

Fostering the representation of African cultures in Wikipedia language editions...

Dr. Marc Miquel

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, **Catalonia**

Amical Wikimedia (Catalan Wikipedia)

March 18th 2018 Tunis



...with **Wikipedia Cultural Diversity Observatory** **(WCDO)**

Dr. Marc Miquel

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, **Catalonia**

Amical Wikimedia (Catalan Wikipedia)

March 18th 2018 Tunis



I. The Problem

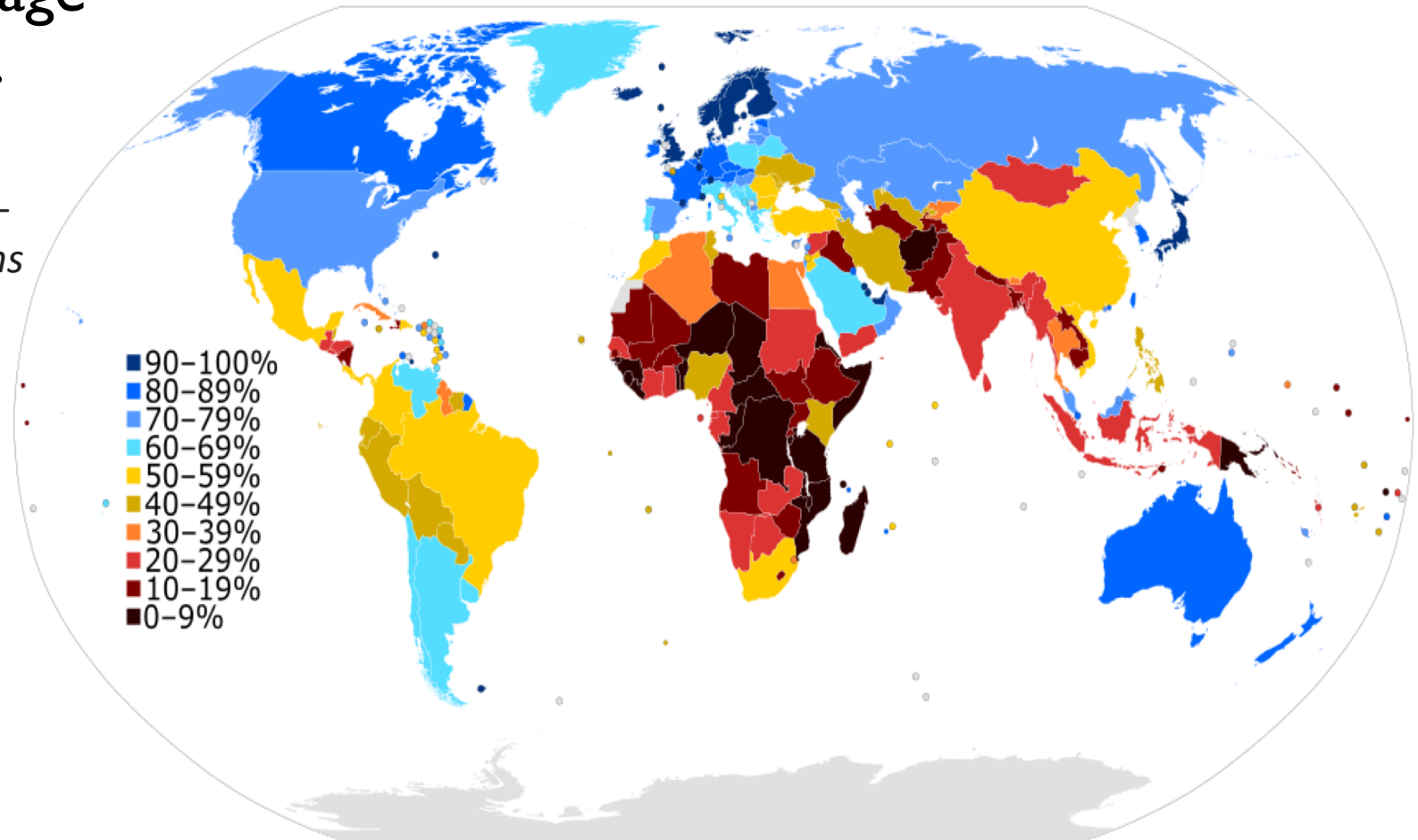
There are only 25 Wikipedia language editions originary from Africa.



They account for the 0.35% of all the Wikipedia articles, or 1.5% if we include Arabic in the list. Many more if we include French...

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.



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

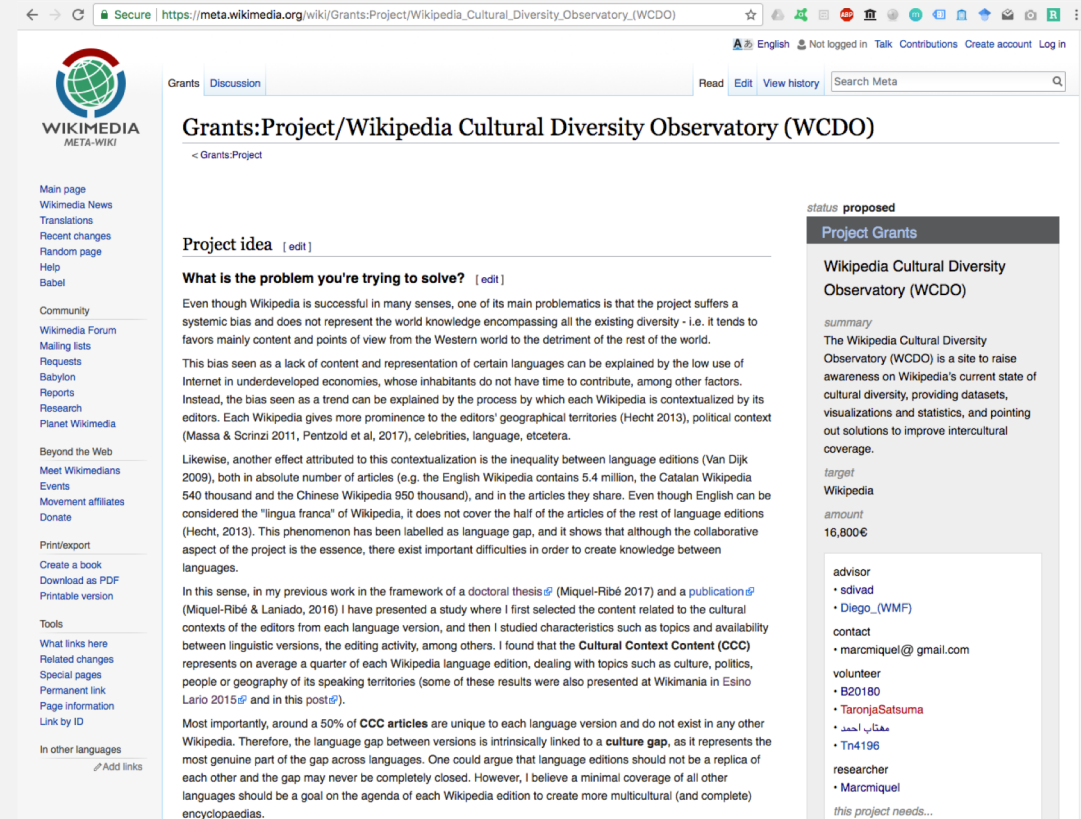


- First, because that many articles that should describe the world's cultural diversity do not exist due to the mentioned reasons.
- Second, because there exists some *language gaps*: Wikipedias do not cover each others content.

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.



The screenshot shows the Wikipedia Cultural Diversity Observatory (WCDO) grant page on Meta-Wiki. The page is titled "Grants:Project/Wikipedia Cultural Diversity Observatory (WCDO)" and is categorized as a "Project Grants" proposal. The status is "proposed" with a target amount of 16,800€.

Project idea [edit]

What is the problem you're trying to solve? [edit]

Even though Wikipedia is successful in many senses, one of its main problematics is that the project suffers a systemic bias and does not represent the world knowledge encompassing all the existing diversity - i.e. it tends to favor mainly content and points of view from the Western world to the detriment of the rest of the world.

This bias seen as a lack of content and representation of certain languages can be explained by the low use of Internet in underdeveloped economies, whose inhabitants do not have time to contribute, among other factors. Instead, the bias seen as a trend can be explained by the process by which each Wikipedia is contextualized by its editors. Each Wikipedia gives more prominence to the editors' geographical territories (Hecht 2013), political context (Massa & Scrinzi 2011, Pentzold et al, 2017), celebrities, language, etcetera.

Likewise, another effect attributed to this contextualization is the inequality between language editions (Van Dijk 2009), both in absolute number of articles (e.g. the English Wikipedia contains 5.4 million, the Catalan Wikipedia 540 thousand and the Chinese Wikipedia 950 thousand), and in the articles they share. Even though English can be considered the "lingua franca" of Wikipedia, it does not cover the half of the articles of the rest of language editions (Hecht, 2013). This phenomenon has been labelled as language gap, and it shows that although the collaborative aspect of the project is the essence, there exist important difficulties in order to create knowledge between languages.

In this sense, in my previous work in the framework of a doctoral thesis (Miquel-Ribé 2017) and a publication (Miquel-Ribé & Laniado, 2016) I have presented a study where I first selected the content related to the cultural contexts of the editors from each language version, and then I studied characteristics such as topics and availability between linguistic versions, the editing activity, among others. I found that the **Cultural Context Content (CCC)** represents on average a quarter of each Wikipedia language edition, dealing with topics such as culture, politics, people or geography of its speaking territories (some of these results were also presented at Wikimania in Esino Lario 2015 and in this post).

Most importantly, around a 50% of **CCC articles** are unique to each language version and do not exist in any other Wikipedia. Therefore, the language gap between versions is intrinsically linked to a **culture gap**, as it represents the most genuine part of the gap across languages. One could argue that language editions should not be a replica of each other and the gap may never be completely closed. However, I believe a minimal coverage of all other languages should be a goal on the agenda of each Wikipedia edition to create more multicultural (and complete) encyclopaedias.

Project Grants

Wikipedia Cultural Diversity Observatory (WCDO)

summary

The Wikipedia Cultural Diversity Observatory (WCDO) is a site to raise awareness on Wikipedia's current state of cultural diversity, providing datasets, visualizations and statistics, and pointing out solutions to improve intercultural coverage.

target

Wikipedia

amount

16,800€

advisor

- sdivad
- Diego_(WMF)

contact

- marcimiquel@gmail.com

volunteer

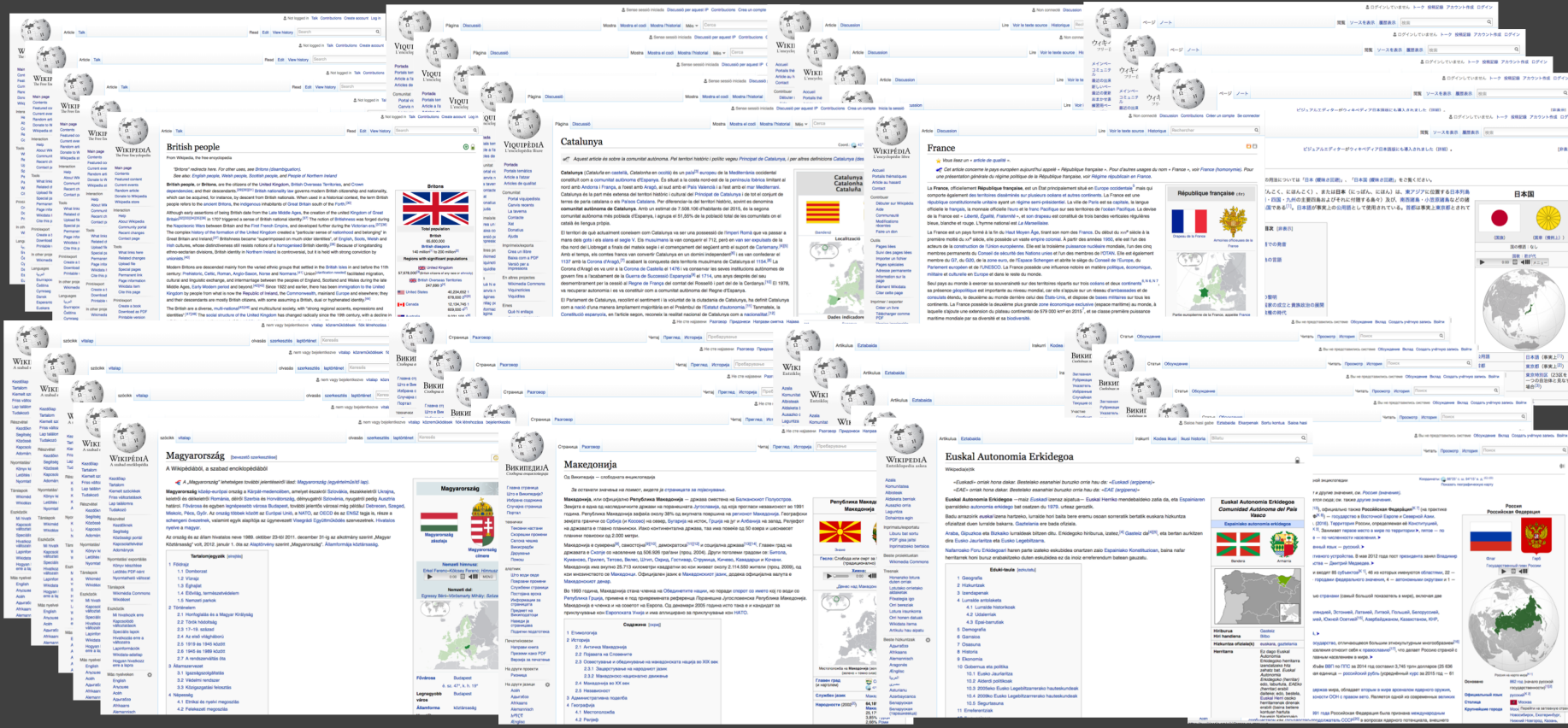
- BZ0180
- TaronjaSatsuma
- مهتاب احمد
- Tn4196

researcher

- Marcimiquel

this project needs...

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



In my research:
I select the **Cultural Context Content (CCC)**, i.e. the articles related to the editors' cultural contexts in each language version (traditions, language, politics, agriculture, biographies, places, events, etcetera).
This means associating each language with 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	Wiki	demon	demon	ISO3166	ISO31662	region	country	ind	lan	official	nt
2	Afar	Qafar	Q193494	Afar	aa			ET	ET-AF	yes	Ethiopia	yes		2 regional	0
3	Somali	Q202800	Afar	aa				ET	ET-SO	yes	Ethiopia	yes		2 regional	0
4	Amhara	Q203009	Afar	aa				ET	ET-AM	yes	Ethiopia	yes		2 regional	0
5	Ali Sabieh	Q821008	Afar	aa				DJ	DJ-AS	yes	Djibouti	yes		5 no	0
6	Arta	Q705941	Afar	aa				DJ	DJ-AR	yes	Djibouti	yes		5 no	0
7	Obock	Q844929	Afar	aa				DJ	DJ-OB	yes	Djibouti	yes		5 no	0
8	Dikhil	Q283979	Afar	aa				DJ	DJ-DI	yes	Djibouti	yes		5 no	0
9	Debubawi K'eyih	Q27728	Afar	aa				ER	ER-DU	yes	Eritrea	yes		5 no	0
10	Semenawi K'eyi B	Q27910	Afar	aa				ER	ER-SK	yes	Eritrea	yes			
11	Abkhazia	Аҟсны	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 Baròh	Q2140	Aceh	ace			ID	ID-SU	yes	Indonesia	yes		6 no	0
14	Republic of Adyge	Адыгэ	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-Cherke	Къарæдзæ-Чæркъес	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;	Botswana	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

My 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.

(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:

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

Wikidata Items that relate to groups of properties such as:

- **Language**
- **Country**
- **Part of**
- **In relation with**
- ...

Wikipedia MySQL db Replicas



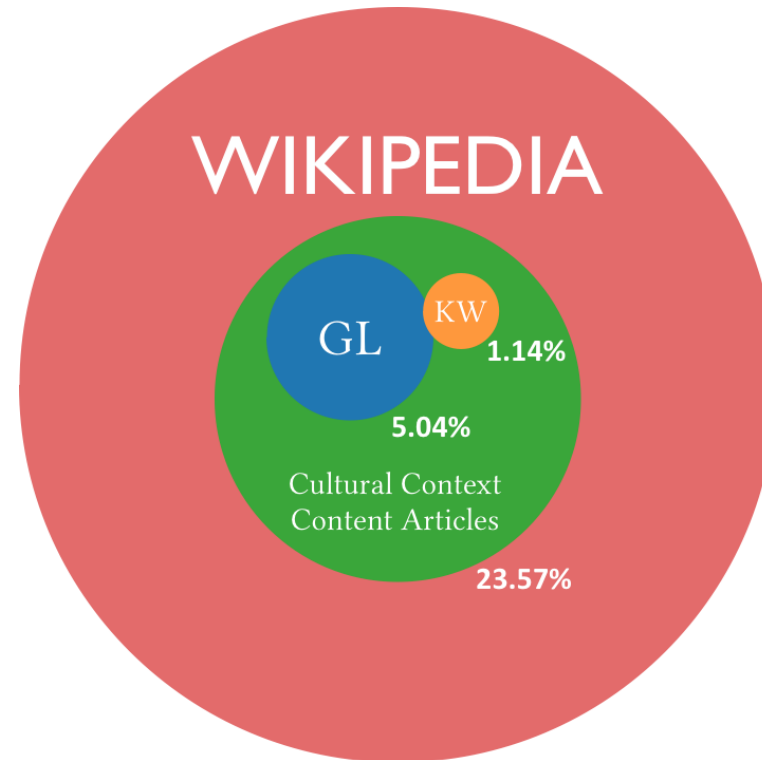
Wikidata XML dump



I create a CCC dataset as rich as possible.

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

I 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.

4.1 CCC Extent (African Languages, March, 2018):

LANGUAGE	LANGUAGE CODE	WP ARTICLES	CCC %	KW %	GL %
Afrikaans	af	49118	24.77	0.97	14.15
Swahili	sw	39844	19.34	0.46	2.35
Yoruba	yo	31774	7.64	0.37	2.58
Amharic	am	14689	2.08	0.19	1.14
Northern Sotho	nso	7890	12.40	1.91	17.20
Somali	so	5091	17.40	3.18	5.87
Shona	sn	3633	11.62	0.85	1.49
Kinyarwanda	rw	1930	5.70	1.50	1.24
Tongan	to	1699	27.55	4.30	12.42
Wolof	wo	1318	11.38	1.29	2.50
Oromo	om	1011	32.05	4.45	7.52
Xhosa	xh	993	1.21	0.50	2.11
Tswana	tn	733	46.79	10.50	51.30

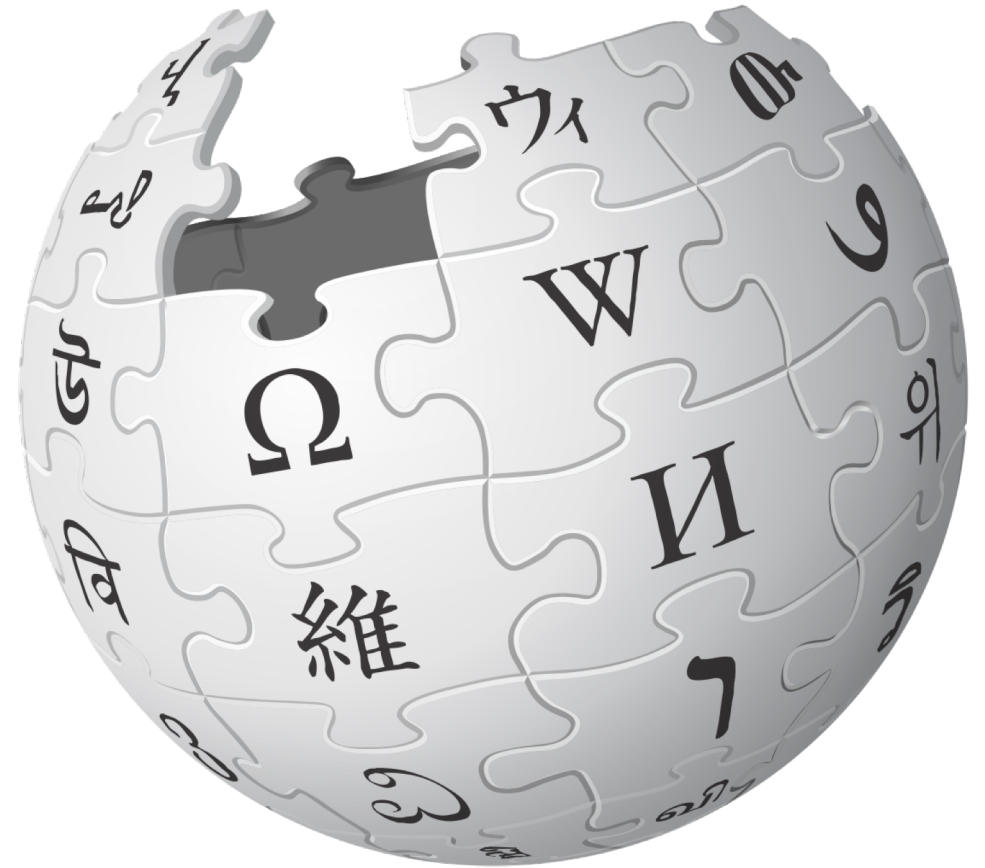
LANGUAGE	LANGUAGE CODE	WP ARTICLES	CCC %	KW %	GL %
Kirundi	rn	706	4.96	0.85	26.77
Tsonga	ts	658	13.22	2.58	0.61
Tumbuka	tum	635	0.47	0.47	0.16
Twi	tw	634	0.32	0.32	0.15
Sotho	st	596	10.23	5.37	19.63
Swazi	ss	462	13.64	3.25	6.49
Akan	ak	445	0.45	0.45	14.16
Venda	ve	326	11.96	2.76	10.12
Tigrinya	ti	296	6.76	1.69	5.41
Sango	sg	281	18.51	1.42	19.93
Afar	aa	2	0.00	0.00	0.00
AVG.	-	6865.17	12.52	2.07	9.39

African languages measured Cultural Context Content (CCC) accounted for about a 13%. This is around half the percentage found for the top Wikipedias measured two years ago.

This 13% of CCC in African language editions is too little.



African Wikipedias are the ones who can explain how Africa is the best way.



CCC articles tend to be more developed. Editors tend to have more access to the sources of information, know the difference points of view on the same topic, among other reasons.

4.2 Culture Gap: the 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.



How about the African languages?

How well are African CCC articles covered?

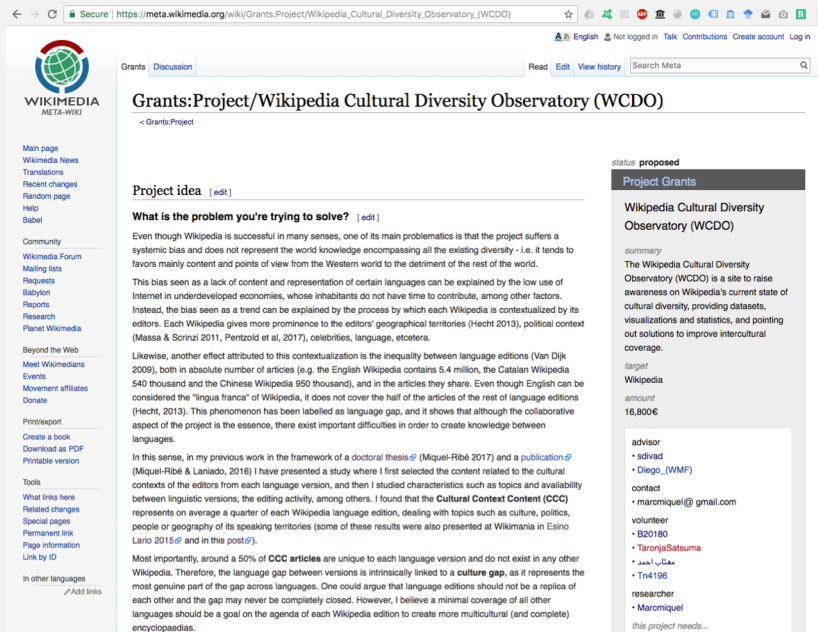
	de	en	es	fr	it	ja	nl	pl	ru	sv	vi	Language	CCC art.
Afrikaans	5.22	13.05	2.96	5.17	3.39	1.49	4.63	3.04	2.83	2.24	0.97	Afrikaans	12.168
Akan	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.23	Akan	2
Amharic	0.08	0.81	0.12	0.73	0.14	0.06	0.07	0.10	0.10	0.28	0.04	Amharic	306
Kinyarwanda	2.07	3.83	2.12	3.01	2.64	1.76	1.61	1.71	1.76	1.50	0.78	Kinyarwanda	110
Kirundi	4.39	4.53	3.97	4.39	4.53	4.39	3.54	4.25	4.11	4.67	0.43	Kirundi	35
Northern S..	1.84	6.29	0.49	1.41	1.32	0.27	1.80	1.23	1.24	0.42	0.13	Northern Soto	978
Oromo	5.05	12.07	3.46	8.90	3.76	2.87	3.36	3.46	3.86	5.24	1.29	Oromo	324
Sango	12.10	17.79	9.96	17.79	17.79	11.03	12.46	16.73	15.30	15.66	2.14	Sango	52
Shona	3.03	4.10	2.53	2.89	2.34	2.23	2.37	2.48	2.73	2.31	1.62	Shona	422
Somali	5.21	12.16	3.73	4.38	4.83	3.61	3.26	3.87	4.01	3.03	1.14	Somali	886
Sotho	7.22	9.90	5.87	7.22	5.54	3.36	6.38	6.88	5.71	4.87	2.18	Sotho	61
Swahili	3.60	8.33	3.40	3.98	3.64	2.75	3.09	3.33	3.25	3.03	2.57	Swahili	7.707
Swazi	6.06	11.91	6.71	6.49	8.23	5.41	4.55	8.44	5.63	4.98	2.38	Swazi	63
Tigrinya	5.74	6.08	5.07	5.74	5.74	5.07	4.73	5.74	5.74	5.41	1.01	Tigrinya	20
Tongan	2.53	7.89	2.18	3.24	1.65	1.12	1.00	2.71	2.94	1.88	0.59	Tongan	468
Tsonga	3.04	6.69	5.02	3.95	3.04	1.82	2.28	2.28	2.89	6.08	5.78	Tsonga	87
Tswana	9.55	39.29	10.78	29.60	27.42	7.64	26.06	27.29	12.01	24.56	1.23	Tswana	343
Tumbuka	0.47	0.32	0.47	0.32	0.32	0.32	0.16	0.47	0.32	0.16		Tumbuka	3
Twi	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.16	0.32	Twi	2
Venda	6.44	10.43	1.53	7.06	5.52	2.15	5.52	4.29	5.22	1.53	1.23	Venda	39
Wolof	5.62	6.75	5.24	7.06	5.69	4.25	4.55	3.72	5.08	3.49	1.06	Wolof	150
Xhosa	1.01	1.11	0.81	0.91	0.81	0.81	0.91	0.71	0.81	0.81	0.60	Xhosa	12
Yoruba	1.22	6.30	0.93	3.02	2.78	0.67	0.69	1.06	1.02	0.89	0.13	Yoruba	2.429

African languages CCC articles are not extensively covered even they are not many.

5. Activism (Creating lists of top priority articles)

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.



The screenshot shows the project page for the Wikipedia Cultural Diversity Observatory (WCDO) on the Meta-Wiki platform. The page title is "Grants:Project/Wikipedia Cultural Diversity Observatory (WCDO)". The main content area is titled "Project Grants" and includes a "Project idea" section. The text under "Project idea" discusses the problem of cultural diversity on Wikipedia, noting that the current state of cultural diversity is not optimal and that the project aims to address this by providing datasets, visualizations, and statistics to improve intercultural coverage. It also mentions that the project is currently in a "proposed" status and has a target amount of 16,800€.

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.

I will ask the communities to generate a **Top 100 Vital Articles lists from CCC**.

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



There is nothing more Wikimedian than multiculturality, embrace it and collaborate across languages.

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

