

# 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)

Lviv, Ukraine 2018



In 2010, I got interested in Wikipedia, especially when I realized that each language has different content.

**So I decided to study it.**





**How is my culture represented in Wikipedia?**

## The Problem

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



Some voices are missing or underrepresented



*"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

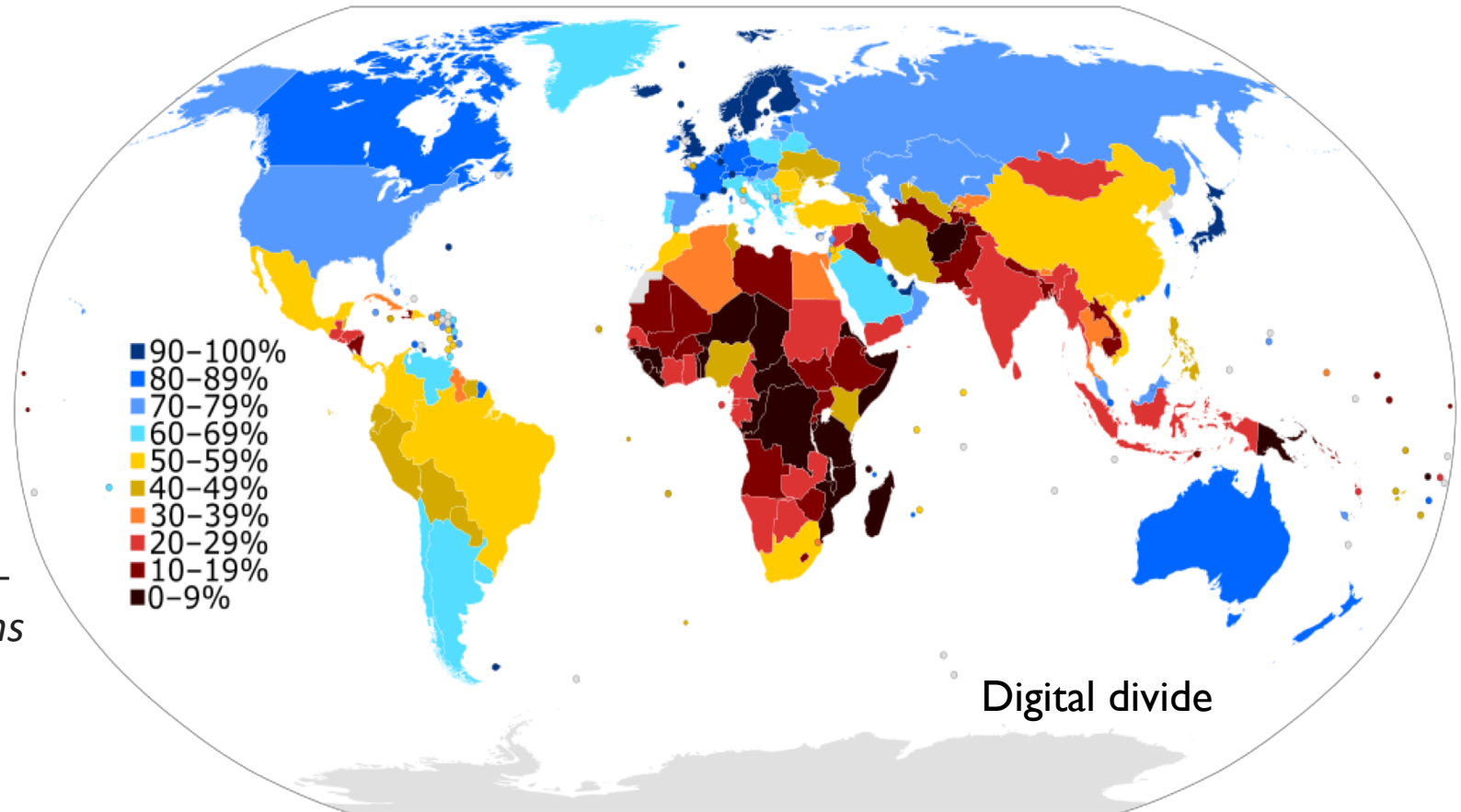




**1. Many articles that should describe the world's cultural diversity do not exist because not everyone has a Wikipedia, or cannot contribute to it.**

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.



**2. Some articles that represent the cultural particularities and contexts specificities exist... But nobody outside their local Wikipedia knows about them.**

**This is where we can actually work.**

**This is how I hope I can help CEE Spring.**

## Proposed Solution

### Wikipedia Cultural Diversity Observatory (WCDO).

“a joint space for **researchers** and **activists** to study and **fight against the knowledge gaps** and promote knowledge equity.”

- **Awareness**
- **Solutions**



<http://wcdo.wmflabs.org> (beta 1)

It is impossible to bridge all the knowledge gaps between languages.



## WCDO Main goal

“Every Wikipedia language edition ensures a minimal coverage of each other language cultural and geographical content.”





I propose every Wikipedia has 100 articles about every other language's cultural and geographical content.

This is 28-30 thousand articles.



## Top CCC articles lists

From each language, those articles from their cultural context which are more relevant according to specific features:

- List = [editors, featured, geolocated, keywords, women, men, created\_first\_three\_years, created\_last\_year, pageviews, discussions]
- Country\_origin (optional) = ISO3166 code
- Lang\_origin = wikicode
- Lang\_target = wikicode

[http://wcdo.wmflabs.org/top\\_ccc\\_articles/?list=men&lang\\_origin=pl&lang\\_target=uk](http://wcdo.wmflabs.org/top_ccc_articles/?list=men&lang_origin=pl&lang_target=uk)

# Top articles in Romanian CCC by editors and their availability in Polish

In this page, you can consult Top CCC articles list from any country or language CCC generated with the latest CCC dataset. The following table shows the Top 500 articles from Romanian CCC by editors and their availability in Polish Wikipedia. They are sorted by the feature **editors**. The rest of columns present complementary features that are explicative of the article relevance. In particular, number of Inlinks from CCC highlights the article importance in terms of how much is required by other articles from the Cultural Context Content.

The available Top CCC articles lists are: list of CCC articles with most number of editors (**Editors**), list of CCC articles with featured article distinction (**Featured**), most bytes and references (weights: 0.8, 0.1 and 0.1 respectively), list of CCC articles with geolocation with most links coming from CCC, list of CCC articles with keywords on title with most bytes (**Bytes**), list of CCC articles categorized in Wikidata as women with most edits (**Women**), list of CCC articles categorized in Wikidata as men with most edits (**Men**), list of CCC articles created during the first three years and with most edits (**First 3Y.**), list of CCC articles created during the last year and with most edits (**Last Y.**), list of CCC articles with most pageviews during the last month (**Pageviews**), list of CCC articles with most edits in talk pages (**Discussions**). The table's last column shows the title the article has in its target language, in blue when it exists and in red as a proposal generated with the Wikimedia Content Translation tool or as an existing Wikidata label in the same language.

It is possible to query any list by changing the URL parameters. You need to specify the list parameter (editors, featured, geolocated, keywords, women, men, created\_first\_three\_years, created\_last\_year, pageviews and discussions), the language target parameter (as lang\_target and the language wikicode), the language origin (as lang\_origin and the language wikicode), and, optionally to limit the scope of the selection, the country\_origin parameter as part of the CCC (as country\_origin and the country ISO3166 code). In case no country is selected, the default is 'all'.

N°	Romanian Title	Editors	Pageviews	Bytes	References	Wikidata Properties	Interwiki Links	Inlinks from CCC	Creation Date	Other Languages	Polish Title
1	<a href="#">România</a>	1064	4774	213456	410	<a href="#">137</a>	<a href="#">298</a>	42303	2003-07-15	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Rumunia</a>
2	<a href="#">București</a>	802	2111	167239	193	<a href="#">63</a>	<a href="#">205</a>	7117	2003-07-26	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Bukareszt</a>
3	<a href="#">Republica Moldova</a>	724	1503	185309	187	<a href="#">123</a>	<a href="#">278</a>	4958	2003-08-02	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Mołdawia</a>
4	<a href="#">FC Dinamo București</a>	712	417	70889	62	<a href="#">14</a>	<a href="#">41</a>	482	2005-05-05	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">FC Dinamo Bukareszt</a>
5	<a href="#">Iași</a>	613	739	110146	74	<a href="#">39</a>	<a href="#">94</a>	2043	2004-01-31	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Jassy</a>
6	<a href="#">Cluj-Napoca</a>	606	1063	115794	107	<a href="#">52</a>	<a href="#">99</a>	1289	2003-11-11	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Kluż-Napoka</a>
7	<a href="#">FC Rapid București</a>	568	282	83241	37	<a href="#">12</a>	<a href="#">33</a>	309	2004-06-05	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">FC Rapid Bukareszt</a>
8	<a href="#">Brașov</a>	547	757	93804	62	<a href="#">41</a>	<a href="#">92</a>	1490	2004-01-31	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Braszów</a>
9	<a href="#">Timișoara</a>	519	675	205305	182	<a href="#">46</a>	<a href="#">94</a>	1410	2004-02-02	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Timișoara</a>

# Top articles in Catalan CCC by men and their availability in Polish

In this page, you can consult Top CCC articles list from any country or language CCC generated with the latest CCC dataset. The following table shows the Top 500 articles from Catalan CCC by men and their availability in Polish Wikipedia. They are sorted by the feature *men*. The rest of columns present complementary features that are explicative of the article relevance. In particular, number of Inlinks from CCC highlights the article importance in terms of how much is required by other articles from the Cultural Context Content.

The available Top CCC articles lists are: list of CCC articles with most number of editors (**Editors**), list of CCC articles with featured article distinction (**Featured**), most bytes and references (weights: 0.8, 0.1 and 0.1 respectively), list of CCC articles with geolocation with most links coming from CCC, list of CCC articles with keywords on title with most bytes (**Bytes**), list of CCC articles categorized in Wikidata as women with most edits (**Women**), list of CCC articles categorized in Wikidata as men with most edits (**Men**), list of CCC articles created during the first three years and with most edits (**First 3Y.**), list of CCC articles created during the last year and with most edits (**Last Y.**), list of CCC articles with most pageviews during the last month (**Pageviews**), list of CCC articles with most edits in talk pages (**Discussions**). The table's last column shows the title the article has in its target language, in blue when it exists and in red as a proposal generated with the Wikimedia Content Translation tool or as an existing Wikidata label in the same language.

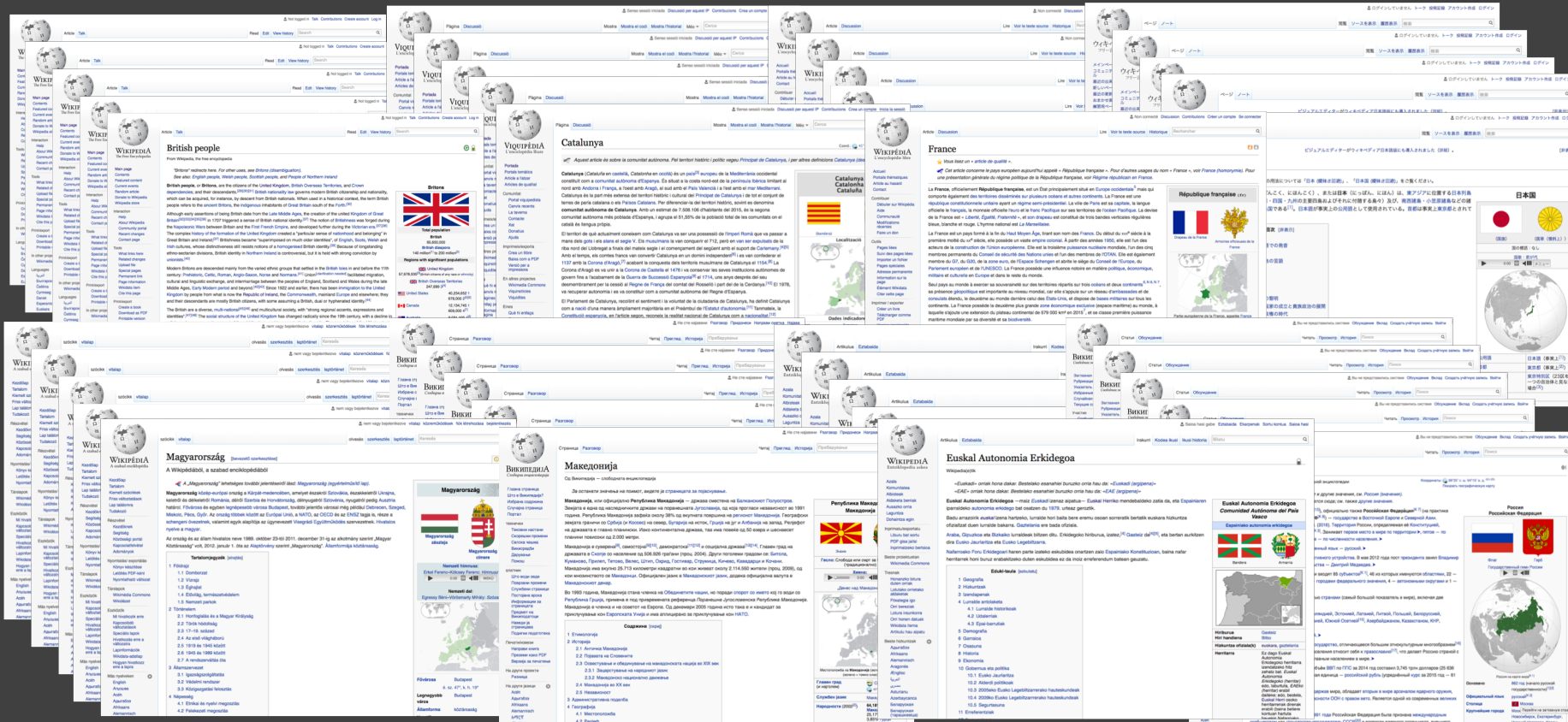
It is possible to query any list by changing the URL parameters. You need to specify the list parameter (editors, featured, geolocated, keywords, women, men, created\_first\_three\_years, created\_last\_year, pageviews and discussions), the language target parameter (as lang\_target and the language wikicode), the language origin (as lang\_origin and the language wikicode), and, optionally to limit the scope of the selection, the country origin parameter as part of the CCC (as country\_origin and the country ISO3166 code). In case no country is selected, the default is 'all'.

Nº	Catalan Title	Edits	Editors	Pageviews	Bytes	References	Wikidata Properties	Interwiki Links	Inlinks from CCC	Creation Date	Other Languages	Polish Title
1	<a href="#">Ramon Llull</a>	1569	449	536	110.2k	98	<a href="#">82</a>	<a href="#">56</a>	136	2003-11-10	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Rajmund Llull</a>
2	<a href="#">Antoni Gaudí i Cornet</a>	1214	371	144	90.2k	57	<a href="#">81</a>	<a href="#">115</a>	131	2003-08-10	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Antoni Gaudí</a>
3	<a href="#">Jacint Verdaguer i Santaló</a>	1081	262	82	93.7k	41	<a href="#">56</a>	<a href="#">33</a>	95	2003-08-11	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Jacint Verdaguer i Santaló</a>
4	<a href="#">Rafael Casanova i Comes</a>	1062	146	2371	268.2k	52	<a href="#">27</a>	<a href="#">24</a>	8	2005-09-16	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	
5	<a href="#">Joan Miró i Ferrà</a>	920	227	55	98.4k	64	<a href="#">114</a>	<a href="#">94</a>	91	2006-01-29	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Joan Miró</a>
6	<a href="#">Salvador Dalí i Domènech</a>	880	366	118	49.3k	52	<a href="#">145</a>	<a href="#">195</a>	46	2003-06-10	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Salvador Dalí</a>
7	<a href="#">Artur Mas i Gavarró</a>	841	265	873	54.0k	95	<a href="#">41</a>	<a href="#">50</a>	98	2004-05-20	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Artur Mas</a>
8	<a href="#">Guifré el Pilós</a>	826	190	152	95.6k	33	<a href="#">28</a>	<a href="#">27</a>	83	2004-12-03	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Wilfred Włochaty</a>
9	<a href="#">Josep Guardiola i Sala</a>	821	272	1884	44.3k	53	<a href="#">55</a>	<a href="#">78</a>	81	2005-10-20	<a href="#">en</a> , <a href="#">fr</a> , <a href="#">de</a> , <a href="#">ru</a>	<a href="#">Josep Guardiola</a>
10	<a href="#">Pablo Picasso</a>	774	308	106	47.7k	56						
11	<a href="#">Jordi Bilbeny</a>	772	153	41	19.5k	39						
12	<a href="#">Joan Maragall i Gorina</a>	758	228	531	72.0k	81						
13	<a href="#">Marc Márquez i Alentà</a>	749	164	189	51.3k	71						
14	<a href="#">Francisco Franco Bahamonde</a>	741	249	184	43.0k	37						
15	<a href="#">Carles Puigdemont i Casamajó</a>	706	184	3713	64.1k	118						



**Why was there such interference?**

**Let me explain you the whole story about CCC...**



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

*ccc\_setup\_language\_context\_mapping.py*

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

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://wcd0.wmflabs.org/language\\_territories\\_mapping/](https://wcd0.wmflabs.org/language_territories_mapping/)

## For example:

Italy	Italia	Q38	Italian	it	Italian	italiano;ita	IT		no	Italy
Istria	Istria	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://wcd0.wmflabs.org/language\\_territories\\_mapping/](https://wcd0.wmflabs.org/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).**

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

Wikipedia MySQL db Replicas



Wikidata JSON dump



**We create a database as rich as possible.**

# Some of these features are reliable CCC, while some other are potential CCC.

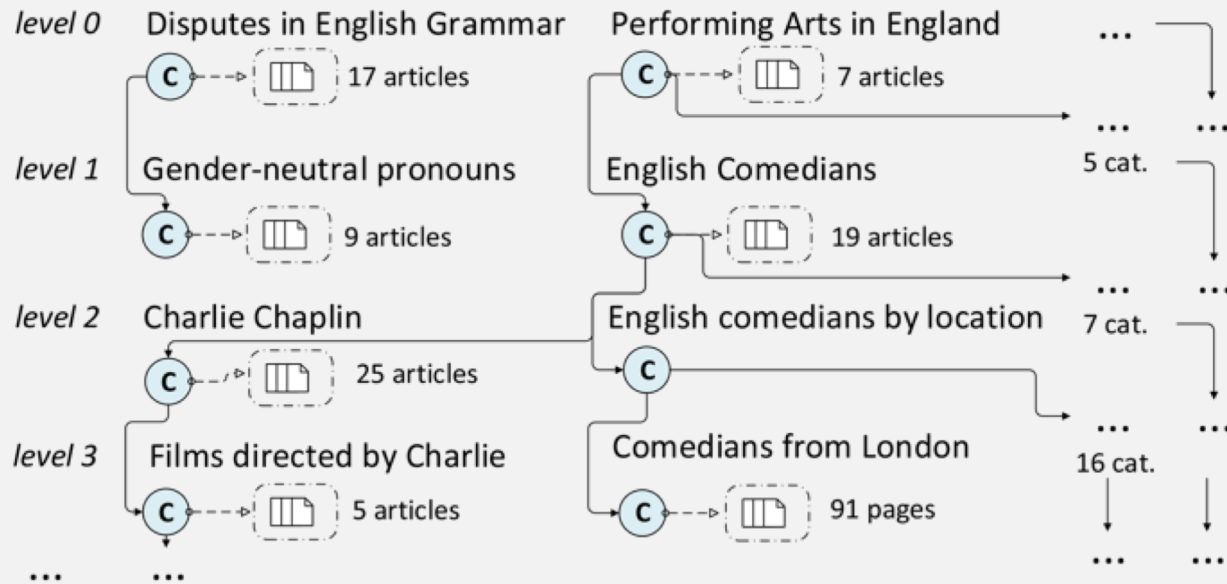
The screenshot shows the Wikipedia article for "English literature". The title "English literature" is prominently displayed at the top. Below the title, there is a brief introductory sentence. The main body of the article contains several paragraphs of text. On the right side, there is a grid of nine portrait images of English-language writers. Below the grid is a caption: "Selected English-language writers: (left to right, top to bottom) Geoffrey Chaucer, William Shakespeare, Jane Austen, Mark Twain, Virginia Woolf, T. S. Eliot, Vladimir Nabokov, Toni Morrison, Salman Rushdie." On the left side, there is a "Contents" table of contents with a "hide" link. The table lists sections such as "Old English literature: c. 450–1066", "Middle English literature: 1066–1500", "English Renaissance: 1500–1660", "Jacobean period: 1603–25", "Late Renaissance: 1625–1660", "Restoration Age: 1660–1700", and "18th century".

The screenshot shows the Wikipedia article for "Times Square". The title "Times Square" is prominently displayed at the top. Below the title, there is a brief introductory sentence. The main body of the article contains several paragraphs of text. On the right side, there is a grid of two images showing Times Square at night. Below the images is a caption: "Broadway show billboards in Times Square, 2009 (top), 2013 (bottom)". On the right side, there is a table with information about Times Square, including its nickname "The Great White Way", its location in New York City, its boundaries, and its subway services. Below the table, there is a "Contents" table of contents with a "hide" link. The table lists sections such as "History", "Early history", "1900s–1930s", "1930s–1950s", and "1960s–1980s".

- **Keyword (demonym/territory name) on title is Reliable CCC**

- **Geolocation in one of the territories is Reliable CCC**

Keywords {English, England, Ireland, Irish, etc.}



### Category crawling using keywords

**Dylan Moran**  
From Wikipedia, the free encyclopedia

**Dylan William Moran** (/ˈmɔːrən/; born 3 November 1971)<sup>[1]</sup> is an Irish comedian, writer, actor and filmmaker. He is best known for his observational comedy, the television sitcom *Black Books* (in which he starred and co-wrote) and his work with *Simon Pegg* in *Shaun of the Dead* and *Run Fatboy Run*. He appeared as one of the two lead characters in the Irish black comedy titled *A Film with Me in It* in 2008.

Moran's most recent film is *Calvary*, an Irish black comedy drama film written and directed by John Michael McDonagh. Moran is a regular performer at national and international comedy festivals including the Edinburgh Festival Fringe, Just for Laughs Montreal Comedy Festival, the Melbourne International Comedy Festival and the Kilkenny Comedy Festival. In 2007, Moran was voted the 17th greatest stand-up comic on Channel 4's 100 Greatest Stand-Ups and again in the updated 2010 list as the 14th greatest stand-up comic. He lives in Edinburgh with his wife, Elaine, and two children.

**Contents** [hide]

- Biography
  - Early life
  - Career
  - Awards and commendations
- Filmography
  - Film
  - Television
- Stand-up DVDs
- References
- External links

**Biography** [edit]

**Early life** [edit]

Moran was born in Navan, County Meath, Ireland.<sup>[1][2][3][4]</sup> He attended St. Patrick's Classical School, where he experimented early on with

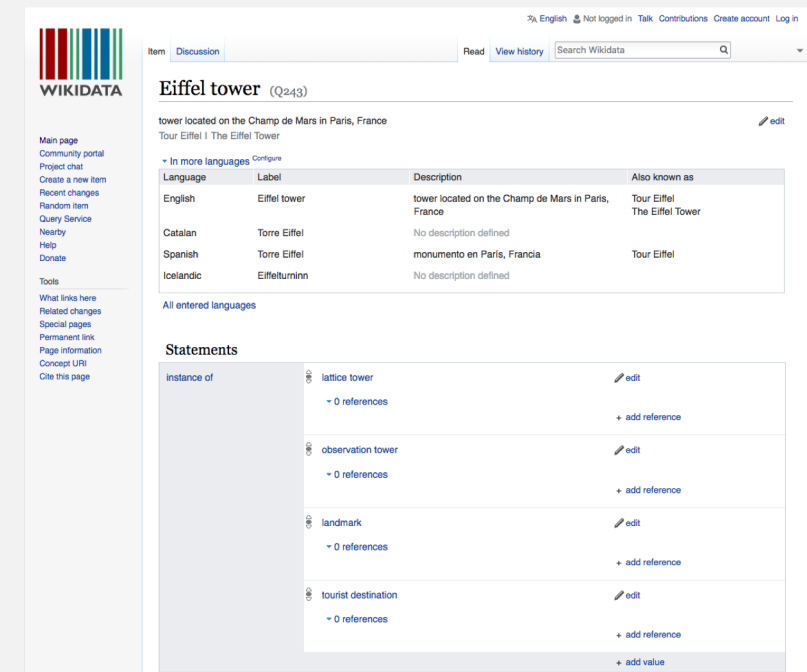
- Being in a subcategory of a category containing a keyword on its title is potential CCC

## Some Wikidata properties are reliable CCC

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

## Some Wikidata properties are potential CCC

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



The screenshot shows the Wikidata page for 'Eiffel tower' (Q243). The page includes a navigation menu on the left, a search bar at the top, and a main content area. The main content area displays the item's description in French, a table of labels in various languages, and a list of statements.

Language	Label	Description	Also known as
English	Eiffel tower	tower located on the Champ de Mars in Paris, France	Tour Eiffel The Eiffel Tower
Catalan	Torre Eiffel	No description defined	
Spanish	Torre Eiffel	monumento en París, Francia	Tour Eiffel
Icelandic	Eiffeltúminn	No description defined	

Statements:

- instance of: lattice tower (0 references)
- observation tower (0 references)
- landmark (0 references)
- tourist destination (0 references)

**With those labelled as potential CCC we cannot be sure how representative the feature is to be included as Cultural Context Content.**

### **Links feature:**

- **Number and percentage of Inlinks/Outlinks (incoming/outgoing links) to CCC is very explicative on how an article is needed to expand CCC or is dedicated to CCC.**

### **Some negative features:**

- **Geolocated articles not in CCC (reliable not in CCC)**
- **Other CCC Wikidata properties (potential not in CCC)**
- **Other Language CCC Category Crawling (potential not in CCC)**
- **Number and Percentage of Inlinks/Outlinks to geolocated articles not in CCC (potential not in CCC)**

**With the features we establish a groundtruth (articles we know they already are CCC and articles we know they are not CCC).**

## 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

The screenshot shows the English Wikipedia article for Noam Chomsky. The article text includes: "Avram Noam Chomsky (born December 7, 1928) is an American linguist, philosopher, cognitive scientist, historian, social critic, and political activist. Sometimes described as 'the father of modern linguistics', Chomsky is also a major figure in analytic philosophy and one of the founders of the field of cognitive science. He holds a joint appointment as Institute Professor Emeritus at the Massachusetts Institute of Technology (MIT) and laureate professor at the University of Arizona...". The article also mentions his work on generative grammar and his vocal opposition to U.S. involvement in the Vietnam War.

The screenshot shows the French Wikipedia article for Alsace. The article text includes: "L'Alsace (prononcé [al.ˈzas]; Elsass en allemand; 'Elsäss en alsacien) est une région culturelle et historique du nord-est de la France à la frontière avec l'Allemagne et la Suisse. Elle est constituée des départements du Bas-Rhin et du Haut-Rhin. Ses habitants sont appelés les Alsaciens. Géographiquement elle se trouve entre le massif des Vosges et le Rhin. Région de l'Europe rhénane, l'Alsace se situe au cœur de la « banane bleue »". The article also features a map of Alsace and its coat of arms.

The screenshot shows the Guarani Wikipedia article for the year 1977. The article text includes: "Oararecha'akue [ Jehajey | editar código ] • Justo Villar - 30 jasytopytĩ". The article also mentions "Omano'akue [ Jehajey | editar código ] • Remberto Giménez - 15 jasykõi • Arsenio Erico - 23 jasypokõi". The article is in the Guarani language and is part of the CCC project.

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\_setup\_language\_context\_mapping.py*** (it creates the language territories equivalences database).

***ccc\_selection.py*** (it creates the main database `ccc_current.db` and the datasets).

***stats\_generation.py*** (it creates the database `wcdo_data.db` with the main statistics).

***meta\_update.py*** (it updates stats in meta with ***pywikibot*** and in the website).

- **Datasets**

They are available at [wcdo.wmflabs.org/datasets](http://wcdo.wmflabs.org/datasets) and at [figshare.com/account/home#/projects/28272](https://figshare.com/account/home#/projects/28272)

- **Code in Github**

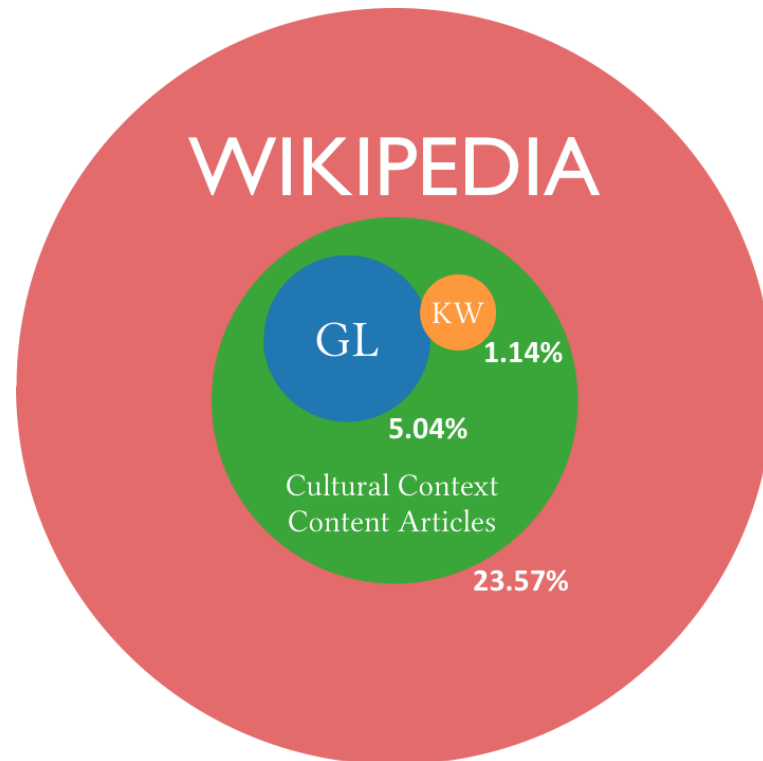
All the code, presentation and files are at: [github.com/marcmiquel/WCDO](https://github.com/marcmiquel/WCDO)

**Do you want to join? E-mail me at [marcmiquel@gmail.com](mailto:marcmiquel@gmail.com)**



**Taking into account the main languages, CCC is about a quarter of the Wikipedia.**

**CCC articles percentage is very variable.**



## **What is the extent of this language CCC?**

**List of Wikipedias by Cultural Context Content**

**List of Language Territories by Cultural Context Content**

[http://wcdo.wmflabs.org/list\\_of\\_wikipedias\\_by\\_cultural\\_context\\_content](http://wcdo.wmflabs.org/list_of_wikipedias_by_cultural_context_content)

[http://wcdo.wmflabs.org/list\\_of\\_language\\_territories\\_by\\_cultural\\_context\\_content](http://wcdo.wmflabs.org/list_of_language_territories_by_cultural_context_content)

# Lists of Wikipedias by Cultural Context Content

This page contains a list of all the current Wikipedia language editions ordered by their number of articles from their Cultural Context Content dataset that relate to territories where the language is spoken as official or as indigeneous.

For each language edition, statistics account for the number of articles of different CCC segments and their percentage computed in relation to the overall total number of Wikipedia articles. This is **(CCC art.)** and **CCC (%)** as the number of CCC articles and percentage, **CCC GL (%)** as the number of articles from CCC that are geolocated, **KW Title (%)** as the number of articles from CCC that contain specific keywords (language name, territory name or demonym) in their titles. Finally, **Region** (continent) and **Subregion** are introduced in order to contextualize the results.

<input type="checkbox"/>	N°	Language	Wiki	Articles	CCC art.	CCC %	Subregion	Region
<input type="checkbox"/>	295	Afar	aa	1	0	0	Sub-Saharan /	Africa
<input type="checkbox"/>	219	Abkhaz	ab	3471	143	4.12	Western Asia	Asia
<input type="checkbox"/>	91	Acehnese	ace	7478	4689	62.7	South-eastern	Asia
<input type="checkbox"/>	264	Adyghe	ady	544	32	5.88	Eastern Europe	Europe
<input type="checkbox"/>	66	Afrikaans	af	52138	12510	23.99	Sub-Saharan /	Africa
<input type="checkbox"/>	234	Akan language	ak	617	92	14.91	Sub-Saharan /	Africa
<input type="checkbox"/>	74	Alemannic	als	24820	10420	41.98	Western Europe	Europe
<input type="checkbox"/>	178	Amharic	am	14815	395	2.67	Sub-Saharan /	Africa
<input type="checkbox"/>	90	Aragonese	an	33081	4845	14.65	Southern Europe	Europe
<input type="checkbox"/>	197	Old English	ang	3058	262	8.57	Northern Europe	Europe
<input type="checkbox"/>	10	Arabic	ar	588607	240034	40.78	Western Asia	Asia

# List of Language Territories by Cultural Context Content

This page contains each Wikipedia language edition Cultural Context Content divided in its territories according to the language territories mapping. Articles are assigned to territories according to the different strategies that have been used to include them into CCC. The label Not Assigned is for those articles which were not possible to classify.

For each territory, statistics account for the number of articles of different CCC segments and their percentage computed in relation to the overall total number of Wikipedia articles. This is **(CCC art.)** and **CCC (%)** as the number of CCC articles and percentage, **CCC GL (%)** as the number of articles from CCC that are geolocated, **KW Title (%)** as the number of articles from CCC that contain specific keywords (language name, territory name or demonym) in their titles. Finally, **Region** (continent) and **Subregion** are introduced in order to contextualize the results.

<input type="button" value="FILTER ROWS"/>											
<input type="checkbox"/>	Qitem	Territory name	Language	Wiki	CCC art.	CCC %	CCC GL art.	CCC art. KW	ISO3166	ISO3166-2	subregion
<input type="checkbox"/>	Q23334	Abkhazia	Abkhaz	ab	81	56.64	42	8	GE	GE-AB	Western Europe
<input type="checkbox"/>	Not Assigned		Abkhaz	ab	62	43.36	54	7			
<input type="checkbox"/>	Q2140	Sumatera	Acehnese	ace	32	0.68	31	0	ID	ID-SU	South East Asia
<input type="checkbox"/>	Q1823	Aceh	Acehnese	ace	4059	86.56	3246	1293	ID	ID-AC	South East Asia
<input type="checkbox"/>	Not Assigned		Acehnese	ace	598	12.75	449	158			
<input type="checkbox"/>	Q5328	Karachay-Cherkessia	Adyghe	ady	2	6.25	1	0	RU	RU-KC	Eastern Europe
<input type="checkbox"/>	Q3734	Republic of Chechnya	Adyghe	ady	14	43.75	11	0	RU	RU-AD	Eastern Europe
<input type="checkbox"/>	Q3680	Krasnodar Krai	Adyghe	ady	2	6.25	2	0	RU	RU-KDA	Eastern Europe
<input type="checkbox"/>	Not Assigned		Adyghe	ady	14	43.75	4	7			
<input type="checkbox"/>	Q963	Botswana	Afrikaans	af	81	0.65	56	20	BW		Sub-Saharan Africa
<input type="checkbox"/>	Q258	South Africa	Afrikaans	af	11969	95.68	4290	399	ZA		Sub-Saharan Africa

## Few results from the List of the Wikipedias and territories:

- CCC % is not related to language size.
- Bot-created Wikipedias CCC % is very low.
- From CEE, biggest are Crimean Tatar, Taraškievica and Russian, which oscillate between 32-44%.
- *Imperial* languages tend to have a high CCC % (English has).
- Dutch and Swedish Wikipedias are big and dilute the CCC%.
- Isolated cultures like Japanese has also a high CCC %.
  
- And the most important...

**The extent of CCC from certain language editions is too little (non-western, especially African and Asian).**

**Each Wikipedia editors are the ones who can explain their context in its the best way.**



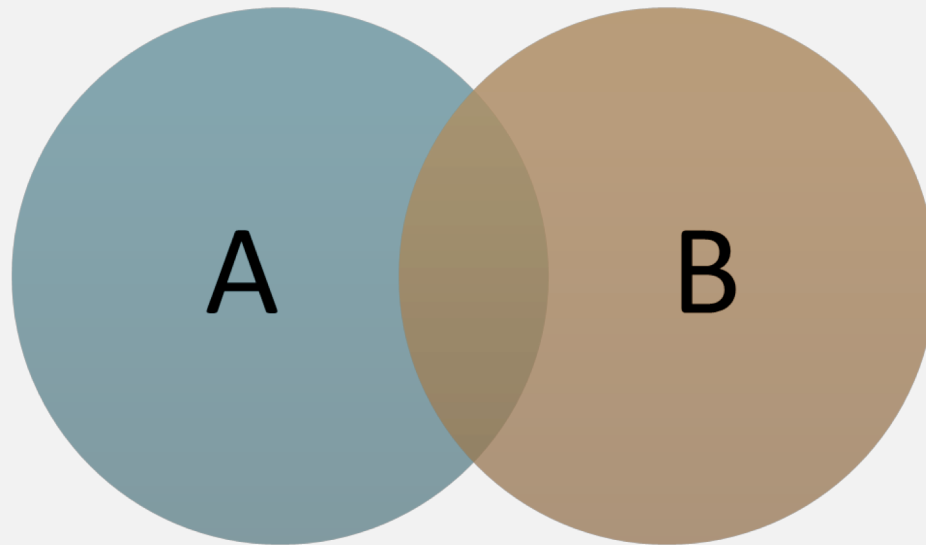
CCC articles tend to be more developed. Editors tend to have more access to the sources of information, know the difference points of view on the same topic, among other reasons.

# INTERSECTIONS AND INCREMENTS

Once the **CCC dataset** is obtained, we can compute the different intersections with other groups in order to understand its extent and scope.

## Main entities:

- CCC Segments (Geolocated, those with keywords on title)
- People (Men and Women)
- Geolocated articles (Continents, world regions and countries)
- Articles without interlanguage links
- All Wikipedia articles
- ...





**What articles do languages have in common?**



## **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.**



## How do languages cover each others' CCC?

These are panels to obtain a general view on the coverage and spread for the entire CCC.

[https://wcdo.wmflabs.org/ccc\\_spread](https://wcdo.wmflabs.org/ccc_spread)

[https://wcdo.wmflabs.org/ccc\\_coverage](https://wcdo.wmflabs.org/ccc_coverage)

# What is the weight of each language CCC in other languages? (Spread)

## Wikipedia Language Editions' CCC Spread

This page shows some statistics that explain how well each Wikipedia language edition [Cultural Context Content \(CCC\)](#) articles are spread across other languages.

The following table is particularly useful in order to understand the content **culture gap** between language editions, that is the imbalances across languages editions in content representing cultural context. Specifically, it shows which language CCC is more popular among all Wikipedia language editions by counting the CCC spread articles, i.e. articles from one language CCC that exist in other language editions.

Languages are sorted in alphabetic order by their Wikicode, and the columns present the following statistics: (**CCC %**) the number of CCC articles and the percentage it occupies in the language computed in relation to their total number of articles, the **first five other languages** with more spread articles from the language CCC and the percentage they occupy computed in relation to their corresponding total number of articles, the relative spread (**R. Spread**) of a language CCC across all the other languages computed as the average of the percentage they occupy in each other language edition, the total spread (**T. Spread**) of a CCC across all the other languages computed as the percentage in relation to all languages articles (not counting the own), and finally, the total number of language CCC articles (Spread Art.) that exists across all the other language editions.

<input type="checkbox"/>	Language	CCC art.	CCC no IW	n°1	n°2	n°3	n°4	n°5	R.Spread	T.Spread	Spread Art.	Region	Subregion
<input type="checkbox"/>	Afar	0	0	ab (0.0%)	ace (0.0%)	ady (0.0%)	af (0.0%)	ak (0.0%)	0	0	0	Africa	Sub-Sahar
<input type="checkbox"/>	Abkhaz	143	12.59	ru (0.0%)	ka (0.1%)	en (0.0%)	cs (0.0%)	hy (0.0%)	0	0	1857	Asia	Western As
<input type="checkbox"/>	Acehnese	4689	12.33	sv (0.1%)	ceb (0.0%)	nl (0.1%)	id (0.4%)	ms (0.4%)	0	0	10648	Asia	South-east
<input type="checkbox"/>	Adyghe	32	3.12	ru (0.0%)	en (0.0%)	es (0.0%)	uk (0.0%)	pl (0.0%)	0.1	0	1353	Europe	Eastern Eui
<input type="checkbox"/>	Afrikaans	12510	38.36	en (0.1%)	de (0.1%)	fr (0.1%)	nso (29.8%	nl (0.1%)	1	0.1	52839	Africa	Sub-Sahar
<input type="checkbox"/>	Akan langu	92	0	en (0.0%)	de (0.0%)	sv (0.0%)	fr (0.0%)	pl (0.0%)	0.1	0	1346	Africa	Sub-Sahar
<input type="checkbox"/>	Alemannic	10420	7.57	de (0.4%)	fr (0.4%)	en (0.1%)	it (0.5%)	nl (0.4%)	1.2	0.4	197688	Europe	Western Eu
<input type="checkbox"/>	Amharic	395	46.58	en (0.0%)	fr (0.0%)	sv (0.0%)	ceb (0.0%)	es (0.0%)	0	0	2292	Africa	Sub-Sahar
<input type="checkbox"/>	Aragonese	4845	24.81	es (0.2%)	ca (0.3%)	en (0.0%)	fr (0.1%)	eu (0.5%)	0.3	0.1	46866	Europe	Southern E
<input type="checkbox"/>	Old English	262	0.38	en (0.0%)	pl (0.0%)	fr (0.0%)	nl (0.0%)	sv (0.0%)	0.2	0	12894	Europe	Northern Ei
<input type="checkbox"/>	Arabic	240034	57.33	en (1.5%)	fr (2.6%)	fa (6.5%)	de (1.7%)	ru (2.4%)	9.5	2.7	1268514	Asia	Western As

# How well do language editions cover other languages' CCC? (Coverage)

## Wikipedia Language Editions' CCC Coverage

This page shows some statistics that explain how well each Wikipedia language edition covers the [Cultural Context Content \(CCC\)](#) articles from the other language editions.

The following table is particularly useful in order to understand the content culture gap between language editions, that is the imbalances across languages editions in content representing each language cultural context. Specifically, it shows how well each language edition covers the other language editions CCC by counting the CCC covered articles, i.e. articles from other language CCC that exist in one particular language edition.

Languages are sorted in alphabetic order by their Wikicode, and the columns present the following statistics: the number of articles in the Wikipedia language edition (**Articles**), the **first five other languages CCC** in terms of most articles covered and the percentage of coverage computed according to the total number of CCC articles of those language edition, the relative coverage (**R. Coverage**) of all languages CCC computed as the average of each language edition CCC percentage of coverage, the total coverage (**T. Coverage**) of all languages CCC computed as the percentage of coverage of all the articles that belong to other language editions CCC, and the total number of articles covered (**Covered Art.**) that belong other language editions CCC.

<input type="checkbox"/>	Language	Articles	No CCC IW	n°1	n°2	n°3	n°4	n°5	R.Coverage	T.Coverage	Covered Art	Region	Subregion
<input type="checkbox"/>	Afar	1	335	en (0.0%)	nl (0.0%)	simple (0.0)	ab (0.0%)	ace (0.0%)	0	0	4	Africa	
<input type="checkbox"/>	Abkhaz	3471	125.9	ar (0.2%)	ru (0.1%)	ka (1.2%)	da (0.2%)	sv (0.0%)	0.7	0	3425	Asia	
<input type="checkbox"/>	Acehnese	7478	83.3	id (1.8%)	ms (2.2%)	en (0.0%)	simple (0.9)	ar (0.1%)	1.9	0.1	8963	Asia	
<input type="checkbox"/>	Adyghe	544	201.7	ru (0.0%)	ar (0.0%)	en (0.0%)	simple (0.1)	fr (0.0%)	0.9	0	1243	Europe	
<input type="checkbox"/>	Afrikaans	52138	51.4	en (0.4%)	ar (2.2%)	simple (6.9)	de (0.4%)	fr (0.5%)	8.4	0.6	57789	Africa	
<input type="checkbox"/>	Akan langu	617	97.2	fr (0.0%)	en (0.0%)	simple (0.1)	ar (0.0%)	de (0.0%)	0.7	0	1058	Africa	
<input type="checkbox"/>	Alemannic	24820	69	de (1.3%)	fr (0.9%)	ar (0.8%)	en (0.1%)	lb (8.2%)	3.3	0.4	34396	Europe	
<input type="checkbox"/>	Amharic	14815	104.7	ar (0.6%)	en (0.0%)	simple (1.7)	fr (0.1%)	es (0.1%)	3.3	0.1	11699	Africa	
<input type="checkbox"/>	Aragonese	33081	44.8	es (2.7%)	fr (0.9%)	en (0.1%)	oc (15.4%)	ca (2.3%)	3.8	0.5	47066	Europe	
<input type="checkbox"/>	Old English	3058	100.8	en (0.0%)	simple (1.3)	ar (0.2%)	es (0.1%)	de (0.0%)	1.5	0.1	6182	Europe	
<input type="checkbox"/>	Arabic	588607	19.2	en (4.3%)	fr (4.5%)	fa (17.2%)	es (5.2%)	simple (43.	17.4	4	382487	Asia	

## Conclusions

We possibly cannot bridge all the gaps... but we can focus on the Top CCC articles lists.

- You can dig so much into what is valuable to your context.
- TV Shows, Monuments, Political figures...?
- ...



**The big Wikipedias** should aim at covering the **minimum of each others' cultures**. I am more concerned about the Top CCC articles gap than the entire Culture Gap.

**The small Wikipedias** should aim at **creating articles that might fill the lists of Top CCC articles**. This is the first group of articles the world should care about.



- **The big Wikipedias**

**Create 100 articles for each language edition CCC.**

We should reach this minimum amount.

Few results not represented (yet):

- **CCC % in terms of pageviews is always bigger than CCC % articles.**
- **The Top 500 CCC articles in terms of pageviews are 20-50% of the pageviews in CCC.**

- **The small Wikipedias**

**Create 100 articles for each CCC segment so they appear in the lists**

By segment it means: geolocated articles, keywords on title (demonym and territory name), women and men.

**Represent your own culture, at least the most important parts.**

**Let's create a virtuous circle!**

I am going to give you a **Control Panel**



# I am going to give you three Control Panels

## Countries Top CCC articles coverage by Catalan Wikipedia

This page shows some statistics of language edition (when it is a region)

Some languages are mapped to order to create lists for countries selection process, and later, they speaking territories), whose territories

These lists are created by ranking plain CCC or geolocated articles from CCC, list of CCC articles with featured article distinction from CCC, list of CCC articles with featured article distinction categorized in Wikidata as men with last year and with most edits **Last Y.**

The following table is useful to assess country level. Countries are sorted in alphabetic order by their Wikicode. **Sum Covered Articles** present the percentage of articles covered by the language. The last column, **Lists Coverage Idx.** at

The challenge is to reach 100 articles

### Country

Afghanistan (Persian CCC)	
<a href="#">Afghanistan (Lahnda languages CCC)</a>	
<a href="#">Afghanistan (Pashto CCC)</a>	
Afghanistan (Uzbek CCC)	

## Languages Top 100 CCC articles lists coverage by Catalan Wikipedia

This page shows some statistics that

These lists are created by ranking the plain CCC or geolocated articles) or articles with featured article distinction from CCC, list of CCC articles with featured article distinction categorized in Wikidata as men with last year and with most edits **Last Y.**

The following table is useful to assess sorted in alphabetic order by their Wikicode. **Sum Covered Articles** present the percentage of articles covered by the language.

The challenge is to reach 100 articles

Language	Wiki	Editors	Featured	Geolocated	Keywords	Women	Men	First 3Y	Last Y	Page views	Talk Edits	Sum Spread Articles	World Subregion
Abkhaz	ab	7%											
Acehnese	ace	13%											
Adyghe	ady	3/12											
Afrikaans	af	79%											
Akan language	ak	11/5											
Alemannic	als	83%											
Amparic	am	22%											
Arabic	an	96%											
Old English	ang	78%											

## Catalan Wikipedia Top 100 CCC article lists spread across the rest of Wikipedias

This page shows some statistics that explain how well the first each Catalan Wikipedia Top 100 CCC articles (only the first 100) are spread across the other language editions.

These lists are created by ranking the articles according to specific features and sometimes giving them weights. These different features are usually based on the content type (e.g. plain CCC or geolocated articles) or article characteristics (number of Bytes). The Top CCC articles lists are: list of CCC articles with most number of editors (**Editors**), list of CCC articles with featured article distinction (**Featured**), most bytes and references (weights: 0.8, 0.1 and 0.1 respectively), list of CCC articles with geolocation with most links coming from CCC, list of CCC articles with keywords on title with most bytes (**Bytes**), list of CCC articles categorized in Wikidata as women with most edits (**Women**), list of CCC articles categorized in Wikidata as men with most edits (**Men**), list of CCC articles created during the first three years and with most edits (**First 3Y.**), list of CCC articles created during the last year and with most edits (**Last Y.**), list of CCC articles with most pageviews during the last month (**Pageviews**), list of CCC articles with most edits in talk pages (**Discussions**).

The following table is useful in order to assess how known the Top 100 CCC articles from Catalan Wikipedia language are in the entire Wikipedia project. Languages are sorted in alphabetic order by their Wikicode, and columns present the number of articles from each list covered by the language. The last column **Sum Spread Articles** the overall sum of articles from the Top 100 of each list spread across a specific language.

The challenge is to reach 100 articles spread (Sum Spread Articles) across each other Wikipedia language edition!

Language	Wiki	Editors	Featured	Geolocated	Keywords	Women	Men	First 3Y	Last Y	Page views	Talk Edits	Sum Spread Articles	World Subregion
Afar	aa	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0	Sub-Saharan Africa
Abkhaz	ab	4%	1%	1%	1%	0%	1%	4%	0%	0%	4%	4	Western Asia
Acehnese	ace	3%	2%	1%	1%	0%	0%	2%	0%	0%	2%	4	South-eastern Asia
Adyghe	ady	2%	1%	1%	1%	0%	0%	2%	0%	0%	2%	2	Eastern Europe
Afrikaans	af	22%	7%	9%	6%	2%	12%	17%	0%	5%	15%	38	Sub-Saharan Africa
Akan language	ak	1%	0%	0%	0%	0%	0%	1%	0%	0%	1%	1	Sub-Saharan Africa
Alemannic	als	16%	4%	6%	4%	0%	11%	13%	0%	4%	11%	24	Western Europe
Amparic	am	18%	4%	10%	6%	0%	3%	18%	0%	2%	11%	23	Sub-Saharan Africa
Arabic	an	71%	18%	79%	12%	10%	36%	79%	0%	13%	44%	196	Southern Europe
Old English	ang	6%	3%	2%	2%	0%	3%	7%	0%	0%	5%	9	Northern Europe

# How do languages cover each others Top CCC articles?

These are panels to obtain a general view on the coverage and spread of the Top CCC.

- **Languages Top CCC articles coverage**

[https://wcdo.wmflabs.org/languages\\_top\\_ccc\\_articles\\_coverage/?lang=ca](https://wcdo.wmflabs.org/languages_top_ccc_articles_coverage/?lang=ca)

- **Countries Top CCC articles coverage**

[https://wcdo.wmflabs.org/countries\\_top\\_ccc\\_articles\\_coverage/?lang=ca](https://wcdo.wmflabs.org/countries_top_ccc_articles_coverage/?lang=ca)

- **Languages Top CCC articles spread**

[https://wcdo.wmflabs.org/languages\\_top\\_ccc\\_articles\\_spread/?lang=ca](https://wcdo.wmflabs.org/languages_top_ccc_articles_spread/?lang=ca)

Lang = wikicode

## Future steps

There are several possibilities for improving WCDO mainly from the data analysis and visualization point of view. The project has just set its base and now it would be time to think of further implementations:

- Pageviews (what is the extent of pageviews in CCC in every language?).
- Topical coverage (what topics are in each CCC, religion, sport, history...?).
- Temporal tracking of the gaps (are we doing better this year than the previous?).
- Past month (are we dedicating efforts to languages that require it?).
- Maps and reports for geolocated articles coverage (where are the gaps?).
- Editors (who creates CCC, anonymous or admins?).
- Monthly Newsletter with popular CCC articles.

Please, answer in the survey (at the end of the page): <https://wcdo.wmflabs.org>

## Closing: WCDO Goals

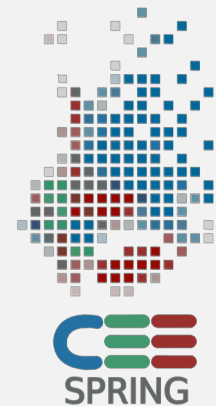
- Every Wikipedia language community is **aware** and knows about the knowledge inequalities in the entire Wikipedia project.
- Every Wikipedia language edition ensures a **minimal coverage** of each other language cultural and geographical content.
- Every Wikimedia **event includes sections and activities** dedicated to mitigate the cultural knowledge gaps and derived knowledge inequalities.
- Every Wikimedian is able to understand better **her own cultural context** and its representation on her home language edition.
- Every Wikimedian values the **importance of representing her own culture** so the rest of language editions users can import and learn from it.



# Get Involved

## Some existing projects that may benefit from WCDO:

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



**Project dissemination is always welcome!**

**There is nothing more Wikimedian than multiculturalism.**

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

# Wikipedia Cultural Diversity Observatory (WCDO)



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

**Marc Miquel, PhD**

{[marcmiquel@gmail.com](mailto:marcmiquel@gmail.com)}

Username:marcmiquel

Pompeu Fabra University, Barcelona, **Catalonia**

Amical Wikimedia (Catalan Wikipedia)

October 13th 2018 **Lviv, Ukraine**



# Thank you very much!

Marc Miquel, PhD

{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)

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:

