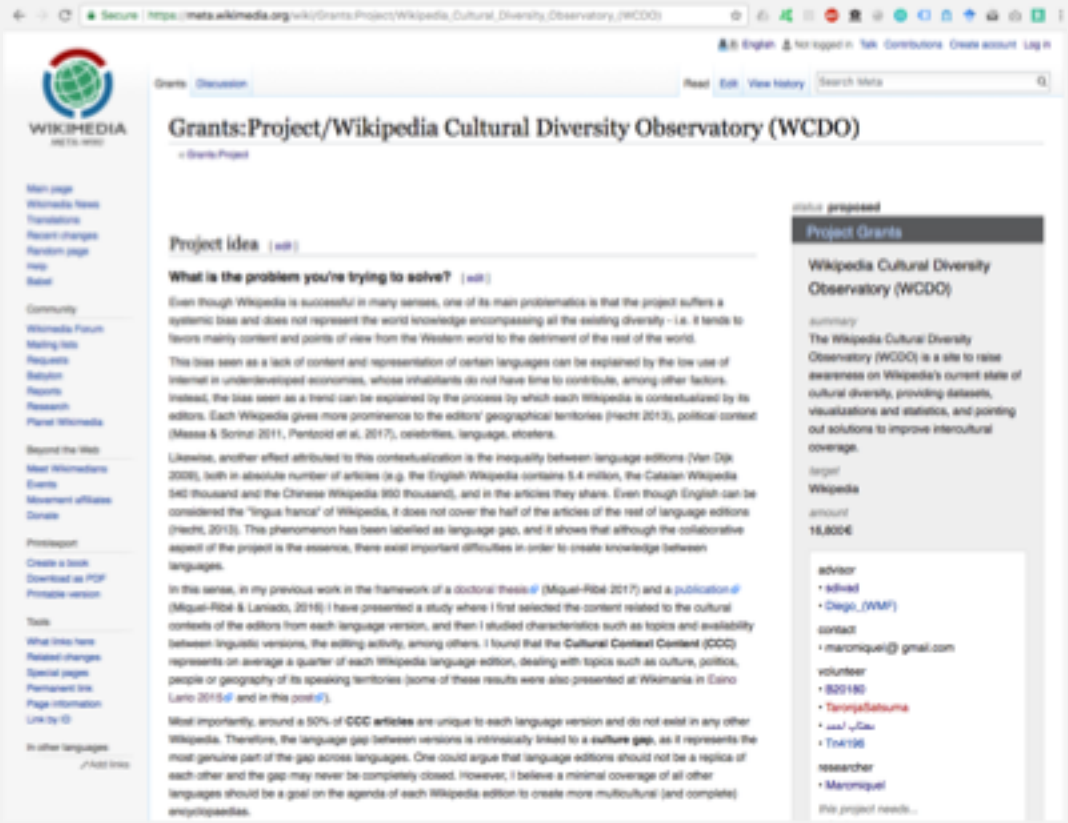
Wikipedias do not cover each others' content.

But this is something we can work on it. This is the scope of this project.

2. Proposed Solution

Wikipedia Cultural Diversity Observatory (WCDO).

Project aimed at **raising awareness** on the current state of cultural diversity in each language and, at the same time, **providing tools** to improve interlanguage collaboration for intercultural coverage.



https://meta.wikimedia.org/wiki/Grants:Project/Wikipedia_Cultural_Diversity_Observatory_(WCDO)

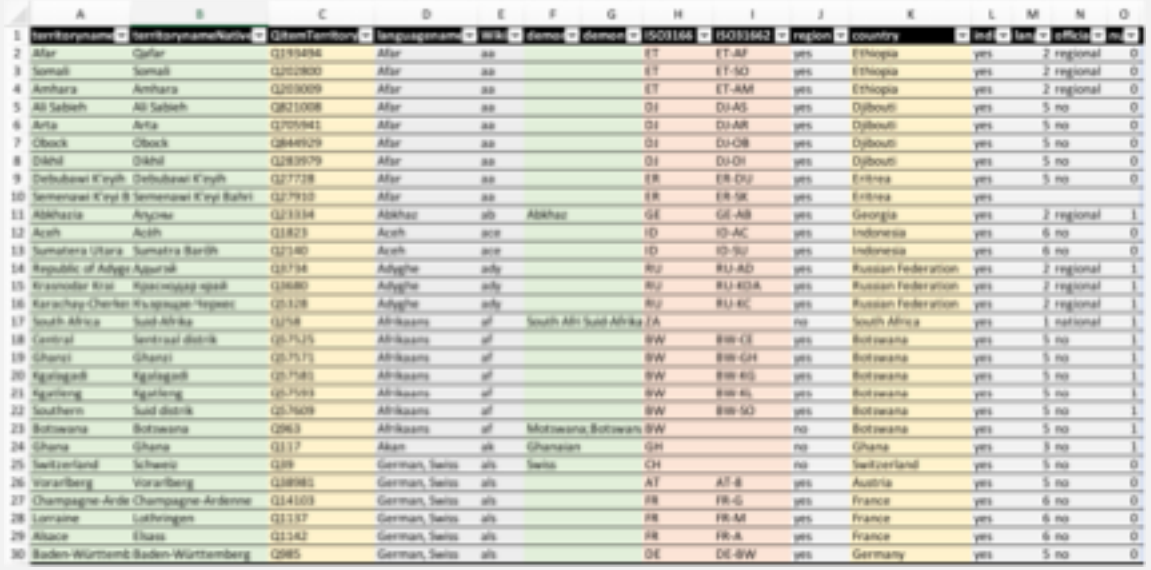
In this research:

We select the **Cultural Context Content (CCC)**, i.e. the articles related to the editors' cultural contexts in each language edition (traditions, language, politics, agriculture, biographies, places, events, etcetera).

This means associating each language to the territories where it is spoken officially or where is native, and then, collecting articles that relate to each territory.

3. Methodology

This requires (i) creating a database with Language-Territories Mapping and (ii) employing different retrieval strategies to extract content from each language edition and label it as CCC.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	territoryname	territorynameNative	QitemTerritory	languagename	WikL	demo	demoo	ISO3166	ISO3166-2	region	country	ind	lan	officia	re
2	Afar	Qafar	Q1959494	Afar	aa			ET	ET-AF	yes	Ethiopia	yes	2 regional		0
3	Somali	Somali	Q1012800	Afar	aa			ET	ET-SO	yes	Ethiopia	yes	2 regional		0
4	Amhara	Amhara	Q1033009	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	Q27718	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	Aceh	Q1823	Aceh	ace			ID	ID-AC	yes	Indonesia	yes	6 no		0
13	Sumatera Utara	Sumatra Barât	Q2140	Aceh	ace			ID	ID-SU	yes	Indonesia	yes	6 no		0
14	Republic of Adyge	Адыгэ	Q6734	Adyghe	ady			RU	RU-AD	yes	Russian Federation	yes	2 regional		1
15	Krasnodar Krai	Краснодар край	Q36680	Adyghe	ady			RU	RU-KDA	yes	Russian Federation	yes	2 regional		1
16	Karachay-Cherkes	Къарачай-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	Q5752	Afrikaans	af			BW	BW-CE	yes	Botswana	yes	5 no		1
19	Ghanzi	Ghanzi	Q5771	Afrikaans	af			BW	BW-GH	yes	Botswana	yes	5 no		1
20	Kgalagadi	Kgalagadi	Q5748	Afrikaans	af			BW	BW-KG	yes	Botswana	yes	5 no		1
21	Kgatleng	Kgatleng	Q5759	Afrikaans	af			BW	BW-KL	yes	Botswana	yes	5 no		1
22	Southern	Suid distrik	Q5769	Afrikaans	af			BW	BW-SO	yes	Botswana	yes	5 no		1
23	Botswana	Botswana	Q963	Afrikaans	af	Motswana	Batswana	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

Language Territories mapping spreadsheet with 1783 rows.

(i) Wikidata Language Qitem, Language name, Language name in Native language, the ISO 639 code, the 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.

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/Language_Territories_Mapping

For example:

Italy	Italia	Q38	Italian	it	Italian	italiano;ita	IT		no	Italy
Istria	Istriana	Q58268	Italian	it			HR	HR-18	yes	Croatia
San Marino	San Marino	Q238	Italian	it	Sammarinese	sammarin	SM		no	San Marino
Piran	Pirano	Q1382	Italian	it			SI	SI-090	yes	Slovenia
Izola	Isola	Q15877	Italian	it			SI	SI-040	yes	Slovenia
Graubünden	grigioni	Q11925	Italian	it			CH	CH-GR	yes	Switzerland
Ticino	ticino	Q12724	Italian	it			CH	CH-TI	yes	Switzerland
Vatican City	vaticano	Q237	Italian	it			VA		no	Vatican State

Sweden	Sverige	Q34	Swedish	sv	Swedish	svensk;sve	SE		no	Sweden
Åland s	Åland	Q5689	Swedish	sv	Ålandic	Ålänning	FI	FI-01	yes	Finland
Kymenlaakso	Kymmenedalen	Q5698	Swedish	sv			FI	FI-09	yes	Finland
Ostrobothnia	Österbotten	Q5702	Swedish	sv			FI	FI-12	yes	Finland

Territories where the language is spoken as **native or with official status**

(i) Wikidata Language Qitem, Language name, Language name in Native language, the ISO 639 code, the 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.

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/Language_Territories_Mapping

(ii) The different retrieval strategies to extract content from each language edition and label it as CCC are the following.

Wikipedia articles with characteristics such as:

1. **Geolocation coordinates**
2. **Specific keywords on their titles (language name, territory name, and demonym).**
3. **Contained in categories with keywords on their titles or in categories contained by these (in an iterative category graph crawling).**

Wikipedia MySQL db Replicas

Wikidata Items that relate to groups of properties such as:

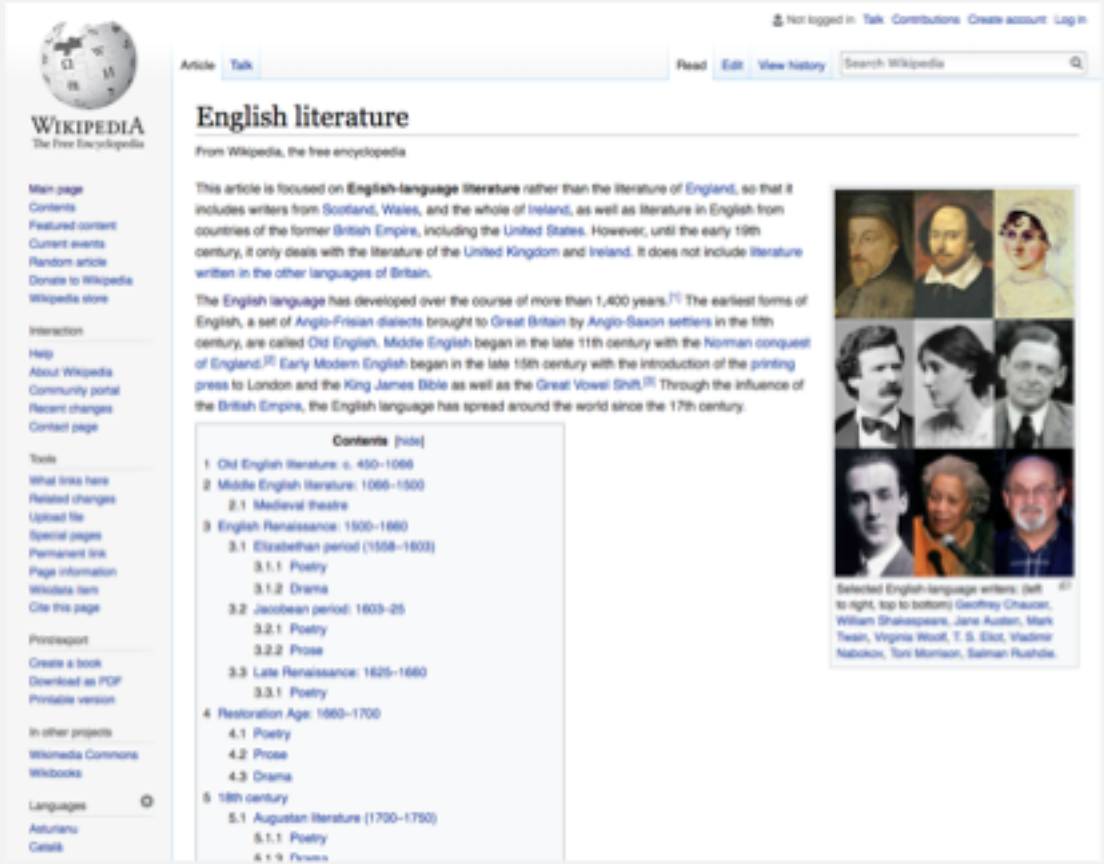
- **Language**
- **Location**
- **Country**
- **Part of**
- **In relation with**
- **…**

Wikidata JSON dump

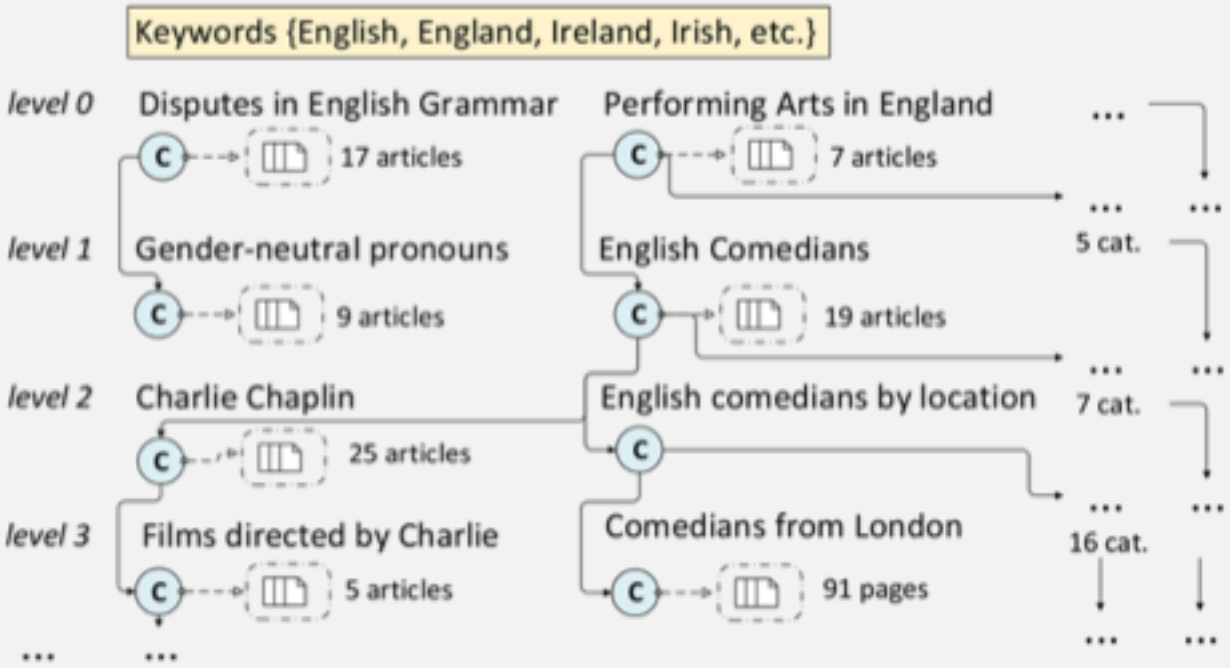
We create a database as rich as possible.

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/Cultural_Context_Content

Some of these strategies are strong, while some other are weak.



- **Keyword (demonym/territory name) on title is strong**
- **Geolocation in one of the territories is strong**



Category crawling using keywords

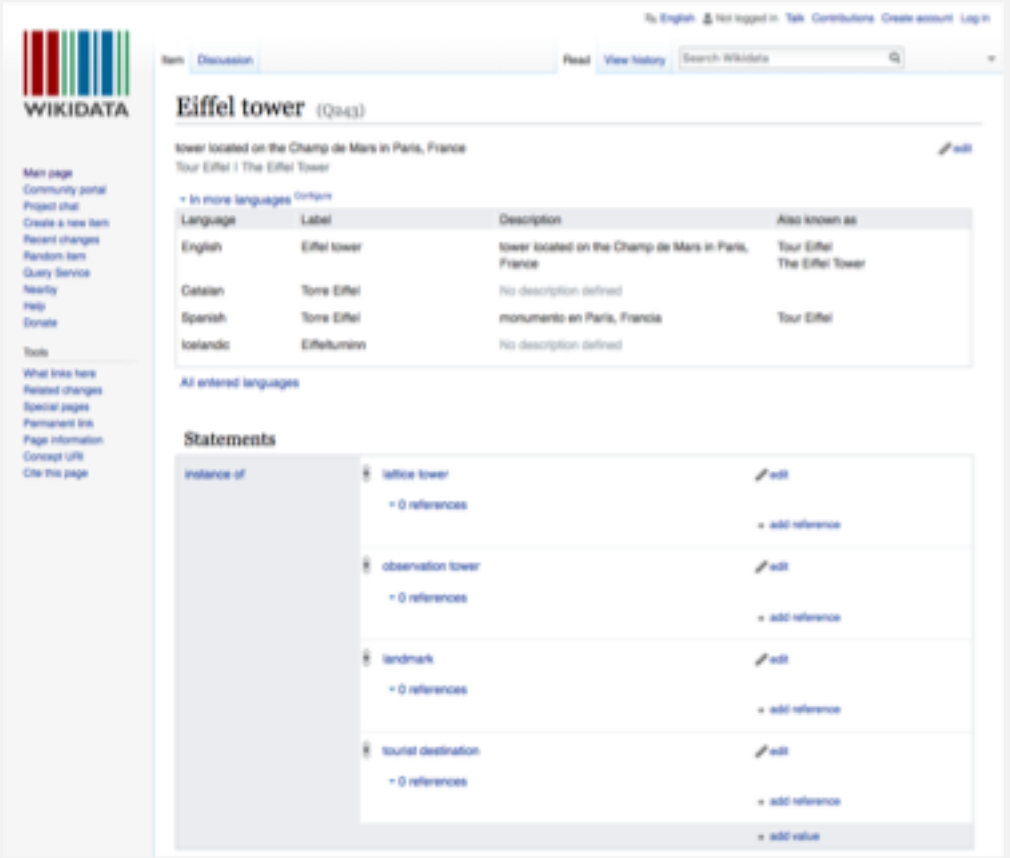
- **Being in a subcategory of a category containing a keyword on its title is weak**

Some Wikidata properties are strong

- **Location properties (location, located in administrative,…).**
- **Country properties (country of citizenship, of origin).**
- **Language properties (official language, native language…).**

Some Wikidata properties are weak

- **Affiliation properties (member of, educated at, employer,…).**
- **Has part (contains administrative entity, has part).**
- **Language properties (language of work, language used,…).**



Those labelled as weak are because we cannot be sure how representative the feature is to be included as Cultural Context Content.

MACHINE LEARNING CLASSIFIER

We have a rich database with all the articles of all the Wikipedias with these features.

Those tagged with a strong feature are considered the Cultural Context Content groundtruth. We are sure they are CCC.

For every Wikipedia article we compute the number of incoming and outgoing links to the CCC groundtruth, as well as the percent they represent from the total number of incoming and outgoing links.

RANDOM FOREST Classifier (implemented using scikit-learn).

- **Training Data:** The Cultural Context Content groundtruth as a positive training set. while the rest of articles (some tagged with other features such as category crawling, wikidata properties and some untagged) are sampled 10x and introduced as negative training set. This is called Negative Sampling.
- **Testing Data:** We take those which have at least one CCC feature (weak ones: category crawling and some wikidata properties) and test them against the classifier in order to obtain the good ones.

The positive articles from the classifier and the initial CCC groundtruth constitute the final CCC.
We run a manual assessment (blind) to determine the quality of the selection and the results were in average a 5% false positive and 5% false negative.

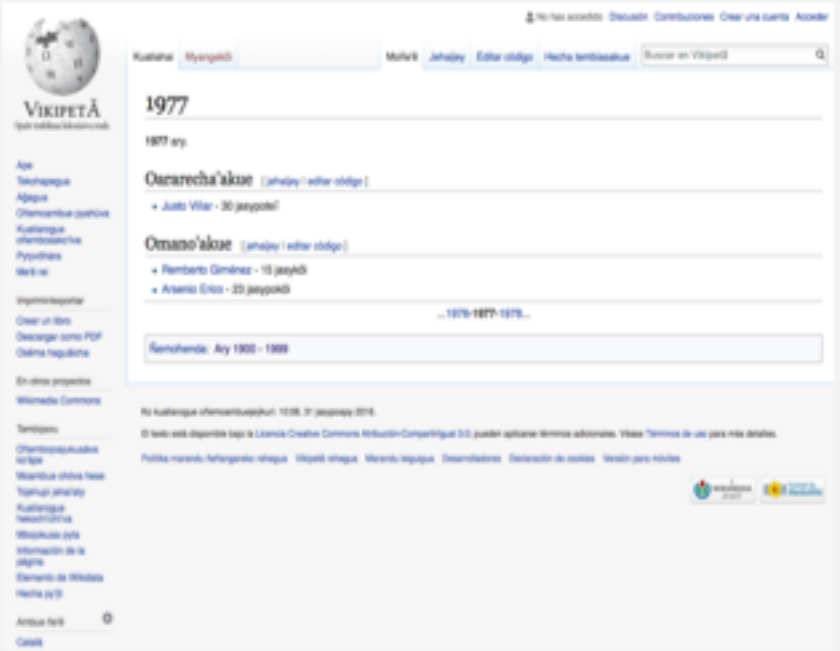
CCC IS A CONTINUUM

Should Chomsky be in Catalan CCC?
He received a Catalan Gov. prize but…

What about Leo Messi?

Should Alsace be part of the German CCC?

It used to be part of the German Empire



Year 1977 should not be part of Guarani CCC,
Even the article in this language contains only events related to Guarani CCC…

The classifier 'decides' whether it should be in or not according to the features.
Not enough outlinks to CCC or no category from the category crawling? Probably out.

Project's Technical Overview

- **Wikimedia Cloud Server at Toolforge**

Server: https://tools.wmflabs.org/admin/tool/wcdo
Phabricator: https://phabricator.wikimedia.org/T193984

Execution:
crontab (cron job in shell) to execute the scripts on a monthly basis.

Python scripts:
ccc_selection.py (it creates the main database ccc_current.db and the datasets).
wcdo_creation.py (it creates the database wcdo_data.db and updates stats in meta with ***pywikibot***).

- **Datasets**

They are available at wcdo.wmflabs.org and at figshare.com/account/home#/projects/28272

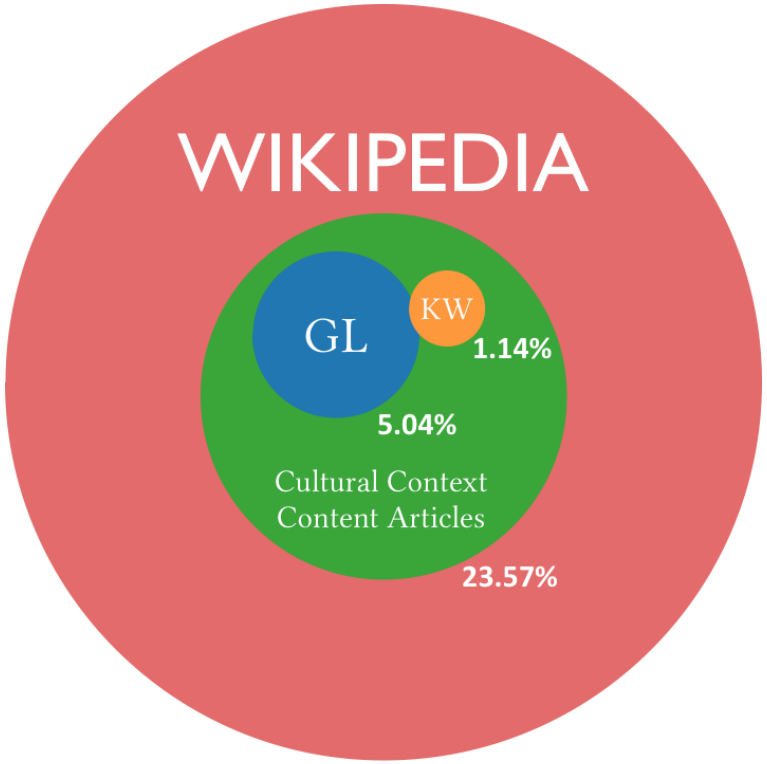
- **Code in Github**

All the code, presentation and files are at: github.com/marcmiquel/WCDO

Do you want to join? E-mail me at marcmiquel@gmail.com

4. WCDO Results (Top language editions, October, 2016):

We used the first three of the mentioned strategies with 40 Wikipedia language editions (the first 30 in number of articles and 10 to increase diversity).



CCC articles were about a quarter of each Wikipedia language edition.

4. WCDO Results

4.1 CCC Extent (Top Languages, July, 2018):

Language	Wiki	Articles	CCC Extent (%)	CCC Geolocated (%)	Keywords Title (%)	CCC People (%)	CCC Female-Male %
English	en	5,623,263	2,764,131 (49.16%)	561,901 (9.99%)	150,811 (2.68%)	982,094 (17.46%)	17.8% - 82.2%
German	de	2,177,877	742,689 (34.1%)	221,567 (10.17%)	21,103 (0.97%)	344,970 (15.84%)	14.4% - 85.6%
French	fr	1,979,006	612,187 (30.93%)	140,582 (7.1%)	37,000 (1.87%)	228,814 (11.56%)	14.8% - 85.2%
Japanese	ja	1,101,749	580,727 (52.71%)	73,902 (6.71%)	10,835 (0.98%)	158,935 (14.43%)	25.8% - 74.2%
Russian	ru	1,465,638	504,961 (34.45%)	216,142 (14.75%)	4,934 (0.34%)	180,894 (12.34%)	11.9% - 88.1%
Spanish	es	1,363,891	434,687 (31.87%)	85,006 (6.23%)	28,131 (2.06%)	149,711 (10.98%)	16.5% - 83.5%
Swedish	sv	3,780,430	326,657 (8.64%)	111,121 (2.94%)	16,114 (0.43%)	82,454 (2.18%)	20.8% - 79.2%
Polish	pl	1,275,946	300,292 (23.53%)	121,268 (9.5%)	12,819 (1.0%)	93,693 (7.34%)	14.7% - 85.3%
Italian	it	1,436,670	284,213 (19.78%)	54,003 (3.76%)	11,543 (0.8%)	109,102 (7.59%)	11.9% - 88.1%
Arabic	ar	573,631	205,186 (35.77%)	23,210 (4.05%)	17,645 (3.08%)	44,554 (7.77%)	14.6% - 85.4%
Portuguese	pt	994,69	197,363 (19.84%)	24,496 (2.46%)	11,124 (1.12%)	59,720 (6.0%)	16.4% - 83.6%
Dutch	nl	1,932,561	185,939 (9.62%)	43,959 (2.27%)	7,803 (0.4%)	70,585 (3.65%)	16.4% - 83.6%
Ukrainian	uk	781,042	178,487 (22.85%)	52,019 (6.66%)	2,770 (0.35%)	60,114 (7.7%)	12.4% - 87.6%

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/List_of_Wikipedias_by_Cultural_Context_Content

Language	Wiki	Articles	CCC Extent (%)	CCC Geolocated (%)	Keywords Title (%)	CCC People (%)	CCC Female-Male %
Standard Chinese	zh	1,002,864	162,889 (16.24%)	58,365 (5.82%)	5,642 (0.56%)	28,199 (2.81%)	27.2% - 72.8%
Persian	fa	623,57	137,805 (22.1%)	51,384 (8.24%)	3,747 (0.6%)	19,890 (3.19%)	11.6% - 88.4%
Bokmål	no	485,173	112,338 (23.15%)	27,313 (5.63%)	1,895 (0.39%)	38,601 (7.96%)	21.5% - 78.5%
Indonesian	id	427,821	110,834 (25.91%)	17,149 (4.01%)	2,545 (0.59%)	10,167 (2.38%)	26.6% - 73.4%
Czech	cs	406,304	102,237 (25.16%)	38,339 (9.44%)	2,845 (0.7%)	38,074 (9.37%)	12.9% - 87.1%
Catalan	ca	580,536	102,122 (17.59%)	50,996 (8.78%)	4,087 (0.7%)	31,384 (5.41%)	14.9% - 85.1%
Finnish	fi	435,444	101,236 (23.25%)	16,471 (3.78%)	929 (0.21%)	43,158 (9.91%)	19.5% - 80.5%
Hungarian	hu	428,546	93,981 (21.93%)	18,149 (4.24%)	6,667 (1.56%)	43,248 (10.09%)	13.5% - 86.5%
Turkish	tr	307,699	90,085 (29.28%)	13,054 (4.24%)	5,931 (1.93%)	22,128 (7.19%)	14.0% - 86.0%
Korean	ko	408,397	86,200 (21.11%)	0 (0.0%)	4,587 (1.12%)	20,003 (4.9%)	27.0% - 73.0%
Danish	da	237,878	80,300 (33.76%)	17,955 (7.55%)	2,490 (1.05%)	28,275 (11.89%)	18.1% - 81.9%
Romanian	ro	384,762	76,729 (19.94%)	27,089 (7.04%)	3,892 (1.01%)	20,680 (5.37%)	15.8% - 84.2%
Hindi	hi	125,701	71,907 (57.2%)	14,383 (11.44%)	3,906 (3.11%)	8,817 (7.01%)	24.1% - 75.9%

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/List_of_Wikipedias_by_Cultural_Context_Content

4.2 Culture Gap: most CCC articles not available across languages

About a 60% of the content language gaps are due to CCC.

Big languages like English or geographically close languages are the ones covering best the smaller languages.

What is the weight of each language cultures in other languages? (Cultural Spread)

Language	Target nº1	Target nº2	Target nº3	Target nº4	Target nº5	Relative Spread Idx	Total Spread Idx	Total Spread Art.
English	fr (23.38%)	de (17.93%)	it (24.86%)	es (25.17%)	pl (20.06%)	23.73	16.08	6,659,099
French	en (4.84%)	de (7.42%)	it (10.09%)	es (10.03%)	nl (6.86%)	9.74	7.02	3,164,724
German	en (3.21%)	fr (5.73%)	it (5.97%)	pl (6.18%)	ru (5.29%)	5.74	3.97	1,782,079
Spanish	en (3.01%)	fr (4.69%)	ca (15.66%)	nl (4.16%)	it (5.29%)	8.45	3.88	1,771,146
Russian	uk (16.13%)	en (1.57%)	ce (46.48%)	hy (25.89%)	pl (4.69%)	5.77	2.93	1,334,444
Italian	en (1.65%)	fr (3.54%)	de (2.61%)	es (3.34%)	ru (2.71%)	4.33	2.43	1,107,508
Basic English	en (0.85%)	fr (1.84%)	de (1.5%)	it (2.26%)	es (2.33%)	8.65	2.34	1,096,284
Japanese	en (1.79%)	zh (6.75%)	fr (2.51%)	ko (11.56%)	it (2.47%)	2.89	1.76	810,066
Arabic	en (1.13%)	fr (1.89%)	fa (4.5%)	it (1.7%)	de (1.09%)	4.19	1.54	716,752
Swedish	ceb (1.57%)	en (1.0%)	war (3.75%)	vi (3.32%)	de (1.38%)	2.93	1.6	691,43
Portuguese	en (1.01%)	es (2.56%)	fr (1.68%)	nl (1.56%)	de (1.24%)	2.62	1.38	635,336
Polish	en (1.62%)	fr (2.16%)	de (1.47%)	ru (1.97%)	uk (2.7%)	1.95	1.11	507,401
Hungarian	en (0.53%)	eo (9.82%)	fr (1.18%)	de (0.96%)	it (1.35%)	1.87	1.02	477,292

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/Culture_Gap_(spread)

How well do language editions cover other languages' cultures? (Cultural Coverage)

Language	Articles	Target nº1	Target nº2	Target nº3	Target nº4	Target nº5	Relative Coverage	Total Coverage	Coverage Art.
English	5,623,263	fr (44.47%)	de (24.31%)	es (38.92%)	ja (17.38%)	it (32.67%)	56.58	29.54	2,225,836
French	1,979,006	en (16.74%)	de (15.28%)	es (21.36%)	it (24.64%)	ja (8.57%)	38.88	15.38	1,490,074
German	2,177,877	en (14.13%)	fr (26.41%)	es (13.92%)	it (19.97%)	ru (9.43%)	35.62	13.89	1,327,091
Italian	1,436,670	en (12.92%)	fr (23.69%)	de (11.55%)	es (17.49%)	ja (6.12%)	34.25	12.09	1,210,466
Russian	1,465,638	en (8.96%)	fr (15.66%)	uk (45.88%)	de (10.44%)	es (12.67%)	35.18	11.71	1,146,697
Spanish	1,363,891	en (12.42%)	fr (22.35%)	de (7.9%)	it (16.03%)	pt (17.72%)	30.36	11.07	1,092,391
Dutch	1,932,561	en (8.46%)	fr (21.65%)	es (18.51%)	de (10.04%)	id (33.36%)	32.14	10.63	1,074,710
Polish	1,275,946	en (9.26%)	fr (17.98%)	de (10.61%)	ru (11.84%)	es (11.78%)	31.0	10.71	1,070,839
Swedish	3,780,430	en (9.01%)	fr (18.36%)	es (15.04%)	de (7.31%)	simple (54.68%)	28.23	9.59	956,848
Portuguese	994,69	en (9.13%)	fr (16.02%)	es (16.36%)	de (6.87%)	it (11.87%)	25.77	8.65	873,797
Standard Chinese	1,002,864	en (5.75%)	ja (11.65%)	fr (10.88%)	es (10.51%)	ru (9.03%)	29.07	7.66	776,609
Ukrainian	781,042	ru (24.95%)	en (4.01%)	fr (12.12%)	de (5.14%)	es (6.54%)	24.66	7.59	768,373
Cebuano	4,692,347	en (5.04%)	sv (22.55%)	fr (11.88%)	es (11.11%)	de (3.21%)	23.47	6.4	656,968

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/Culture_Gap_(coverage)

4.3 CCC Vital articles (Lists of Articles from Cultural Context Content)

- Top 100 in number of editors
- Top 1000 in number of editors
- Top 100 most viewed during the last month
- Top 100 most discussed (edits in Talk pages)
- Top 100 geographical with most incoming links
- Top 100 keywords (demonym and territory names) in their titles with most Bytes
- Top 100 featured articles
- Top 100 articles most edited from those created during the first 3 years
- Top 100 articles most edited from those created during the past 3 months

“Top 100 in number of editors” is probably the most important list.
The number of editors is a good indicator of the article relevance.

CCC Vital articles (Top 100 most edited women in Catalan Wikipedia)

- Mercè Rodoreda i Gurguí
- Joaquima de Vedruna
- Caterina Albert i Paradís
- Rita Barberà Nolla
- Ángeles Santos Torroella
- Carme Forcadell i Lluís
- Joana Raspall i Juanola
- Concepció Badia i Millàs
- Pilar Rahola i Martínez
- Laia Sanz i Pla-Giribert
- Montserrat Caballé i Folch
- Alicia Sánchez-Camacho Pérez
- Maria del Mar Bonet
- Margarida Xirgu i Subirà
- Ada Colau i Ballano
- Maria Mercè Marçal i Serra
- Muriel Casals i Couturier
- Concha García Campoy
- Teresa Forcades i Vila
- Victòria dels Àngels
- Arantxa Sánchez Vicario
- Montserrat Roig i Fransitorra
- Carme Karr i Alfonsetti
- Emma Vilarasau Tomàs
- Mònica Terribas i Sala
- Isabel-Clara Simó i Monllor
- Eulàlia de Barcelona
- Carme Chacón Piqueras
- Isabel Coixet i Castillo
- Irene Rigau i Oliver
- Margarida de Prades
- Gemma Lienas i Massot
- Neus Munté i Fernández
- Maria Antònia Munar i Riutort
- Maria Gay
- Isabel de Villena
- Empar Moliner i Ballesteros
- Carme Riera Guilera
- Núria de Gispert i Català
- Núria Perpinyà i Filella
- Maria Lluïsa Borràs i González
- Anna Gabriel i Sabaté
- Joana Ortega i Alemany
- Maria Àngels Anglada i d'Abadal
- Neus Català i Pallejà
- Montserrat Tura i Camafreita
- Amàlia Garrigós i Hernández
- Núria Picas i Albets
- Meritxell Borràs i Solé
- Olga Xirinacs Díaz
- Anna Lizaran i Merlos
- Eva Piquer i Vinent
- Ana María Matute Ausejo
- Montserrat Abelló i Soler
- Alícia de Larrocha i de la Calle
- Maria Antònia Oliver Cabrer
- Marta Rovira i Vergés
- Maria Aurèlia Capmany i Farnés
- Joana Serrat i Tarré
- Teresa Pàmies i Bertran
- Lola Anglada
- Teresa Rebull
- Eulàlia Lledó i Cunill

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/Catalan_Wikipedia/CCC_Vital_articles_women

CCC Vital articles (Top 100 most edited women in English Wikipedia)

- Britney Spears
- Beyoncé
- Mariah Carey
- Christina Aguilera
- Madonna (entertainer)
- Kelly Clarkson
- 7Hillary Clinton
- Diana, Princess of Wales
- Rihanna
- Sarah Palin
- Hilary Duff
- Serena Williams
- Carrie Underwood
- Lady Gaga
- Lindsay Lohan
- Marilyn Monroe
- Jennifer Lopez
- Pink (singer)
- Elizabeth II
- Nicole Scherzinger
- Cher
- Janet Jackson
- Ashley Tisdale
- Ann Coulter
- Paris Hilton
- Margaret Thatcher
- Avril Lavigne
- Whitney Houston
- Ayn Rand
- Mickie James
- Priyanka Chopra
- Taylor Swift
- Raven-Symoné
- Gwen Stefani
- Aaliyah
- Trish Stratus
- Lita (wrestler)
- Katy Perry
- Leona Lewis
- Vanessa Hudgens
- Jessica Simpson
- Ashanti (singer)
- Scarlett Johansson
- Kylie Minogue
- Fergie (singer)
- Jennifer Aniston
- Elizabeth I of England
- Alicia Keys
- Ashlee Simpson
- Celine Dion
- Brenda Song
- Nelly Furtado
- Emma Watson
- Asin
- Kelly Rowland
- Amy Winehouse
- Genie (feral child)
- J. K. Rowling
- Natalie Portman
- Oprah Winfrey
- Ciara
- Demi Lovato
- Kesha

https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/English_Wikipedia/CCC_Vital_articles_women

5. Get Involved (Creating lists of top priority articles, especially)

Wikipedia Cultural Diversity Observatory (WCDO).
Prioritized translations. Automatically generate lists of 100 **Vital articles for every language** so they are the first that every other language should have.

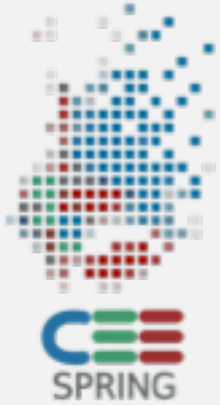
These lists of Top 100 articles would ensure that each Wikipedia language edition has a minimal and strategical coverage of the whole available Wikipedia project cultural diversity.

28,800 articles challenge!

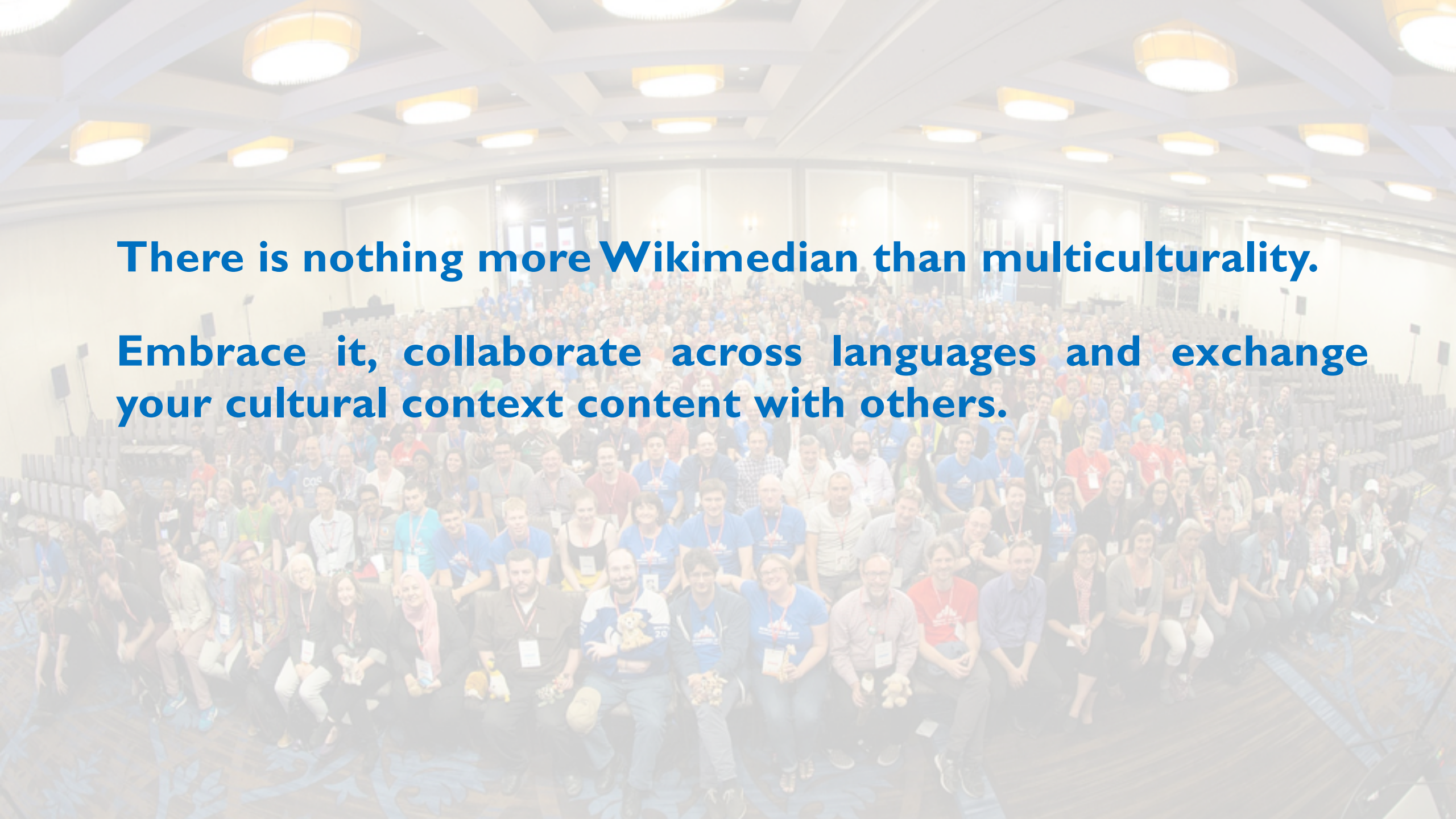
https://meta.wikimedia.org/wiki/Wikipedia_Cultural_Diversity_Observatory/Get_involved

Some existing projects may benefit from WCDO:

- Wikimedia CEE Spring
- Intercultur Wikimedia España
- Catalan Culture Challenge
- WikiArabia
- Systemic bias project (English, Deutsch, Esperanto, Arabic, Dutch and Russian)



Create content across languages, everyone benefits!



There is nothing more Wikimedian than multiculturality.

Embrace it, collaborate across languages and exchange your cultural context content with others.

Wikipedia Cultural Diversity Obsevatory (WCDO)

[https://meta.wikimedia.org/wiki/WCDO]

Dr. Marc Miquel

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, **Catalonia**

Amical Wikimedia (Catalan Wikipedia)

July 18th 2018 **Cape Town, South Africa**

Thank you very much!

Dr. Marc Miquel

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, **Catalonia**

Amical Wikimedia (Catalan Wikipedia)

March 18th 2018 Tunis

References (if you want to know more or engage)

Miquel-Ribé, M., & Laniado, D. (2016, July). Cultural identities in wikipedias. In *Proceedings of the 7th 2016 International Conference on Social Media & Society* (p. 24). ACM.

Miquel-Ribé. M. (2017). *Identity-based motivation in digital engagement: the influence of community and cultural identity on participation in wikipedia* (Doctoral dissertation, Universitat Pompeu Fabra).

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

Greetings to:

WIKIMEDIA FOUNDATION

WIKIMANIA CAPE TOWN JULY 18 - 22, 2018

eurecat Centre Tecnològic de Catalunya

AMICAL WIKIMEDIA