

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 Foundation – Project Grantee



Wikipedia is a multilingual "free-access, free-content Internet encyclopedia"*



Russian version of the article 'VKontakte'



Catalan version of the article 'VKontakte'

The Problem

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



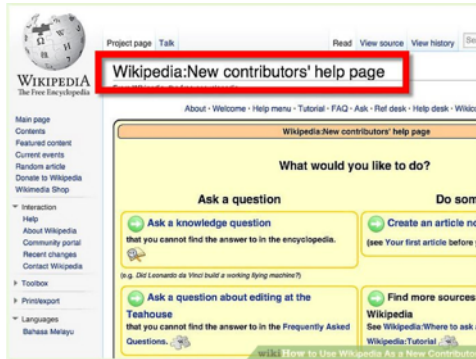
Why is Wikipedia failing at gathering the human cultural diversity?

- Some contexts and their cultural expressions are not in it. (**Representation**)
- Some are represented but remain exclusive to some languages. (**Sharing**)



➔ Missing pieces

Who creates Wikipedia?



Online Communities



Chapters (Indian, German, Italian, Catalan, Armenian, South African...)



Wikimedia Foundation teams
(Infrastructure, product, funding...)

Wikimedia Movement

Proposed Solution

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”.



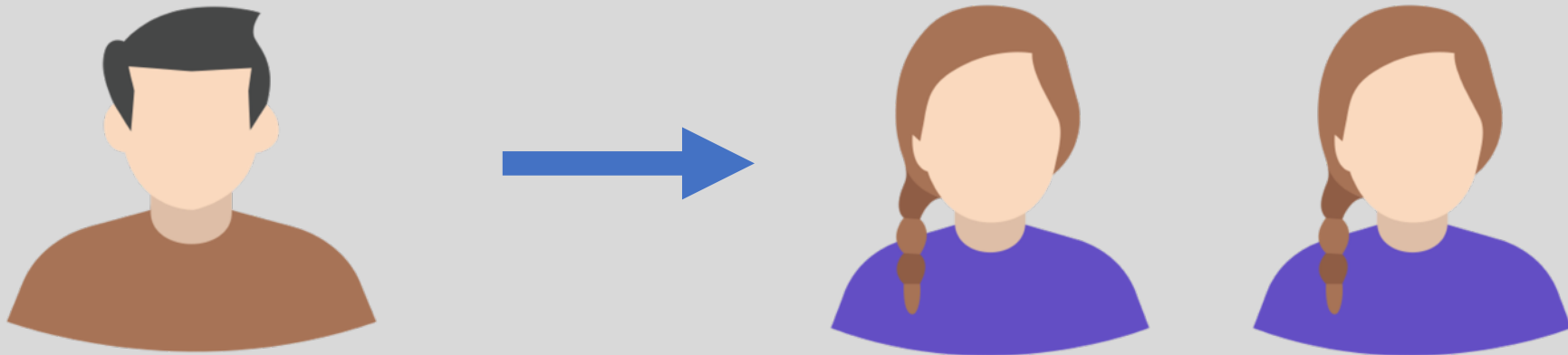
It started in January 2018. Its work lines are:

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

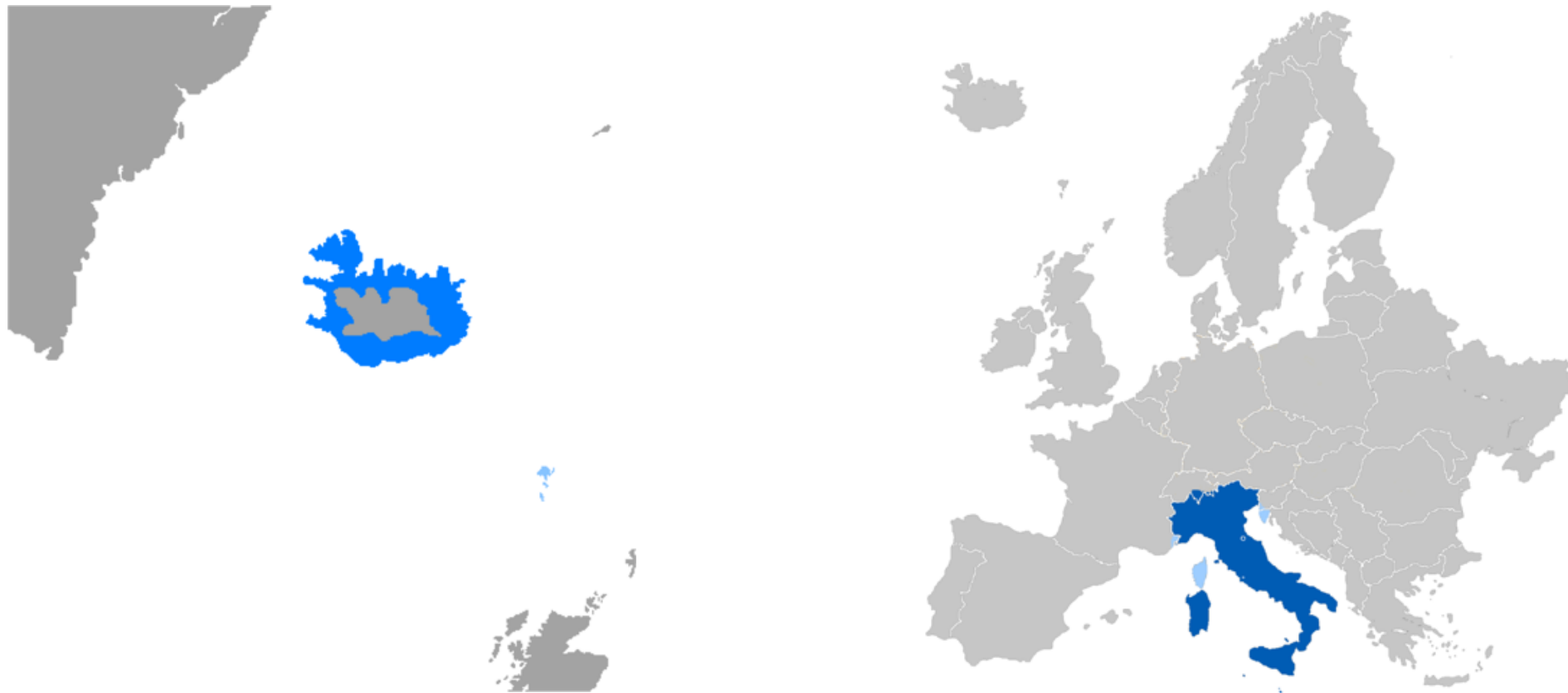
<http://wcdo.wmflabs.org>

Awareness (metrics and visualizations): studying the Wikipedia content

Why WCDO? Because to understand cultural diversity and fix it, we need a cartography and metrics to understand better the topical distribution.



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



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.



For each cultural context, we aim at selecting the **Cultural Context Content (CCC)**, i.e. traditions, language, politics, agriculture, biographies, places, events, etcetera.

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	Wik	demon	demon	ISO3166	ISO31662	region	country	ind	lan	official	nu
1	Afar	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-DIU	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	Q2823	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 Adyge	Q3794	Adyge	ady			RU	RU-AD	yes	Russian Federation	yes	2	regional	1
14	Krasnodar Krai	Q3680	Adyge	ady			RU	RU-KDA	yes	Russian Federation	yes	2	regional	1
15	Karachay-Cherk	Q5328	Adyge	ady			RU	RU-KC	yes	Russian Federation	yes	2	regional	1
16	South Africa	Q258	Afrikaans	af	South Afr	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		GH		no	Ghana	yes	3	no	1
24	Switzerland	Q39	German, Swiss	als	Swiss		CH		no	Switzerland	yes	5	no	0
25	Vorarlberg	Q38981	German, Swiss	als			AT	AT-B	yes	Austria	yes	5	no	0
26	Champagne-Arde	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-Württemb	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://wcdo.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.



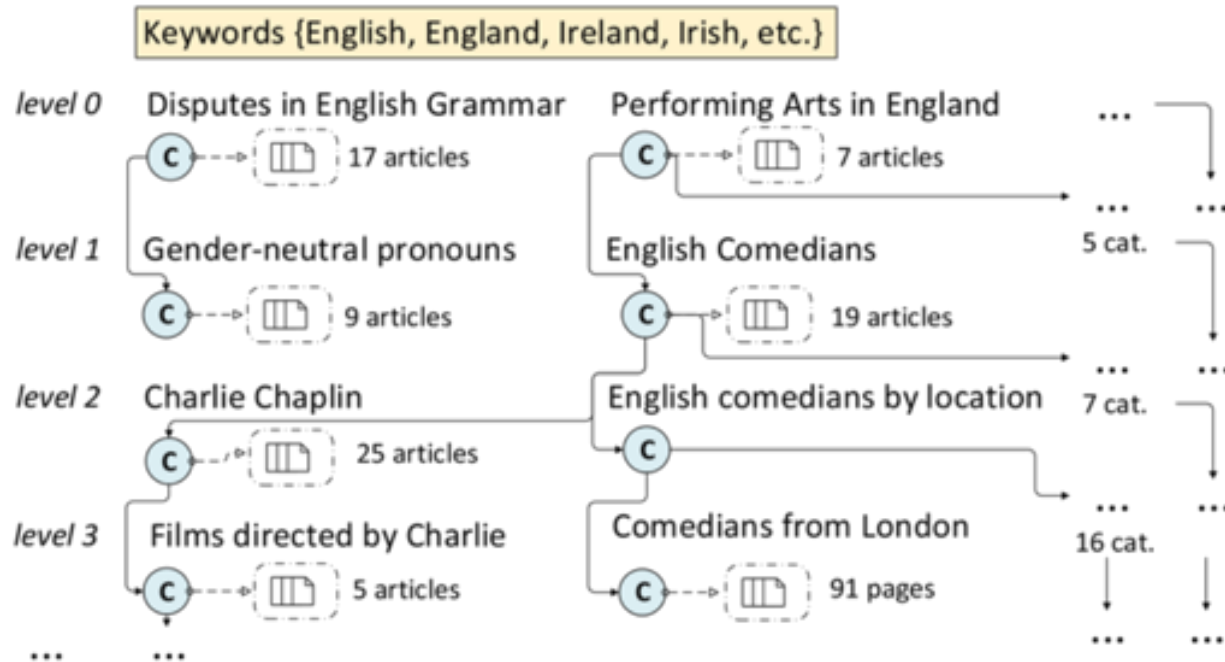
The screenshot shows the Wikipedia article for "Times Square". The title is "Times Square" with a subtitle "From Wikipedia, the free encyclopedia". The coordinates are "40°45'28"N 73°59'08"W". The article text describes Times Square as a major commercial intersection, tourist destination, and entertainment center in Midtown Manhattan. It mentions its location at the junction of Broadway and Seventh Avenue, its size (stretching from West 42nd to West 47th Streets), and its history, including its renaming in 1904 and its role in the annual New Year's Eve ball drop. The sidebar on the left contains various navigation links such as "Main page", "Contents", "Featured content", "Current events", "Random article", "Donate to Wikipedia", "Wikipedia store", "Interaction", "Help", "About Wikipedia", "Community portal", "Recent changes", "Contact page", "Tools", "What links here", "Related changes", "Upload file", "Special pages", "Permanent link", "Page information", "Wikidata item", "Cite this page", "Print/export", "Create a book", "Download as PDF", "Printable version", "In other projects", "Wikimedia Commons", "Languages", and "Català", "Español", "Euskara", "Français", "Galego".



The screenshot shows the Wikipedia article for "English literature". The title is "English literature" with a subtitle "From Wikipedia, the free encyclopedia". The article text focuses on English-language literature rather than the literature of England, mentioning 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. It discusses the development of the English language over more than 1,400 years, from Old English to Middle English, and the influence of the British Empire on the spread of the English language. The sidebar on the left contains various navigation links such as "Main page", "Contents", "Featured content", "Current events", "Random article", "Donate to Wikipedia", "Wikipedia store", "Interaction", "Help", "About Wikipedia", "Community portal", "Recent changes", "Contact page", "Tools", "What links here", "Related changes", "Upload file", "Special pages", "Permanent link", "Page information", "Wikidata item", "Cite this page", "Print/export", "Create a book", "Download as PDF", "Printable version", "In other projects", "Wikimedia Commons", "Wikibooks", "Languages", and "Asturianu", "Català".

• Geolocation in one of the territories

• Keyword (demonym/territory name) on title



Category crawling using keywords

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

Dylan Moran

From Wikipedia, the free encyclopedia

Dylan William Moran (/ˈmɒrən/; 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
 - 1.1 Early life
 - 1.2 Career
 - 1.3 Awards and commendations
- Filmography
 - 2.1 Film
 - 2.2 Television
- Stand-up DVDs
- References
- External links

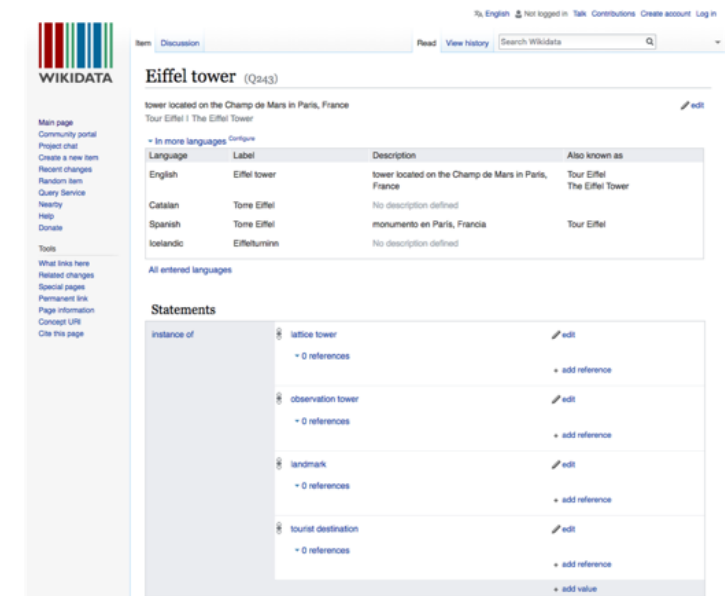
Biography [edit]

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

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' (Q2413). The page includes a sidebar with navigation links, a main content area with the item's label and description, and a table of statements with their references.

Wikidata

Item: **Eiffel tower** (Q2413)

tower located on the Champ de Mars in Paris, France
Tour Eiffel | The Eiffel Tower

• In more languages [Configure](#)

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	Eiffeltúnninn	No description defined	

All entered languages

Statements

Instance of	Statement	References	Action
	lattice tower	0 references	edit
		add reference	
	observation tower	0 references	edit
		add reference	
	landmark	0 references	edit
		add reference	
	tourist destination	0 references	edit
		add reference	
		add value	

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



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/

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 extent in non-western languages (African and Asian) is on average much smaller (Miquel-Ribé and Laniado, 2019).

CCC articles tend to be more developed in number of Bytes, images, and categories (Miquel-Ribé, 2017).



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 considering a Wikipedia language edition cultural diversity as the coverage of all the others' CCC.



Strategy (goals and priorities): focusing on the essential

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.

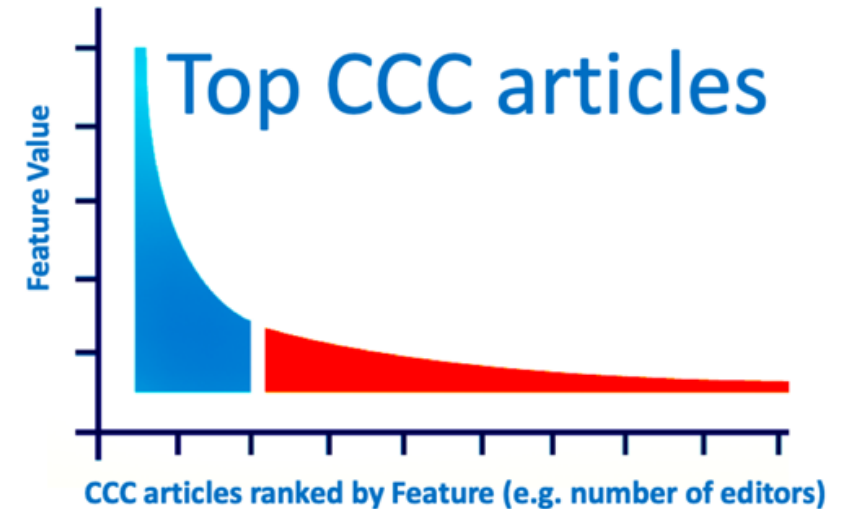
28-30 thousand articles to cover a minimum of Wikipedia cultural diversity.



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://wcdo.wmflabs.org/top_ccc_articles/?list=men&lang_origin=pl&lang_target=uk

http://wcdo.wmflabs.org/top_ccc_articles

Top 500 CCC articles list "Men" from Catalan CCC to Standard Chinese

The following table shows the Top 500 articles list Men from Catalan CCC and its article availability in Standard Chinese Wikipedia. Articles are sorted by the feature **Men**. The rest of columns present complementary features that are explicative of the article relevance (number of editors, edits, pageviews, Bytes, Wikidata properties or Interwiki links). In particular, number of inlinks from CCC (incoming links from the CCC group of articles) highlights the article importance in terms of how much it is required by other articles. The column named Other Language present Interwiki links to the article version when available in the four languages closer to the target language (those that cover best this language and therefore it is likely their editors consult it).

The table's last column shows the article title in its target language, in **blue** when it exists, in **red** as a proposal generated with the Wikimedia Content Translation tool or as an existing Wikidata label in the same language, and **empty** when the article does not exist or there is no title proposition available.

The available Top CCC articles lists are: list of CCC articles with most pageviews during the last month (**Pageviews**), list of CCC articles with most number of editors (**Editors**), 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 edits in talk pages (**Discussions**), 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 (**Geolocated**), list of CCC articles with keywords on title with most bytes (**Keywords**), list of CCC articles categorized in Wikidata as women with most edits (**Women**) and list of CCC articles categorized in Wikidata as men with most edits (**Men**).

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'.

Select the parameters

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

List

Language origin

Country origin

Language target

Men

Catalan (ca)

Select a country (optional)

Standard Chinese (zh)

QUERY RESULTS!

Nº	Catalan Title	Edits	Editors	Pageviews	Bytes	References	Wikidata Properties	Interwiki Links	Inlinks from CCC	Creation Date	Other Languages	Standard Chinese Title
1	Ramon Llull	1569	449	536	110.2k	98	82	56	136	2003-11-10	en , sv	拉蒙·福利
2	Antoni Gaudí i Cornet	1214	371	144	90.2k	57	81	115	131	2003-08-10	en , sv , zh_yue	安东尼·高迪
3	Jacint Verdaguer i Santaló	1081	262	82	93.7k	41	56	33	95	2003-08-11	en , sv	
4	Rafael Casanova i Comes	1062	146	2371	268.2k	52	27	24	8	2005-09-16	en	
5	Joan Miró i Ferrà	920	227	55	98.4k	64	114	94	91	2006-01-29	en , sv , ceb , zh_yue	胡安·米羅
6	Salvador Dalí i Domènech	880	366	118	49.3k	52	145	195	46	2003-06-10	en , sv , zh_yue	萨尔瓦多·达利
7	Artur Mas i Gavarró	841	265	873	54.0k	95	41	50	98	2004-05-20	en , sv	阿图尔·马斯
8	Guillem el Piadè	826	190	152	95.6k	33	28	27	83	2004-12-03	en , sv	
9	Josep Guardiola i Sala	821	272	1884	44.3k	53	55	70	21	2005-12-29	en , sv , zh_yue	佩普·瓜迪奥拉
10	Pablo Picasso	774	308	106	47.7k	56	167	217	93	2005-06-26	en , sv , zh_yue	巴勃罗·毕加索

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

- **Countries Top CCC articles coverage**

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

Lang = wikicode

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

This page shows some st

These lists are created by plain CCC or geolocated 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 use alphabetic order by their 1 articles from the Top 100

The challenge is to reach

Language	Wik
Afar	aa
Abkhaz	ab
Acehnese	ace
Adyghe	ady
Afrikaans	af
Akan language	ak
Alemannic	als
Amharic	am
Andonese	an
Old English	ang

This page shows some statis language edition (when it is a

Some languages are mapped order to create lists for count selection process, and later, 1 speaking territories), whose b

These lists are created by ran plain CCC or geolocated artik articles with featured article d from CCC, list of CCC article categorized in Wikidata as m last year and with most edits

The following table is useful t country level. Countries are s CCC) row would be for Catal columns, **Lists Coverage Idx**

The challenge is to reach 10

Country

- Afghanistan (Persian CCC)
- Afghanistan (Lahnda language CCC)
- Afghanistan (Pashto CCC)
- Afghanistan (Uzbek CCC)

Countries Top CCC articles coverage by Catalan Wikipedia

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
Amharic	am	22%	9%	7/98	1/25	0/0	2/8	4/5	1/1	17%	12%	27.6	23	Sub-Saharan Africa
Andonese	an	96%	69%	43%	36%	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.

Strategy (goals and priorities)

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.

Wikipedia Cultural Diversity Observatory (WCDO).

As seen, these four work lines complement each other.

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



Mission:

Align the Wikimedia movement to represent and share the existing cultural diversity in the Wikipedia language editions.

Reaching Maturity on Cultural Diversity in Wikimedia

Having a **bigger picture** may help the Wikimedia Foundation.

Its different teams at advancement, software development and community engagement should break the barriers that stop editors during the **representation** and the **sharing processes**.



Wikimedia Foundation teams
(infrastructure, product, advancement...)

- Knowing the degree of success at representation and sharing can be key in order to allocate resources to specific events, community engagement programs, among others.

Reaching Maturity on Cultural Diversity in Wikipedia communities

Communities and Chapters benefit directly from each of the different work lines we exposed.



Chapters (Indian, German, Italian, Catalan, Armenian, South African...)



Online Communities

We can consider a maturity model in order to understand what point any community has reached in terms of cultural context representation and sharing according to how many elements they have incorporated:

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

Cultural Diversity Maturity Model in Wikipedia Language Communities

Level	01 Unintentional →	02 Spontaneous →	03 Organized →	04 Controlled →	05 Distributed
Situation	Few editors translating general encyclopedic articles (New York, Mona Lisa, etc.). No cultural context representation.	Editors represent their own cultural context and translate articles to cover cultural diversity individually.	Events to represent own context (e.g. Wiki Loves Monuments), spread it (e.g. Catalan Culture Challenge) and cover others' (e.g. Asian Month).	While the use of metrics shows the knowledge gaps but its use is incipient, the community organization is mature and has capacity.	Cultural diversity has dedicated events and is also cross-section. Editors follow the stats on the depth of the gaps and regularly use the tools to bridge them.
Incorporated elements	None	Discourse	Discourse Organization	Discourse Organization Awareness	Discourse Organization Awareness Strategy
Barriers	Editing barriers i.e. digital divide, sociocultural barriers.	Lack of community building and offline support.	Difficulty to assess the impact and gaps.	Metrics and tools to find gaps are not integrated in the editors workflow.	
Tools to reach next level	Lack of editors	Organization	Quantification	Strategic goals	
Community Example	Some African languages.	Maltese and Walloon, among others.	Catalan, Spanish, Italian, among others.	CEE languages (e.g. Ukrainian and German) among others.	None yet

Conclusions

- Improving cultural diversity in Wikipedia is a daunting goal that we believe it can only be tackled with research.
- It implies different areas of work besides research (discourse, community programs, tool development, etcetera).
- WCDO is transversal project as it provides resources to both the Wikimedia Foundation departments, chapters and communities.
- Current tool prototypes and dashboards will be improved over time as they need to be localized and refined with user feedback.
- Datasets and research in academic venues is collateral to this project goal but essential for improving its quality and inspiring further research in the Digital Humanities field.



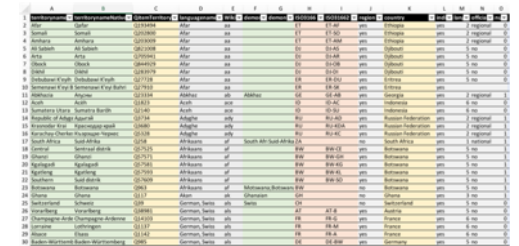
Future work (In progress)

Awareness (metrics and visualizations): mapping all languages and contexts

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

If we need to map all the languages to territories, we can find:

- a) **Potential new Wikipedia languages.**
- b) **Contexts explained by a language of a higher status** instead of the native one (marginalization).



The image shows a screenshot of a Wikipedia language list table. The table has columns for language name, ISO 639-1 code, ISO 639-2 code, and a status column. The status column contains values like 'active', 'inactive', 'extinct', 'unclassified', etc. The table is color-coded with green for active languages and yellow for inactive or unclassified languages. The table is sorted by the number of Wikipedia articles in each language, with the most active languages at the top.

The case of b) 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 language (from Uganda) has a very low CCC %. There is an opportunity to give a digital revitalization of the language through Wikipedia.

Future work (In progress)

Cultural Diversity Observatory Functionalities by the end of 2019



Language-based
Gap Dashboards



Temporal
Monitoring



Geolocated
Articles Map



Monthly
Newsletter

**Remind the gaps situation
(Awareness)**



Top CCC
Article Lists



Article Searcher



Image
Galleries



Multilingual
Users Finder

**Assist the sharing process
(Organization)**



Some tools to organize better and bridge the most important gaps.

Some dashboards on metrics and results to raise awareness.

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 Foundation – Project Grant



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.

