

Languages Matter to Cultural Diversity:

Finding Missing Languages and Bridging the Gaps in Minority Languages

Marc Miquel, PhD

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, Catalonia

Amical Wikimedia (Catalan Wikipedia)

Wikimedia Foundation – Project Grantee



Celtic Knot
**WIKIMEDIA LANGUAGE
CONFERENCE 2019**

4 & 5 July 2019



Wikipedia Cultural Diversity Observatory

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

Marc Miquel, PhD

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, Catalonia

Amical Wikimedia (Catalan Wikipedia)

Wikimedia Diversity WG 2030



Celtic Knot
**WIKIMEDIA LANGUAGE
CONFERENCE 2019**

4 & 5 July 2019



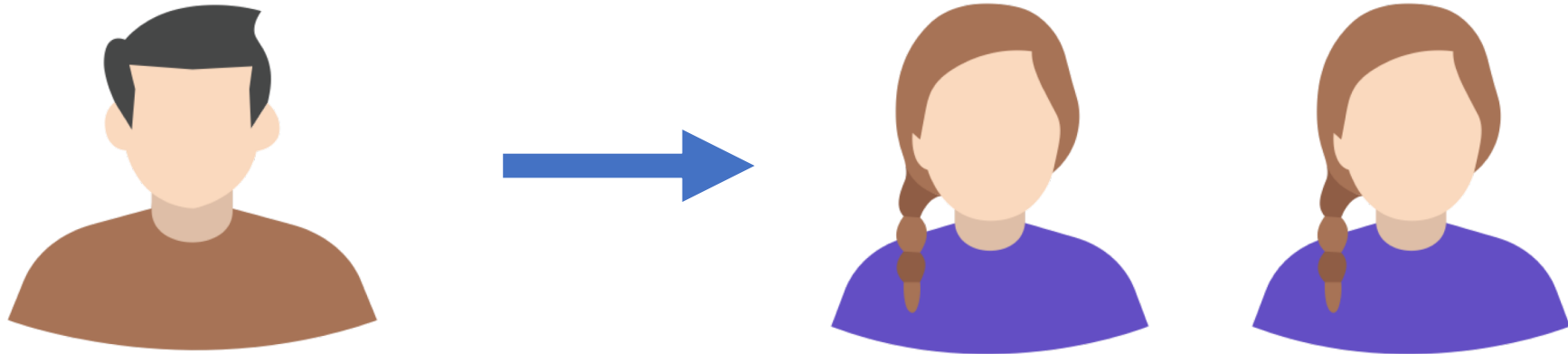
The Problem

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



What can we do?

We cannot solve cultural diversity like we solve the content gender gap, creating two women for every man.



Proposed Solution

Wikipedia Cultural Diversity Observatory (WCDO).

“a joint space for **researchers, developers and activists** to study and **fight against the knowledge gaps** and increase cultural diversity in contents”.

Its work lines are:

- **Discourse**
- **Awareness (metrics and visualizations)**
- **Organization (events and tools)**
- **Strategy (goals and priorities)**

<http://wcd0.wmflabs.org>



Why is Wikipedia failing at gathering the human cultural diversity?

- Some languages, contexts and their concepts are not in Wikipedia. (**Representation**)
- Some are in Wikipedia but remain exclusive to some language editions. (**Sharing**)

				ꣳ					
			Ვ	Ლ		Ი	Კ		
Რ	Ს	Ტ	Უ	Ფ	Ქ	Ღ	Ყ	Შ	Ჩ
B	Წ	Ჭ	Ხ	Ჯ	Ჰ	Ჱ	Ჲ	Ჳ	Ჴ
Ჵ	Ჶ	Ჷ	Ჸ	Ჹ	Ჺ	᲻	᲼	Ჽ	Ჾ
Ჿ	᳀	᳁	᳂	᳃	᳄	᳅	᳆	᳇	᳈
				᳉	᳊	᳋			
				᳌	᳍	᳎			

→ Missing pieces

We can work on sharing.

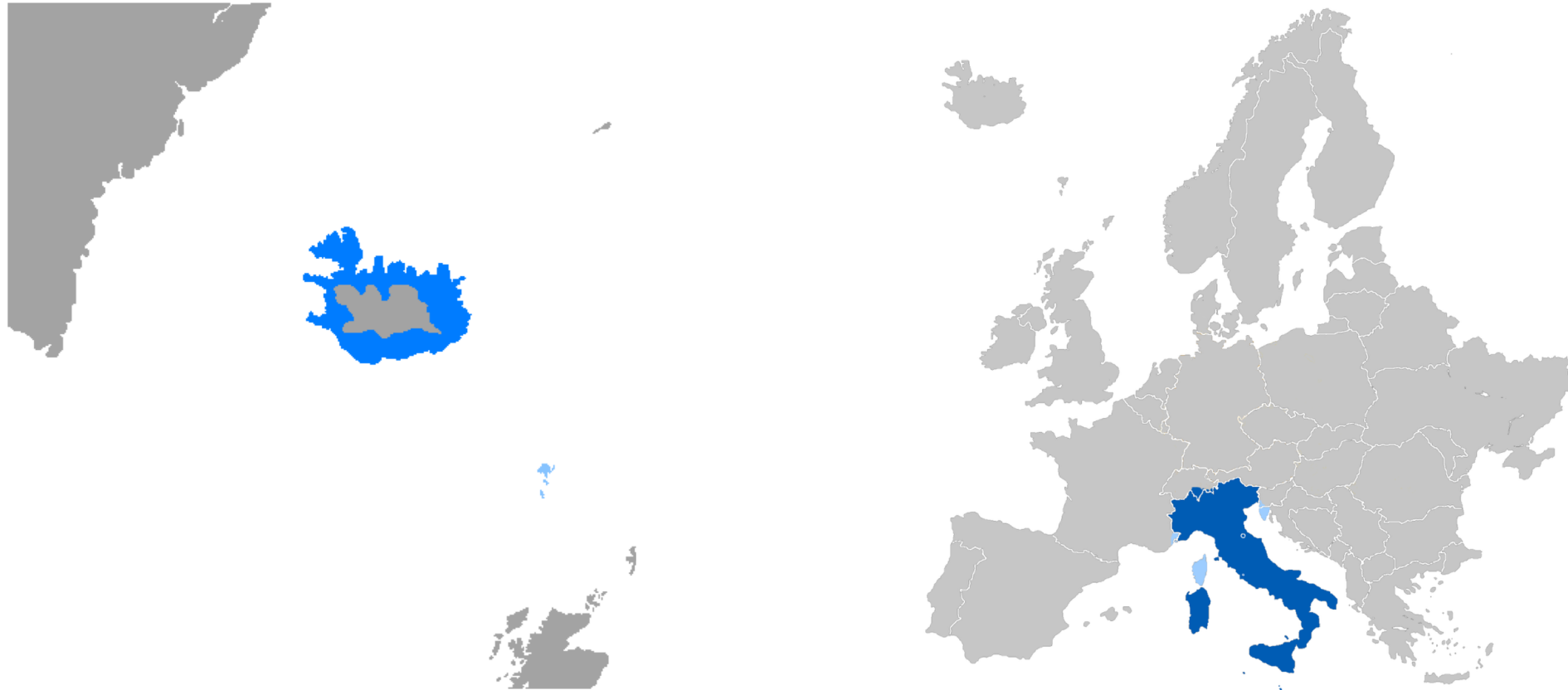
Concepts from one group of people/context **represented** in their native Wikipedia language edition but **not shared** with other languages.



Cartography



For each Wikipedia language edition, we aim at selecting the **Cultural Context Content (CCC)**, i.e. traditions, language, politics, agriculture, biographies, places, events, etcetera, related to the territories where the language is spoken.



Icelandic Cultural Context only relates to concepts from Iceland.

Italian Cultural Context includes articles about everything related to Italy, San Marino, Vaticano, Canton Ticino, Istria among others.

Method to collect Cultural Context Content

We created a method (Miquel-Ribé, 2017; Miquel-Ribé & Laniado, 2019) that 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'eyih B. Semenawi K'eyih Bahri	Q27910	Afar	aa				ER	ER-SK	yes	Eritrea	yes				
10	Abkhazia	Аҧсны	Q23334	Abkhaz	ab	Abkhaz		GE	GE-AB	yes	Georgia	yes		2	regional	1
11	Aceh	Acèh	Q1823	Aceh	ace			ID	ID-AC	yes	Indonesia	yes		6	no	0
12	Sumatera Utara	Sumatra Baròh	Q2140	Aceh	ace			ID	ID-SU	yes	Indonesia	yes		6	no	0
13	Republic of Adyghe	Адыгэ	Q3734	Adyghe	ady			RU	RU-AD	yes	Russian Federation	yes		2	regional	1
14	Krasnodar Krai	Краснодар край	Q3680	Adyghe	ady			RU	RU-KDA	yes	Russian Federation	yes		2	regional	1
15	Karachay-Cherkessia	Къарачае-Черкес	Q5328	Adyghe	ady			RU	RU-KC	yes	Russian Federation	yes		2	regional	1
16	South Africa	Suid-Afrika	Q258	Afrikaans	af	South Afri	Suid-Afrika	ZA		no	South Africa	yes		1	national	1
17	Central	Sentraal distrik	Q57525	Afrikaans	af			BW	BW-CE	yes	Botswana	yes		5	no	1
18	Ghanzi	Bhanzi	Q57571	Afrikaans	af			BW	BW-GH	yes	Botswana	yes		5	no	1
19	Kgalagadi	Kgalagadi	Q57581	Afrikaans	af			BW	BW-KG	yes	Botswana	yes		5	no	1
20	Kgatleng	Kgatleng	Q57593	Afrikaans	af			BW	BW-KL	yes	Botswana	yes		5	no	1
21	Southern	Suid distrik	Q57609	Afrikaans	af			BW	BW-SO	yes	Botswana	yes		5	no	1
22	Botswana	Botswana	Q963	Afrikaans	af	Motswana;Botswana	BW			no	Botswana	yes		5	no	1
23	Ghana	Ghana	Q117	Akan	ak	Ghanaian	GH			no	Ghana	yes		3	no	1
24	Switzerland	Schweiz	Q39	German, Swiss	als	Swiss	CH			no	Switzerland	yes		5	no	0
25	Vorarlberg	Vorarlberg	Q38981	German, Swiss	als			AT	AT-8	yes	Austria	yes		5	no	0
26	Champagne-Ardenne	Champagne-Ardenne	Q14103	German, Swiss	als			FR	FR-G	yes	France	yes		6	no	0
27	Lorraine	Lothringen	Q1137	German, Swiss	als			FR	FR-M	yes	France	yes		6	no	0
28	Alsace	Elsass	Q1142	German, Swiss	als			FR	FR-A	yes	France	yes		6	no	0
29	Baden-Württemberg	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://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: geolocated, keywords, category graph and Wikidata properties.

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search Wikipedia

Times Square

From Wikipedia, the free encyclopedia

Coordinates: 40°45′28″N 73°59′09″W﻿ / ﻿﻿ / ﻿

For other uses, see [Times Square \(disambiguation\)](#).

Times Square is a major commercial intersection, tourist destination, entertainment center and neighborhood in the **Midtown Manhattan** section of **New York City** at the junction of **Broadway** and **Seventh Avenue**. It stretches from West 42nd to West 47th Streets.^[1] Brightly adorned with billboards and advertisements, Times Square is sometimes referred to as "The Crossroads of the World",^[2] "The Center of the Universe",^[3] "the heart of The Great White Way",^{[4][5][6]} and the "heart of the world".^[7] One of the world's busiest pedestrian areas,^[8] it is also the hub of the **Broadway Theater District**^[9] and a major center of the world's **entertainment industry**.^[10] Times Square is one of the world's most visited tourist attractions, drawing an estimated 50 million visitors annually.^[11] Approximately 330,000 people pass through Times Square daily,^[12] many of them tourists,^[13] while over 460,000 pedestrians walk through Times Square on its busiest days.^[7]

Formerly known as Longacre Square, Times Square was renamed in 1904 after *The New York Times* moved its headquarters to the then newly erected Times Building – now **One Times Square** – the site of the annual **New Year's Eve ball drop** which began on December 31, 1907, and continues today, attracting over a million visitors to Times Square every year.^{[14][15]}

Times Square functions as a **town square**, but is not a **square** in the geometric sense of a polygon; it is more of a bowtie shape, with two triangles emanating roughly north and south from 45th Street,^[16] where **Seventh Avenue** intersects **Broadway**. Broadway runs diagonally, crossing through the horizontal and vertical **street grid** of Manhattan laid down by the **Commissioners' Plan of 1811**, and that intersection creates the "bowtie" shape of Times Square.^[17]

The southern triangle of Times Square has no specific name,^[18] but the northern triangle is called **Father Duffy Square**. It was dedicated in 1937 to Chaplain **Francis P. Duffy** of New York City's U.S. 69th Infantry Regiment and is the site of a memorial to him, along with a statue of **George M. Cohan**,^[19] as well as the TTKTS reduced-price ticket booth run by the **Theatre Development Fund**. Since 2008, the booth has been backed by a red, sloped, triangular set of bleacher-like stairs, which is used by people to sit, talk, eat, and take photographs.

Times Square

Neighborhood in Manhattan

Broadway show billboards in Times Square, 2009 (top), 2013 (bottom)

Nickname(s): The Great White Way
The Crossroads of the World

State New York

City New York City

Borough Manhattan

Boundaries Broadway, 7th Avenue, 42nd and 47th Streets

Subway services 1, 2, 3, 7, <7>, A, C, E, N, Q, R, W, and S trains at Times Square–42nd Street station

Bus routes M7, M20, M42, M50, M104

Historical features Duffy Square
George Michael Cohan statue
One Times Square

Contents [hide]

- History
 - Early history
 - 1900s–1930s
 - 1930s–1950s
 - 1960s–1980s

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search Wikipedia

English literature

From Wikipedia, the free encyclopedia

This article is focused on **English-language literature** rather than the literature of **England**, so that it includes writers from **Scotland**, **Wales**, and the whole of **Ireland**, as well as literature in English from countries of the former **British Empire**, including the **United States**. However, until the early 19th century, it only deals with the literature of the **United Kingdom** and **Ireland**. It does not include *literature written in the other languages of Britain*.

The **English language** has developed over the course of more than 1,400 years.^[1] The earliest forms of English, a set of **Anglo-Frisian dialects** brought to Great Britain by Anglo-Saxon settlers in the fifth century, are called **Old English**. **Middle English** began in the late 11th century with the **Norman conquest of England**.^[2] **Early Modern English** began in the late 15th century with the introduction of the **printing press** to London and the **King James Bible** as well as the **Great Vowel Shift**.^[3] Through the influence of the **British Empire**, the English language has spread around the world since the 17th century.

Contents [hide]

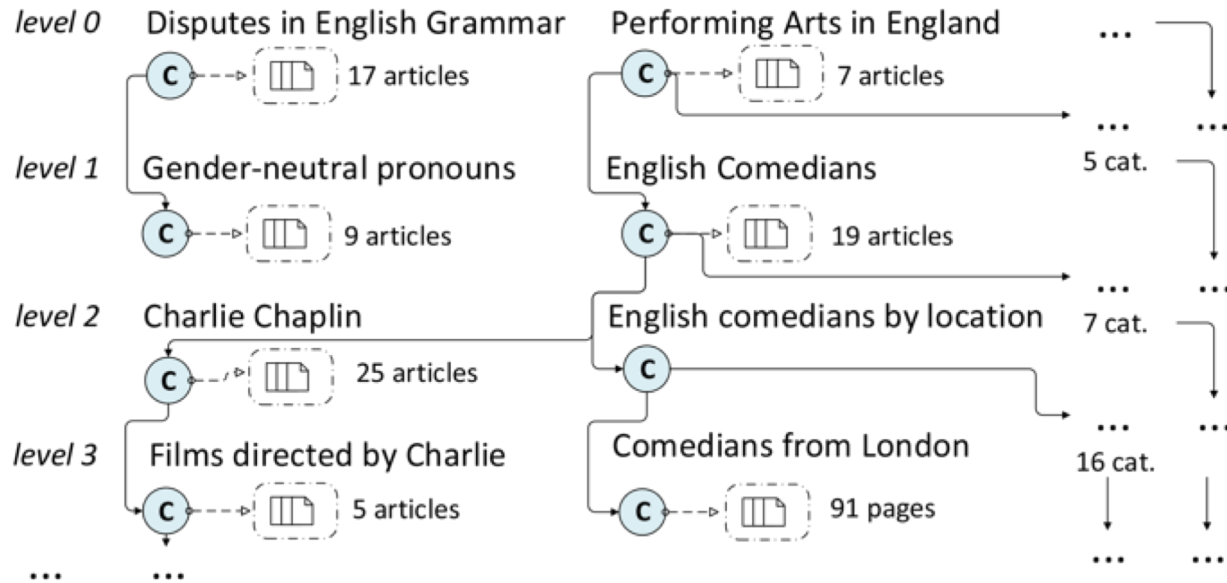
- Old English literature: c. 450–1066
- Middle English literature: 1066–1500
 - Medieval theatre
- English Renaissance: 1500–1660
 - Elizabethan period (1558–1603)
 - Poetry
 - Drama
 - Jacobean period: 1603–25
 - Poetry
 - Prose
 - Late Renaissance: 1625–1660
 - Poetry
- Restoration Age: 1660–1700
 - Poetry
 - Prose
 - Drama
- 18th century
 - Augustan literature (1700–1750)
 - Poetry
 - Drama

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.

• **Geolocation in one of the territories**

• **Keyword (demonym/territory name) on title**

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



Category crawling using keywords

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

WIKIPEDIA
The Free Encyclopedia

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search Wikipedia

Dylan Moran

From Wikipedia, the free encyclopedia

Dylan William Moran (/ˈmɔːrən/; born 3 November 1971)^[1] 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.^{[1][2][3][4]} He attended St. Patrick's Classical School, where he experimented early on with

Dylan Moran

Photo taken April 2006

Born 3 November 1971 (age 46)
Navan, County Meath, Ireland

Medium Stand-up, film, television

Nationality Irish

Years active 1992–present

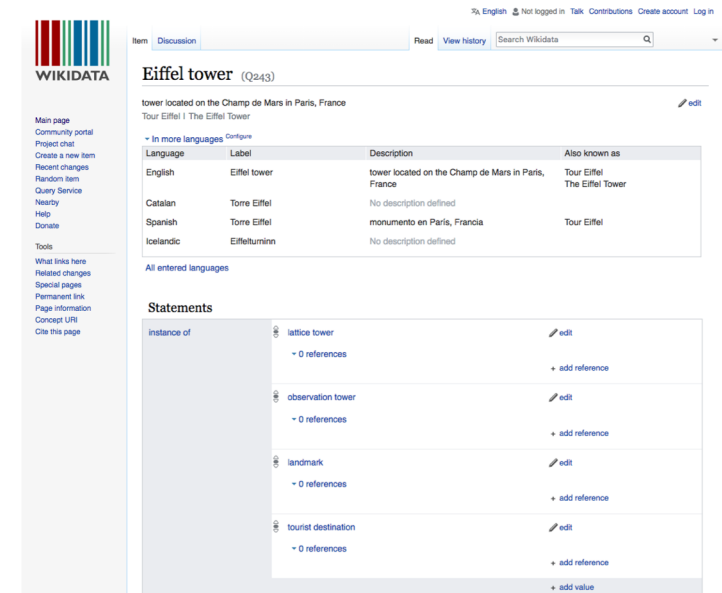
Genres Observational comedy, deadpan, satire, surreal humour

Website www.dylanmoran.com

Youngest winner of Perrier Comedy Award (1996)

Some Wikidata

- Location properties (location, located in administrative,...).
- Country properties (country of citizenship, of origin).
- Language properties (official language, native language...).
- 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 the Eiffel tower (Q243). The page is in English and shows the item's description: "tower located on the Champ de Mars in Paris, France". The page also shows a table of labels and descriptions 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	Eiffeltúrnin	No description defined	

Statements:

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

Link features:

- 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.

Machine Learning Classifier

We have a database with all the articles and features related to the territories.

We introduce it to a Random Forest classifier to obtain the final CCC dataset for each language edition.

The manual assessment (blind) determined a 5%-5% false positive and false negatives.

Datasets



← → ↻ 🔒 https://wcdo.wmflabs.org/datasets/ 🔍 ☆ 🔔 🏠

Index of /datasets/

../		
2018-09/	04-Sep-2018 14:01	-
latest/	04-Sep-2018 14:01	-

Download at:

<https://wcdo.wmflabs.org/datasets/>

https://figshare.com/articles/Cultural_Context_Content_CCC_Datasets/

Cartography



For each Wikipedia language edition, we aim at selecting the **Cultural Context Content (CCC)**, i.e. traditions, language, politics, agriculture, biographies, places, events, etcetera, related to the territories where the language is spoken.

Awareness (metrics and visualizations): extent of CCC

Taking into account the largest Wikipedia language editions, CCC is in average about a quarter of each Wikipedia (Miquel-Ribé and Laniado 2016).

CCC articles tend to be more developed in number of Bytes, images, and categories (Miquel-Ribé, 2017). Every language should represent their context properly (extent). This is healthy.

We have a **problem of representation** considering that the CCC extent in non-western languages (African and Asian) is on average much smaller (Miquel-Ribé and Laniado, 2019).



Awareness (metrics and visualizations): gap between languages

About a 60% of the content language gaps are due to CCC (Miquel-Ribé and Laniado 2018). **Culture gap.**

Big languages like English or geographically close languages are the ones covering best the smaller languages (Miquel-Ribé and Laniado 2016).

We have a **problem of sharing** considering a Wikipedia language edition cultural diversity as the coverage of all the others' CCC.



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

In the Cultural Diversity Observatory we propose every Wikipedia has 100 articles about every other language's cultural and geographical content.

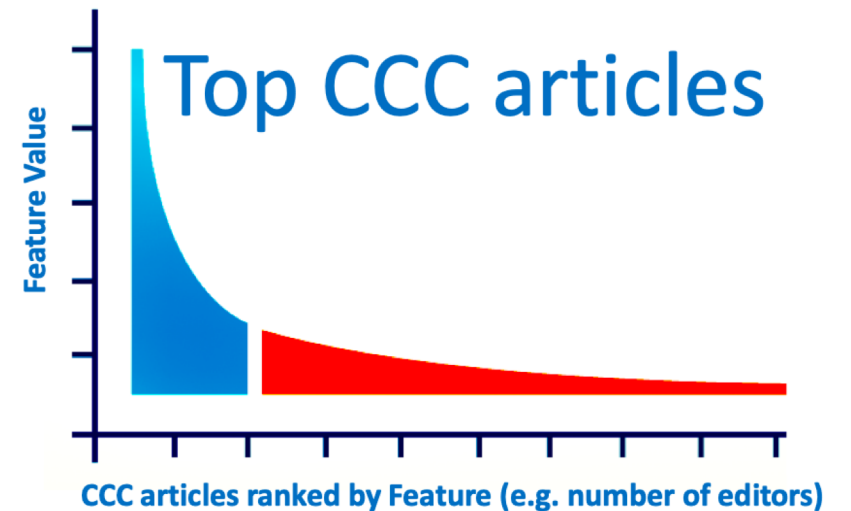
Every language should have this minimum. This is cultural diversity.

28-30 thousand articles to cover a minimum of Wikipedia cultural diversity.

Organization (events and tools): Top CCC articles lists

From each language, those articles from their cultural context which are the **most 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://wcd0.wmflabs.org/top_ccc_articles/?list=men&lang_origin=pl&lang_target=uk

http://wcd0.wmflabs.org/top_ccc_articles

Top 500 CCC articles list "Men" from Cornish CCC in Italian Wikipedia

Select the parameters

[Download Table \(Excel\)](#)

List: Language origin: Country origin: Language target:

N°	Cornish Title	Edits	Editors	Pageviews	Bytes	References	Wikidata Properties	Interwiki Links	Inlinks from CCC	Creation Date	Other Languages	Italian Title
1	Henry Jenner	37	15	8	2.4k	4	44	12	7	2004-09-06	en	Henry Jenner
2	Robert Morton Nance	30	15	1	6.3k	1	25	7	6	2004-09-03	en	Robert Morton Nance
3	Nicholas Williams	27	15	1	0.7k	0	15	5	7	2004-09-06	en	Nicholas Williams (linguista)
4	Rod Lyon	25	11	0	0.7k	0	12	2	4	2004-09-02	en	Rod Lyon
5	Ken George	24	14	0	0.7k	0	7	7	5	2004-09-07	en , es	Ken George
6	A.S.D. Smith	21	12	0	2.0k	0	14	3	5	2004-09-13	en	A.S.D. Smith
7	Andrew George	21	9	1	1.3k	2	28	6	3	2004-09-09	en	Andrew George
8	Meryasek	18	11	0	1.7k	0	11	7	0	2004-12-05	en , fr	Meriasek
9	Jowan Bolitho	17	6	0	1.3k	2	5	1	1	2005-01-16	en	
10	E.G. Retallack Hooper	16	10	0	0.7k	0	6	1	2	2005-01-16	en	
11	Myhal an Gof	15	10	0	0.9k	0	12	9	0	2004-09-15	en , fr , de	
12	Robert Biscoe	15	5	0	0.7k	1	6	1	3	2004-09-28	en	
13	John Keigwin	15	7	0	3.1k	5	19	1	4	2005-04-30	en	

Organization (events and tools): Panels to understand the coverage of Top CCC

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

- **Languages Top CCC articles spread**

https://wcdo.wmflabs.org/languages_top_ccc_articles_spread/?lang=ca

Lang = wikicode

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 well Catalan Wikipedia covers the Top 100 CCC articles from the lists generated from all the other language editions CCC. Languages are sorted in alphabetic order by their Wikicode, and columns present the number of articles from each list covered by English language. The last two columns, **Lists Coverage Idx.** and **Sum Covered Articles** present the percentage of articles from the lists covered and the overall sum of articles from the lists covered by Catalan Wikipedia for each language.

The challenge is to reach 100 articles spread across each language CCC!

Language	Wiki	Editors	Featured
Afar	aa	0%	0%
Abkhaz	ab	4%	1%
Acehnese	ace	3%	2%
Adyghe	ady	2%	1%
Afrikaans	af	22%	7%
Akan language	ak	1%	0%
Alemannic	als	16%	4%
Ambrosian	am	18%	4%
Andalusi	an	71%	18%
Old English	ang	6%	3%

Languages Top 100 CCC articles lists coverage by Catalan Wikipedia

This page shows some statistics that explain how well Catalan Wikipedia language edition covers the Top 100 of the Top CCC articles lists from other Wikipedia 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 to assess how well Catalan Wikipedia covers the Top 100 CCC articles from the lists generated from all the other language editions CCC. Languages are sorted in alphabetic order by their Wikicode, and columns present the number of articles from each list covered by English language. The last two columns, **Lists Coverage Idx.** and **Sum Covered Articles** present the percentage of articles from the lists covered and the overall sum of articles from the lists covered by Catalan Wikipedia for each language.

The challenge is to reach 100 articles covered (Sum Covered Articles) from each language CCC!

Language	Wiki	Editors	Featured	Geolocated	Keywords	Women	Men	First 3Y	Last Y	Page views	Talk Edits	List Coverage Idx.	Sum Covered Articles	World Subregion
Abkhaz	ab	7%	7%	4/87	2/13	0/0	2/11	5/9	1/20	8%	6%	12.7	8	Western Asia
Acehnese	ace	13%	0%	7%	5%	0/1	0/14	9%	0%	8%	4%	4.6	15	South-eastern Asia
Adyghe	ady	3/12	3/12	0/0	3/4	0/0	0/3	1/3	0/0	3/12	3/12	20.8	3	Eastern Europe
Afrikaans	af	79%	28%	20%	12%	15%	22%	63%	4%	39%	29%	31.1	145	Sub-Saharan Africa
Akan language	ak	11/57	11/57	4/29	2/2	1/4	4/13	7/16	0/0	11/57	11/57	29.1	11	Sub-Saharan Africa
Alemannic	als	83%	28%	88%	17%	12%	50%	79%	24%	61%	51%	49.3	295	Western Europe
Ambrosian	am	22%	9%	7/98	1/25	0/0	2/8	4/5	1/1	17%	12%	27.6	23	Sub-Saharan Africa
Andalusi	an	96%	69%	43%	38%	62%	89%	90%	23/61	66%	59%	65	427	Southern Europe
Old English	ang	78%	77%	18/24	5/7	5/6	37/41	29/37	2/2	81%	78%	81.2	85	Northern Europe

Problem: smaller language editions do not even have 100 on their cultural context to fill the lists.

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.

Sharing stage

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.

Representation stage

Why is Wikipedia failing at gathering the human cultural diversity?

a) Concepts from one group of people/context **represented** in their native language but not shared to other languages.



b) Concepts from one group of people/context not represented in their native language but yes in other languages.



b1. Their native language is minoritized but has Wikipedia.

b2. Their native language is minoritized and does not have Wikipedia.

c) Concepts from one group of people/context not represented in their native language and no language at all.



What else can we work on to improve cultural diversity?

We can try to work on language and concepts representation.



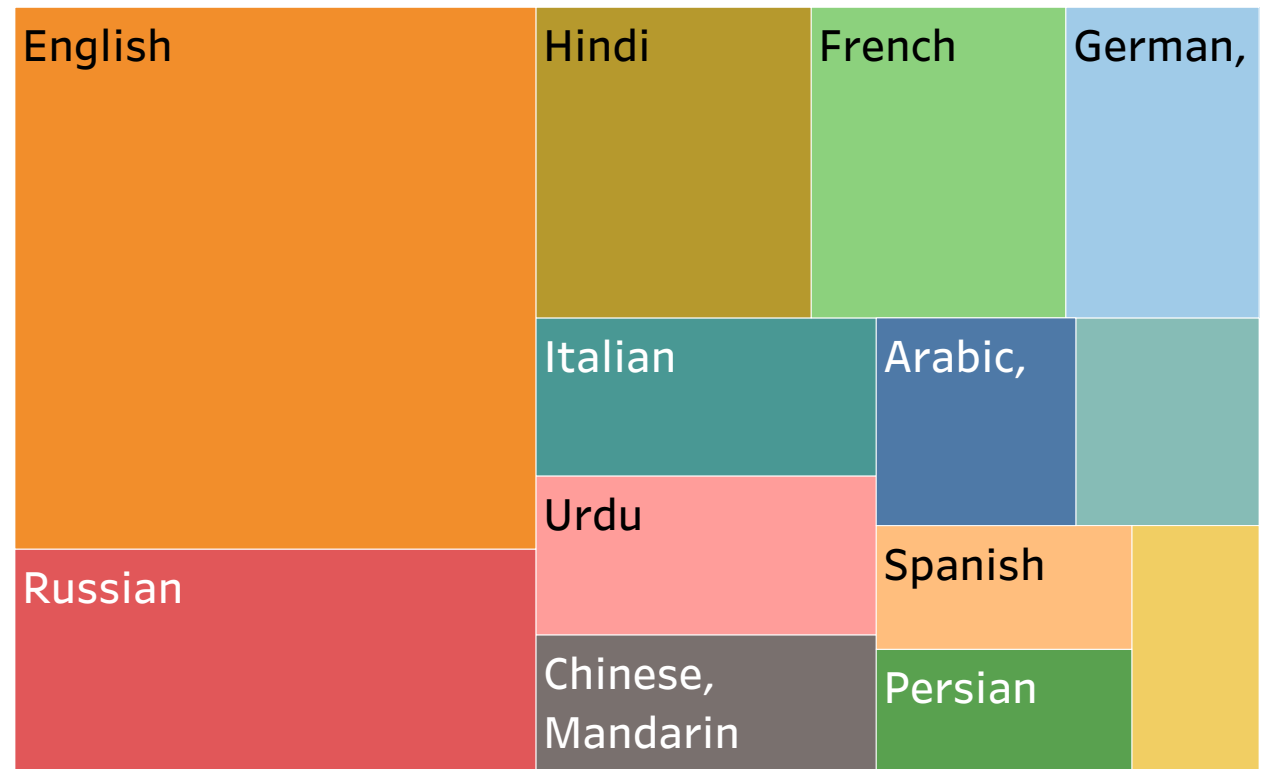
Work in progress

b I. Wikipedia Languages' Missing CCC

b I. Their native **language is minoritized** but has a Wikipedia.

253 out of the 302 Wikipedia languages are overlapped with other languages with a higher social status.

English and Russian are the languages which overlap most with other Wikipedia languages (87 and 37) because they are official at country level.



The is quite usual in African languages, whose content tends to be represented in English or French rather than in their native indigenous languages.

For instance, **Luganda Wikipedia** (from Uganda) has a very low CCC (**just 3.15%**). There is an opportunity to give a digital life to the language through Wikipedia.



- We can create lists of **missing CCC articles** in minoritized languages that exist in higher status languages and encourage them to create them.

For example: **Luganda language** coexists with **English** as a higher status language (“Status 3 wider communication” and “Status 1 national” correspondingly).

Luganda Wikipedia (1171 articles) lacks articles about:

- Geography
- Traditions
- Politicians
- Etc.

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search Wikipedia

Yoweri Museveni

From Wikipedia, the free encyclopedia

The **neutrality of this article is disputed**. Relevant discussion may be found on the **talk page**. Please do not remove this message until conditions to do so are met. *(January 2014)* [Learn how and when to remove this template message](#)

Yoweri Kaguta Museveni (ⓘ pronunciation (help·info); born 15 September 1944) is a Ugandan politician who has been **President of Uganda** since 1986. Museveni was involved in rebellions that toppled notorious Ugandan leaders **Idi Amin** (1971–79) and **Milton Obote** (1980–85) before capturing power in the 80s. In the mid to late 1990s, Museveni was celebrated by the **West** as part of a **new generation of African leaders**. During Museveni's presidency, Uganda has experienced relative peace and significant success in battling **HIV/AIDS**. At the same time, Uganda remains a country suffering from high levels of corruption, unemployment and poverty. Museveni's presidency has been marred by involvement in the **civil war in the Democratic Republic of the Congo** and other **Great Lakes region** conflicts; the rebellion in Northern Uganda by the **Lord's Resistance Army** which caused a drastic

Yoweri Museveni

Museveni in July 2012

9th President of Uganda

President of Uganda does exist

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search Wikipedia

Ruhakana Rugunda

From Wikipedia, the free encyclopedia

Ruhakana Rugunda (born 7 November 1947) is a Ugandan politician who has been **Prime Minister of Uganda** since 2014. A physician by profession, he held a long series of Cabinet posts under President **Yoweri Museveni** beginning in 1986. He served as Uganda's Minister of Foreign Affairs from 1994 to 1996 and as Minister of Internal Affairs from 2003 to 2009. Subsequently, he was Permanent Representative to the **United Nations** from 2009 to 2011 and Minister of Health from 2013 to 2014.

He was appointed as Prime Minister on 18 September 2014. He replaced **Amama Mbabazi**, who was dropped from the Cabinet.^[1]

Contents [hide]

- 1 Early career
- 2 Political career
- 3 Personal
- 4 See also
- 5 References

Rugunda at the World Trade Organization Ministerial Conference of 2015

10th Prime Minister of Uganda

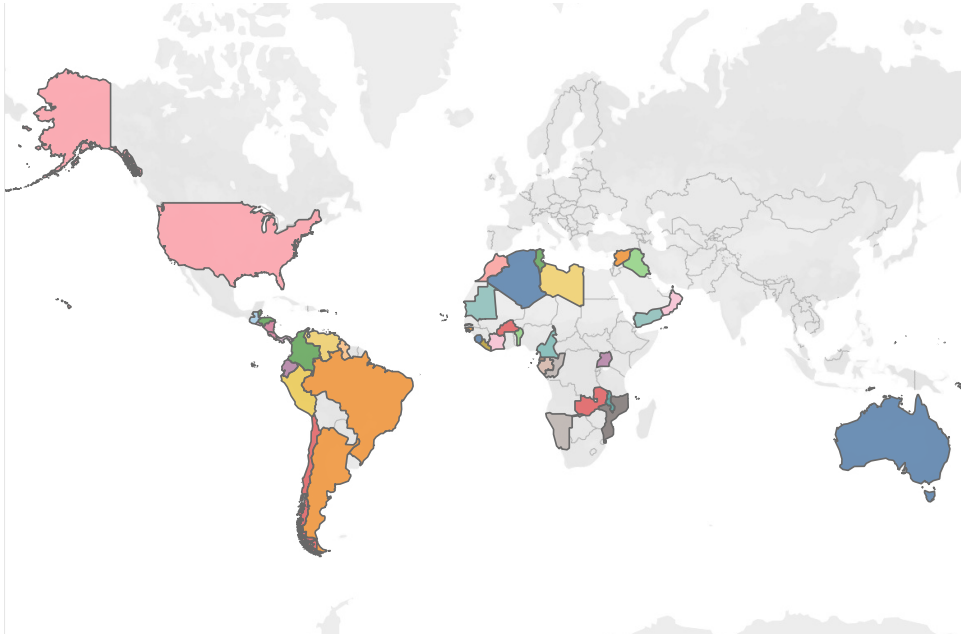
Incumbent

Assumed office
18 September 2014

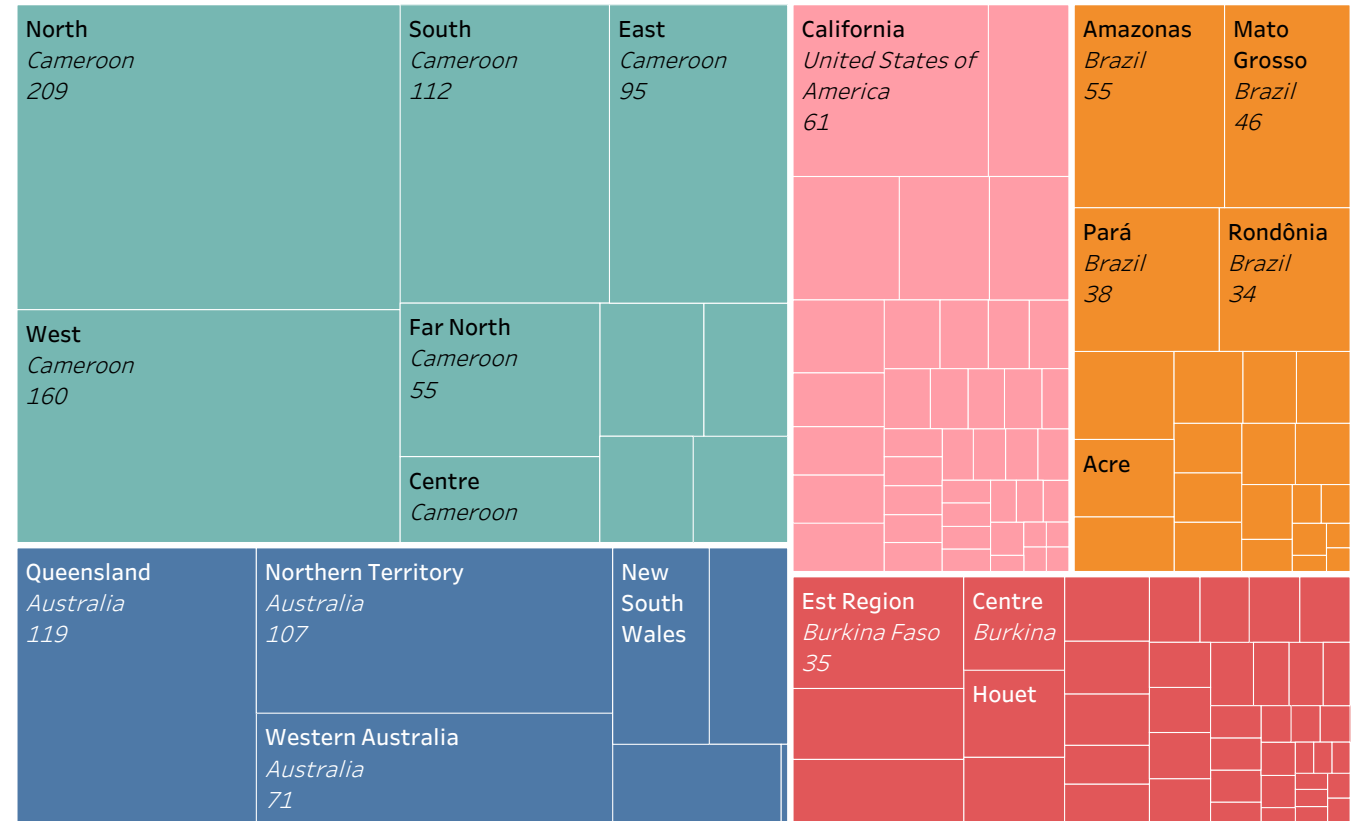
President **Yoweri Museveni**

Prime Minister of Uganda does not exist

b2. Missing territories (not covered by a Wikipedia of an indigenous language)



Countries not covered by an indigenous language with a Wikipedia.



Top five countries (colour) and subdivisions not covered by a Wikipedia of an indigenous language by the number of indigenous languages (size).

c. Missing Languages in Wikipedia

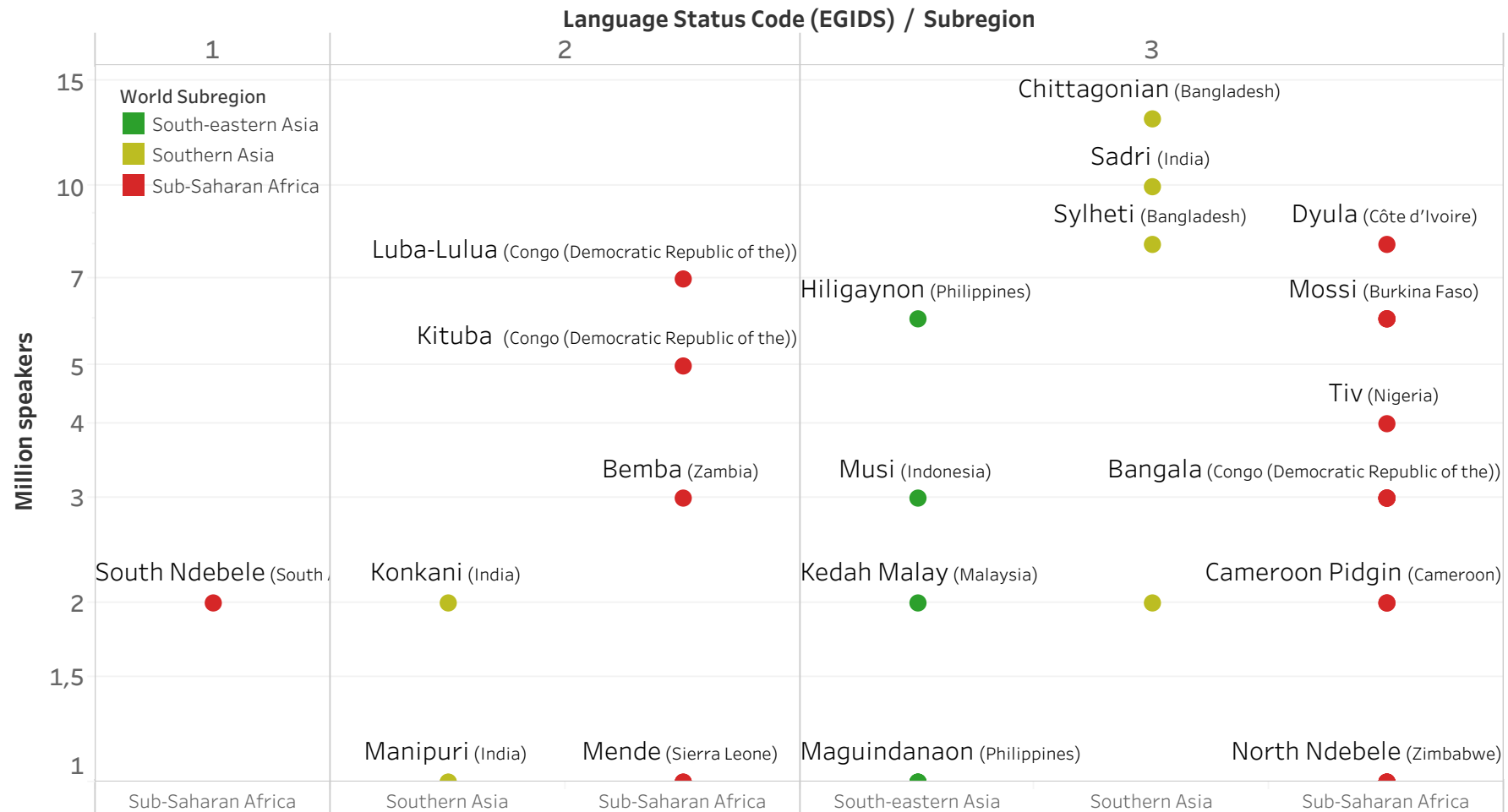
There exist approximately 7000 languages in the planet according to Ethnologue (SIL). There is an opportunity to map all the missing knowledge.

	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	nu
2	Afar	Qafar	Q193494	Afar	aa			ET	ET-AF	yes	Ethiopia	yes	2	regional	0
3	Somali	Somali	Q202800	Afar	aa			ET	ET-SO	yes	Ethiopia	yes	2	regional	0
4	Amhara	Amhara	Q203009	Afar	aa			ET	ET-AM	yes	Ethiopia	yes	2	regional	0
5	Ali Sabieh	Ali Sabieh	Q821008	Afar	aa			DJ	DJ-AS	yes	Djibouti	yes	5	no	0
6	Arta	Arta	Q705941	Afar	aa			DJ	DJ-AR	yes	Djibouti	yes	5	no	0
7	Obock	Obock	Q844929	Afar	aa			DJ	DJ-OB	yes	Djibouti	yes	5	no	0
8	Dikhil	Dikhil	Q283979	Afar	aa			DJ	DJ-DI	yes	Djibouti	yes	5	no	0
9	Debubawi K'eyih	Debubawi K'eyih	Q27728	Afar	aa			ER	ER-DU	yes	Eritrea	yes	5	no	0
10	Semenawi K'eyi B	Semenawi K'eyi Bahri	Q27910	Afar	aa			ER	ER-SK	yes	Eritrea	yes			
11	Abkhazia	Аԥсны	Q23334	Abkhaz	ab	Abkhaz		GE	GE-AB	yes	Georgia	yes	2	regional	1
12	Aceh	Аԥех	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 Adyghe	Адыгæ	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	Къарачæ-Черкæс	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

We need to detect potential languages that can easily become Wikipedias. EGIDS (language social status) sets some priority along the number of speakers.

Potential new Wikipedias

We need to detect potential languages that can easily become Wikipedias with million speakers and the language status code (EGIDS).



Wikipedia Cultural Diversity Observatory (WCDO).

“a joint space for **researchers** and **activists** to study and **fight against the knowledge gaps** and increase cultural diversity in contents”.

Its work lines are:

- Discourse
- Awareness (metrics and visualizations)
- Organization (events and tools)
- Strategy (goals and priorities)

<http://wcdо.wmflabs.org>



Help minoritized and potential languages build the required capacity.



Wikimedia Summit 2019, (Back row L-R) Rebecca, Jeffrey, Simona, Dariusz, Winifred, (Front row L-R) Liang, Oscar, Michal and Sailesh

**We should encourage them not only to build a Wikipedia,
But to create their most unique and specific knowledge about their context.**



Arctic char fishing. By Ansgar Walk.

We need to value their knowledge and difference.

Each language has some 'uniqueness' to contribute.

Wikipedia is their best strategy to pass this knowledge on.



Thank you very much!

Wikipedia Cultural Diversity Observatory (WCDO)

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

Marc Miquel

{marcmiquel@gmail.com}

Username:marcmiquel

Pompeu Fabra University, Barcelona, Catalonia

Amical Wikimedia (Catalan Wikipedia)

Wikimedia WG Diversity 2030



References (if you want to know more)

Miquel-Ribé, M., & Laniado, D. (2016). Cultural identities in wikipeidias. 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.

Miquel-Ribé, M., & Laniado, D. (2019). Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions. *Proceedings of the 13th International AAAI Conference on Web and Social Media. ICWSM*. ACM.

