Wikipedia Cultural Diversity Observatory (WCDO)

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

Marc Miquel, PhD

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, **Catalonia**

Amical Wikimedia (Catalan Wikipedia)

July 18-22th 2018 **Cape Town, South Africa**

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.

1. The Problem

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

Some voices are missing or underrepresented

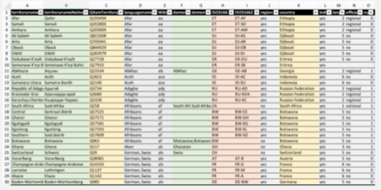
2. Proposed Solution

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.

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.



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

(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 g :h 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

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

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%)	it (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**

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
- VWhitney 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