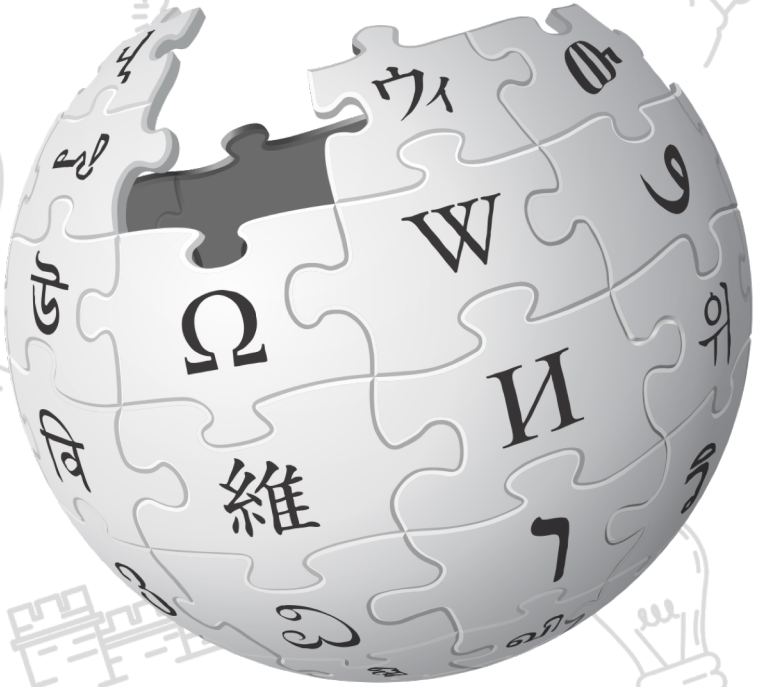
The Wikipedia Diversity Observatory

A Project to Identify and Bridge Content Gaps in Wikipedia

Introduction

The Wikipedia
Diversity Observatory

“The sum of human knowledge”



More than 300 Wikipedia language editions

Content diversity is about representing and sharing concepts across language editions

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

"Local Content" or CCC (Cultural Context Content) is on avg. 25% of the largest 40 Wikipedias (Miquel and Laniado, 2017)

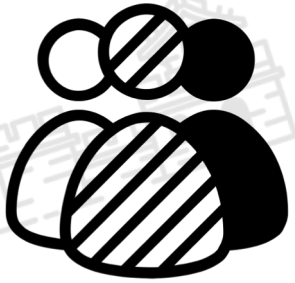
It is often not shared across languages (Culture Gap).

Religion

Sexual Orientation

Geography

Gender

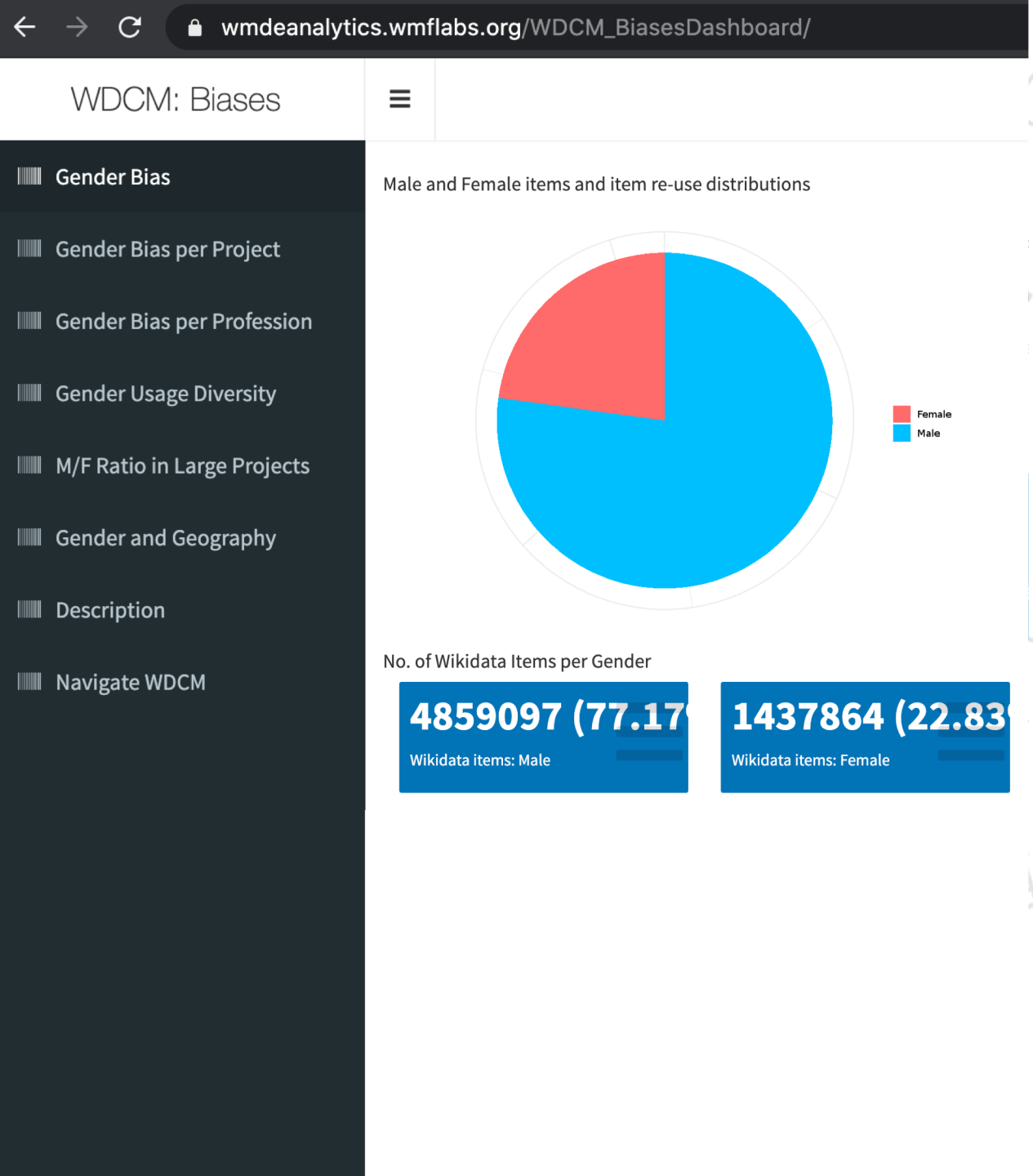


Ethnicity

Culture Gap
Gender Gap
Ethnicity Gap
Geography Gap
…

"Bridging the Knowledge Gaps – The Ubuntu Way Forward" Wikimania 2018

(Geography Gap)



Some tools appeared (Gender Gap)

How can we help editors identify and bridge the gaps on any relevant category?

The Wikipedia Diversity Observatory

Providing datasets, visualizations and tools to work towards more diversity within Wikipedia language editions.

Approach

The Wikipedia
Diversity Observatory

1. We create the database/datasets.

Methodology
(Miquel and Laniado, 2019)

Label every Wikipedia language edition article according to categories relevant to diversity.

- **Wikidata features**

Geography
Language
Gender
…

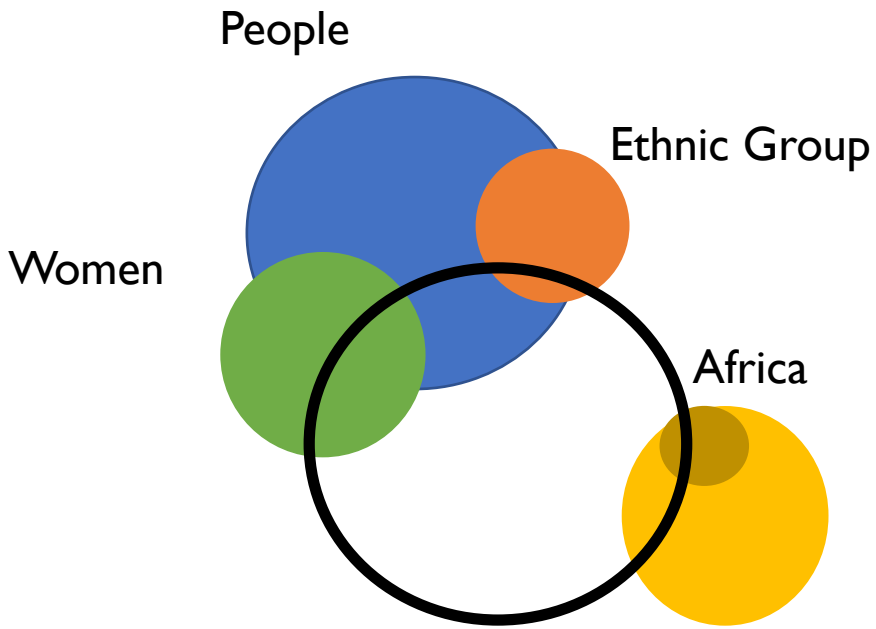
- **Graph Structure (Links)**
- **Category Structure**
- **Machine Learning for "Local Content"**

Datasets (CSV or SQlite3) available at:
https://wcdo.wmflabs.org/datasets/

2. We compute the statistics.

Methodology
(Miquel and Laniado, 2019)

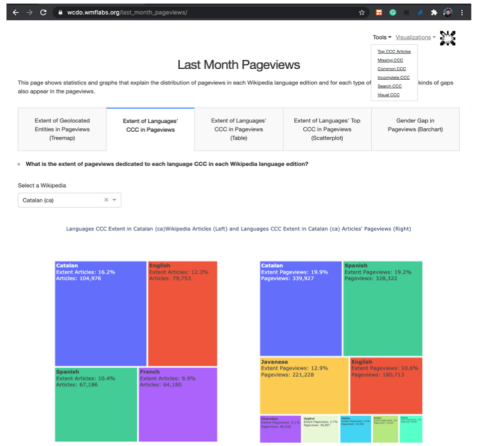
Compute the intersections of different groups of articles and put them in a database that stores all the history.



German Wikipedia
Accumulated articles May 2020

3. We create the dashboards with visualizations and tools.

Website



wcdo.wmflabs.org

Visualizations

- Culture Gap
- Geographic Gap
- Gender Gap
- Last Month Pageviews
- Diversity Over Time
- …

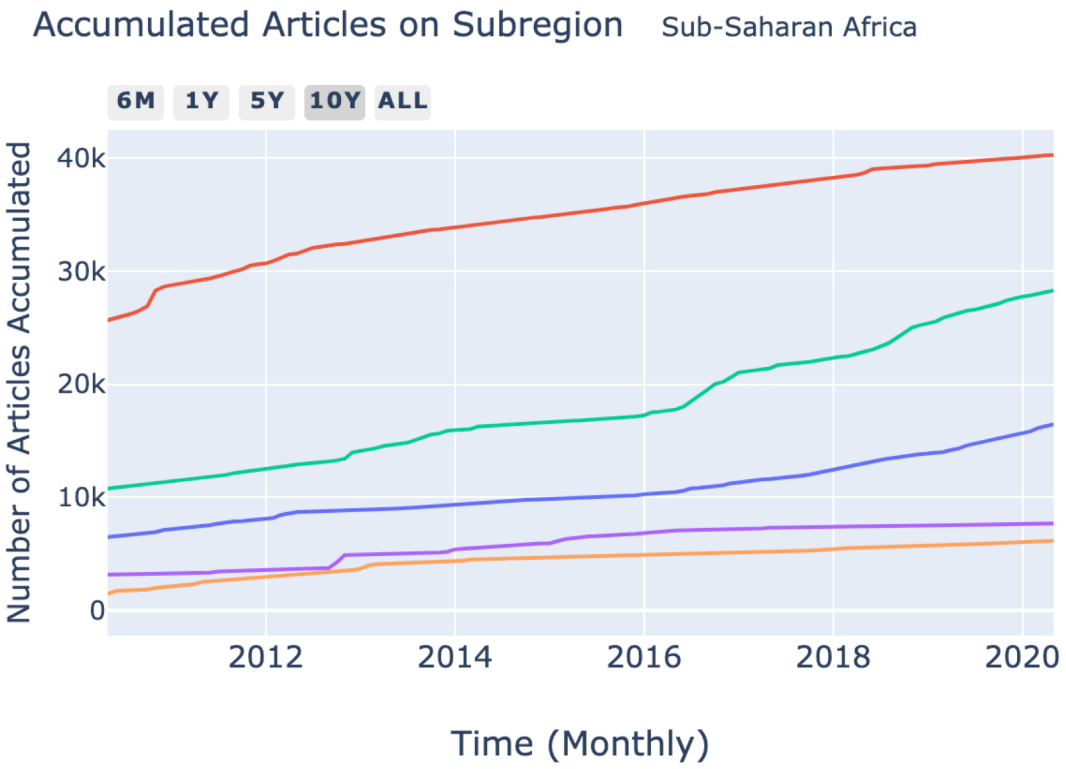
Tools

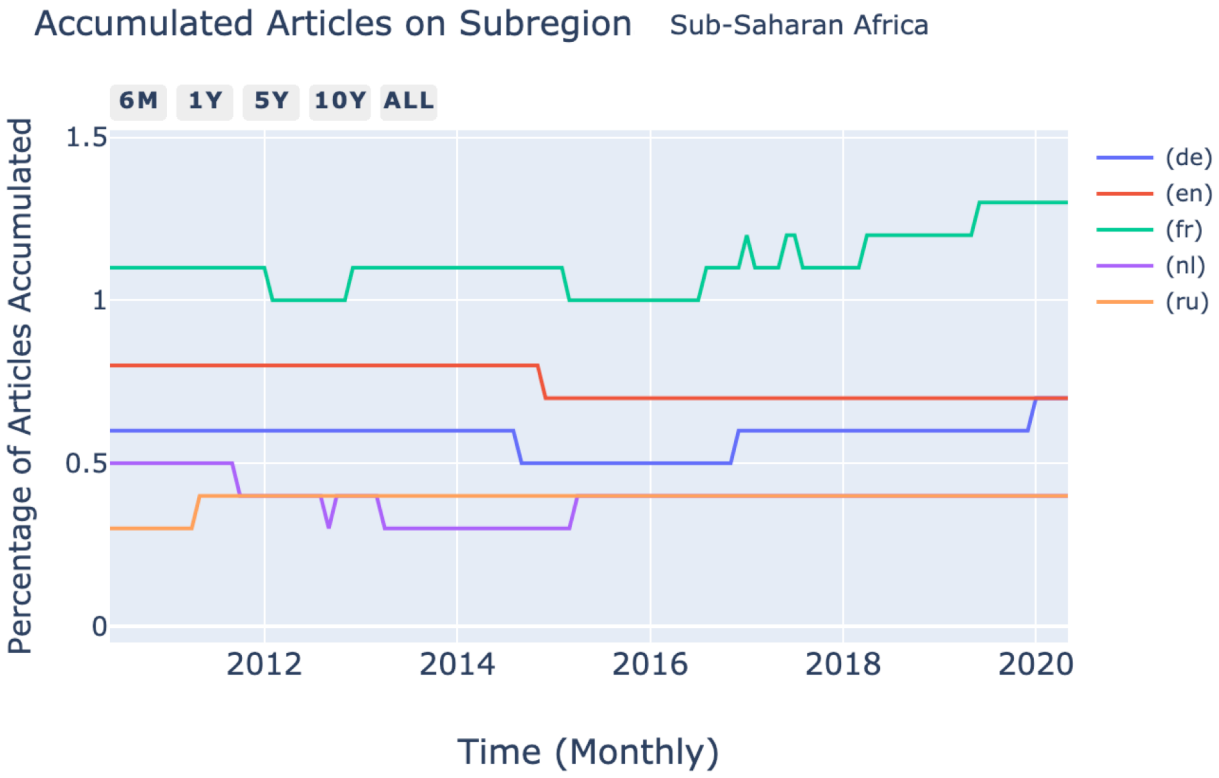
- Top CCC Diversity Lists
- Common CCC
- Missing CCC
- Visual CCC
- Incomplete CCC
- Search CCC

Dashboards:
Visualizations and Tools

The Wikipedia
Diversity Observatory

Diversity Over Time





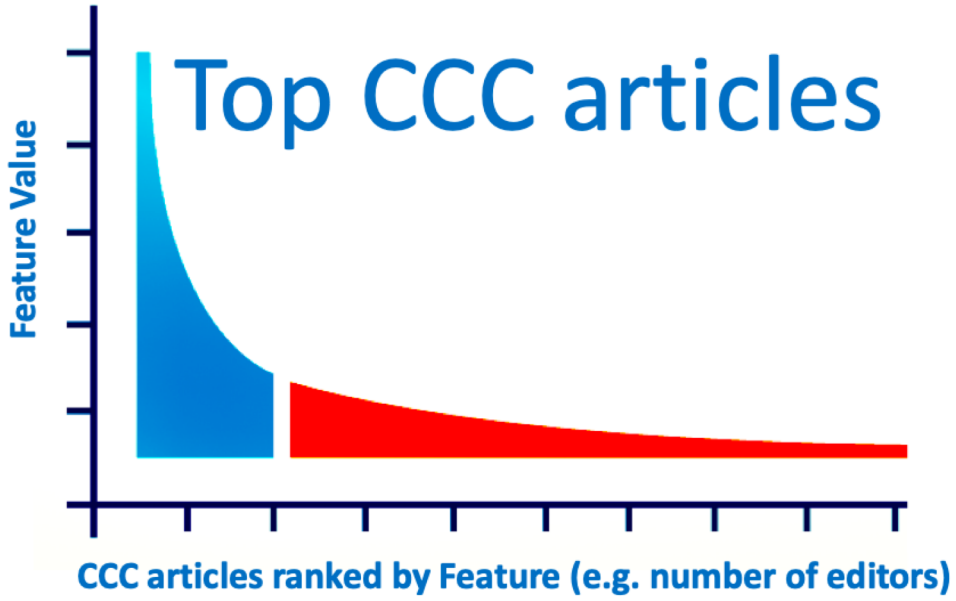
Has Wikimania been useful to encourage the creation of articles geolocated in Subsaharan Africa?

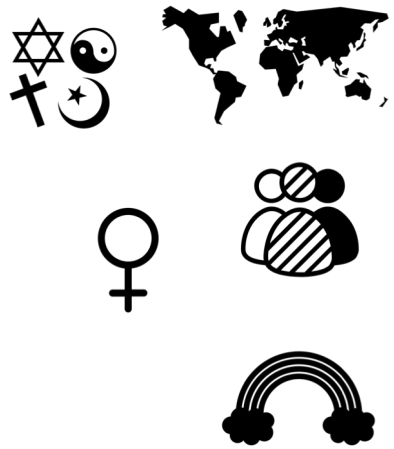
https://wcdo.wmflabs.org/diversity_over_time/

Dashboards:
Visualizations and Tools

The Wikipedia
Diversity Observatory

Top relevant articles about any language local content context segmented by **more than 25 topics**





Lists of 100 – 500 articles that should be in every language edition

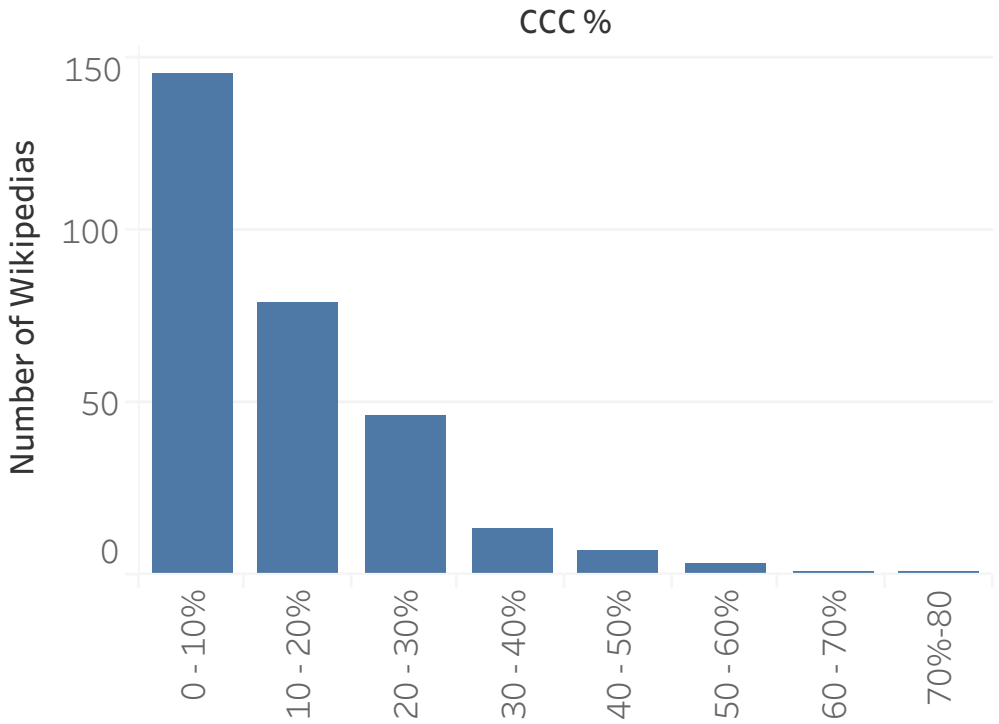
Case 1: Exporting "Local Content" to other languages

Yoruba Top CCC articles list "Women" and its coverage by Catalan Wikipedia

Nº	Yoruba Article Title	Edits	Editors	Creation Date	Related Languages	Catalan Article Title
1	Genevieve Nnaji	62	12	2009-09-24	es, en, fr, it	Genevieve Nnaji (label)
2	Quincy Olasumbo Ayodele	36	3	2016-07-12	en	Quincy Olasumbo Ayodele (translation)
3	Funmilayo Ransome-Kuti	24	6	2009-12-10	es, en, fr, it	Funmilayo Ransome-Kuti
4	Salawa Abeni	21	6	2011-06-18	es, en	Salawa Abeni (translation)
5	Ngozi Okonjo-Iweala	20	7	2008-10-08	es, en, fr	Ngozi Okonjo-Iweala
6	Agbani Darego	17	6	2009-12-19	es, en, fr, it	Agbani Darego (translation)
7	Oreoluwa Lesi	14	2	2018-10-13	en	
8	Chimamanda Ngozi Adichie	13	10	2009-12-26	es, en, fr, it	Chimamanda Ngozi Adichie
9	Nkechi Justina Nwaogu	13	3	2011-06-18	en	Nkechi Justina Nwaogu (label)
10	Onyeka Onwenu	13	4	2011-06-18	en	Onyeka Onwenu (translation)

https://wcdo.wmflabs.org/top_ccc_articles/?list=women&target_lang=ca&source_lang=yo

We have many Wikipedias with underdeveloped local content



145 Wikipedias CCC is below 10% of their content.

[https://wcdo.wmflabs.org/list_of_wikipedias_by_cultural_context_content]

Wolof Wikipedia

(Wolof is spoken in Senegal and Mauritania)

Alex Ferguson,
Scottish Football coach

Available in Wolof

Ronald Reagan,
American president

Available in Wolof

Anna Rita del Piano,
Italian theatre actress

Available in Wolof

Macky Sall,
president of Senegal

Not available in Wolof

Case 2: From a Wikipedia to a Language Local Content

Missing CCC Articles in Wolof Wikipedia on people

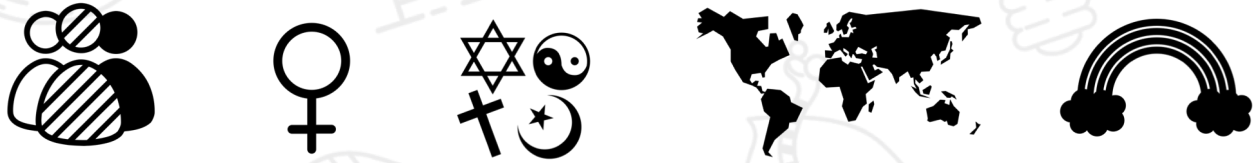
Nº	Language	Title	Editors	Pageviews	Interwiki	Bytes	Lang Label
1	en	Tacko Fall	118	53384	5	10.8k	fr Tacko Fall
2	en	Patrice Evra	1774	7346	63	133.0k	fr Patrice Évra
3	en	Idrissa Gueye	212	5512	38	14.2k	fr Idrissa Gueye
4	en	Patrick Vieira	1570	3271	61	83.1k	wo Patrick Vieira
5	en	El Hadji Diouf	1592	2086	36	55.0k	fr El-Hadji Diouf
6	en	Papiss Cissé	960	1400	34	34.1k	fr Papiss Cissé
7	en	Macky Sall	152	1068	48	31.1k	fr Macky Sall
8	en	Mame Biram Diouf	675	732	37	36.1k	fr Mame Biram Diouf
9	en	Dame N'Doye	438	689	27	17.1k	fr Dame N'Doye
10	en	Gorgui Dieng	126	632	21	19.3k	fr Gorgui Dieng

https://wcdo.wmflabs.org/missing_ccc_articles/? topic=men&source_lang=en&target_lang=wo

Solutions to improve coverage and spread

For every Wikipedia language edition, regardless of its community size and current capacity.

For any category relevant to diversity.



Conclusions

The Wikipedia
Diversity Observatory

- Research to foster content diversity in peer-production through raising awareness and providing tools.
- Strategic to the Wikimedia Movement as it helps coordinating efforts across all Wikipedia language editions (Strategy 2030).
- Integrated in community contests and events like Intercultur, CEE Spring, among others.
- Finding gaps is relevant to partnerships and education programs.

The Wikipedia Diversity Observatory

Providing datasets, visualizations and tools to work towards more diversity within Wikipedia language editions.

Thank you very much!

References (if you want to know more)

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.*

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. **(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. **(2016)**. Cultural identities in wikipedias. In *Proceedings of the 7th 2016 International Conference on Social Media & Society* (p. 24). ACM.