Wikipedia Cultural Diversity Observatory:
Empowering Communities to Bridge the Culture Content Gaps

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

Marc Miquel, PhD
{marcmiquel@gmail.com}
Username:marcmiquel
Pompeu Fabra University, Barcelona, Catalonia
Amical Wikimedia (Catalan Wikipedia)
WG Diversity – Strategy 2030
Wikimania 2018
Has the Wikimania 2018 theme really lead to bridging the knowledge gaps? (especially, the African one)
Wikidata geolocated items August 2019

Northern Europe
23,207,648
19.77%

Eastern Europe
7,595,720
10.35%

Southern Asia
3,094,848
4.69%

Western Europe
10,122,381
15.33%

Eastern Asia
1,868,320
3.18%

Southern Europe
7,943,130
6.01%

Western Asia
2,848,338

Sub-Saharan Africa
16,061,597
4.13%

Northern America
6,092,650
16.61%

Latin America and the Caribbean
34,823,828
9.13%

Australia and New Zealand
1,001,964

Legend:
- Antarctica
- Oceania
- Africa
- Asia
- Americas
- Europe
What happened with African articles before and after the Wikimania?

Wikimania (July 2018, Q3) has had different impact on major European languages.
Wikimedia Italy has raised awareness for some time on Africa.

2006. WikiAfrica is launched by lettera27 Foundation in collaboration with Wikimedia Italia. They make it visible in their communication.
Problem of Cultural Diversity
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 and activists to study and fight against the knowledge gaps and increase cultural diversity in contents”.

Its work lines are:
• Discourse (concepts and discussions)
• Awareness (metrics and observations)
• Organization (events and tools)
• Strategy (goals and priorities)

http://wcdo.wmflabs.org
What do we understand by cultural diversity in Wikipedia?

Cultural diversity is the sum of all the concepts from all the different contexts generated in all languages.

1. Language
2. Territories (Context) / Speakers
3. Concepts (i.e. traditions, language, politics, agriculture, biographies, places, events, etcetera.)

This is a **pragmatic definition** with the aim at collecting Wikipedia articles.
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.
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)*

---

**Missing pieces**
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.

(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]
• **Geolocation**
• **Category Graph**
• **Keywords in title**
• **Some Wikidata properties**
  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,…).

• **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 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.
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.
Strategy (goals and priorities)

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://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 "Women" from Swedish CCC in Catalan Wikipedia

Select the parameters

<table>
<thead>
<tr>
<th>List</th>
<th>Language origin</th>
<th>Country origin</th>
<th>Language target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>Swedish (sv)</td>
<td>Select a country (optional)</td>
<td>Catalan (ca)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N°</th>
<th>Swedish Title</th>
<th>Edits</th>
<th>Editors</th>
<th>Pageviews</th>
<th>Bytes References</th>
<th>Wikidata Properties</th>
<th>Interwiki Links</th>
<th>Inlinks from CCC</th>
<th>Creation Date</th>
<th>Other Languages</th>
<th>Catalan Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Astrid Lindgren</td>
<td>2269</td>
<td>928</td>
<td>1539</td>
<td>68.6k 96</td>
<td>94</td>
<td>97</td>
<td>546</td>
<td>2003-01-23</td>
<td>es, en, fr, it</td>
<td>Astrid Lindgren</td>
</tr>
<tr>
<td>2</td>
<td>Selma Lagerlöf</td>
<td>2225</td>
<td>926</td>
<td>1732</td>
<td>29.0k 78</td>
<td>107</td>
<td>125</td>
<td>426</td>
<td>2003-01-23</td>
<td>es, en, fr, it</td>
<td>Selma Lagerlöf</td>
</tr>
<tr>
<td>3</td>
<td>Mona Sahlin</td>
<td>1512</td>
<td>598</td>
<td>366</td>
<td>29.7k 62</td>
<td>34</td>
<td>28</td>
<td>309</td>
<td>2003-09-13</td>
<td>es, en, fr, it</td>
<td>Mona Sahlin</td>
</tr>
<tr>
<td>4</td>
<td>Carola Häggkvist</td>
<td>1384</td>
<td>646</td>
<td>644</td>
<td>42.8k 31</td>
<td>36</td>
<td>39</td>
<td>376</td>
<td>2004-03-14</td>
<td>es, en, fr, it</td>
<td>Carola Häggkvist</td>
</tr>
<tr>
<td>5</td>
<td>Drottning Kristina</td>
<td>1057</td>
<td>484</td>
<td>613</td>
<td>58.6k 41</td>
<td>62</td>
<td>63</td>
<td>581</td>
<td>2002-12-05</td>
<td>es, en, fr, it</td>
<td>Cristina de Suècia</td>
</tr>
<tr>
<td>6</td>
<td>Amy Deasismont</td>
<td>983</td>
<td>476</td>
<td>314</td>
<td>15.9k 34</td>
<td>29</td>
<td>24</td>
<td>3</td>
<td>2005-03-14</td>
<td>es, en, fr, it</td>
<td>Amy Deasismont</td>
</tr>
<tr>
<td>7</td>
<td>Kronprinsessan Victoria</td>
<td>957</td>
<td>454</td>
<td>1116</td>
<td>30.6k 48</td>
<td>37</td>
<td>58</td>
<td>161</td>
<td>2003-02-01</td>
<td>es, en, fr, it</td>
<td>Victoria de Suècia</td>
</tr>
<tr>
<td>8</td>
<td>Loreen</td>
<td>843</td>
<td>357</td>
<td>260</td>
<td>40.9k 95</td>
<td>39</td>
<td>63</td>
<td>244</td>
<td>2005-06-22</td>
<td>es, en, fr, it</td>
<td>Loreen</td>
</tr>
<tr>
<td>9</td>
<td>Edda Magnason</td>
<td>805</td>
<td>66</td>
<td>98</td>
<td>27.3k 55</td>
<td>23</td>
<td>9</td>
<td>19</td>
<td>2010-01-20</td>
<td>en, fr</td>
<td>Edda Magnason</td>
</tr>
<tr>
<td>10</td>
<td>Agnes Carlsson</td>
<td>770</td>
<td>400</td>
<td>362</td>
<td>38.7k 77</td>
<td>40</td>
<td>27</td>
<td>149</td>
<td>2005-12-02</td>
<td>es, en, fr, it</td>
<td>Agnes Carlsson</td>
</tr>
<tr>
<td>11</td>
<td>Prinsessan Madeleine</td>
<td>754</td>
<td>338</td>
<td>650</td>
<td>17.2k 31</td>
<td>29</td>
<td>45</td>
<td>68</td>
<td>2001-07-28</td>
<td>es, en, fr, it</td>
<td>Magdalena de Suècia</td>
</tr>
</tbody>
</table>
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
Problem: smaller language editions do not even have 100 on their cultural context to fill the lists.
253 out of the 302 Wikipedia languages are from languages which coexist 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 inverse situation

In some African languages, cultural context content tends to be represented in English or French rather than in their native Wikipedia.

For instance, **Luganda Wikipedia** (from Uganda) has a very low CCC (**just 3.15%**).
• 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.

**President of Uganda does not exist**

**Prime Minister of Uganda does not exist**
### Missing CCC Articles List

**Select the Wikipedia**
- Target Language: Luganda (lg)
- Target Country (Optional): Select a country (optional)
- Target Region (Optional): Select a region (optional)

**Filter by content**
- Search Type: Missing articles
- CCC Segment: Select a CCC Segment
- Topic: Select a topic

**Choose source of content**
- Source language(s): English (en)
- Order by feature: Number of Inlinks
- Limit the results: 100

---

<table>
<thead>
<tr>
<th>No.</th>
<th>Language</th>
<th>Title</th>
<th>Number of Editors</th>
<th>Number of Pageviews</th>
<th>Number of Inlinks</th>
<th>Number of Bytes</th>
<th>Label (Lang)</th>
<th>Qitem</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>en</td>
<td>Nile</td>
<td>0</td>
<td>9131</td>
<td>2825</td>
<td>0</td>
<td>en Nile</td>
<td>Q3392</td>
</tr>
<tr>
<td>2</td>
<td>en</td>
<td>Finance Trust Bank</td>
<td>0</td>
<td>42</td>
<td>589</td>
<td>0</td>
<td>en Central Region</td>
<td>Q7877595</td>
</tr>
<tr>
<td>3</td>
<td>en</td>
<td>Central Region, Uganda</td>
<td>0</td>
<td>123</td>
<td>568</td>
<td>0</td>
<td>en Central Region</td>
<td>Q429685</td>
</tr>
<tr>
<td>4</td>
<td>en</td>
<td>Uganda Protectorate</td>
<td>0</td>
<td>444</td>
<td>554</td>
<td>0</td>
<td>en Operation Entebbe</td>
<td>Q1097943</td>
</tr>
<tr>
<td>5</td>
<td>en</td>
<td>Operation Entebbe</td>
<td>0</td>
<td>6000</td>
<td>513</td>
<td>0</td>
<td>en Operation Entebbe</td>
<td>Q33933</td>
</tr>
<tr>
<td>6</td>
<td>lc</td>
<td>Kampala–Jinja Highway</td>
<td>0</td>
<td>34</td>
<td>439</td>
<td>0</td>
<td>lc Kampala–Jinja Highway</td>
<td>Q6359092</td>
</tr>
<tr>
<td>7</td>
<td>ldo</td>
<td>Law Development Centre</td>
<td>0</td>
<td>65</td>
<td>404</td>
<td>0</td>
<td>ldo Law Development Centre</td>
<td>Q6503014</td>
</tr>
<tr>
<td>8</td>
<td>en</td>
<td>Kampala–Hoima Road</td>
<td>0</td>
<td>21</td>
<td>396</td>
<td>0</td>
<td>en Kampala–Hoima Road</td>
<td>Q25226934</td>
</tr>
</tbody>
</table>
Conclusion
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”.

Its work lines are:
• Discourse (concepts and discussions)
• Awareness (metrics and observations)
• Organization (events and tools)
• Strategy (goals and priorities)

http://wcdo.wmflabs.org
Future ideas

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

Diversity Stats and Metrics would guide and help in covering the unrepresented *concepts* and *points of view*, related to all groups:

- Gender
- Culture
- LGBT+
- Historically marginalized communities,
- Among others.

**WG Diversity – Strategy 2030: Diversity Metrics Recommendation**

https://tinyurl.com/diversitymetrics
We should help minoritized, minority and potential languages build a minimal capacity.
We should encourage them not only to build a Wikipedia,

But to create their most unique and specific knowledge about their context: few lists of 100 articles can really make a difference.

We need to value their knowledge and difference.
Each context has some ‘uniqueness’ to contribute to human knowledge that is best explained by their native languages.

Wikipedia is their best strategy to pass this 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 Foundation – Project Grant
References (if you want to know more)


