Talking about…

Wikipedia Diversity Observatory and the Knowledge Gaps Project

Marc Miquel i Ribé

marcmiquel@gmail.com

Universitat Pompeu Fabra, Barcelona

Wikimedia Research Team

Wikimedia Research Sessions, September 22nd, 2020

I’m Marc Miquel

- Researcher (PhD Thesis about Group Identities and Participation in Wikipedia)
- WMF Grantee (Wikipedia Diversity Observatory)
- Amical Wikimedia (Catalan Community Chapter), member 2011, chairman.
- Lead Writer in Wikimedia Strategy 2030 (Diversity WG and final Writers)

This talk

- First, I will explain the Diversity Observatory project and outcomes
- Second, I will review the Knowledge Gaps Taxonomy and propose changes.

Feel free to interrupt whenever you consider. This is a ~~talk~~ conversation.

The Wikipedia Diversity Observatory

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

Diversity Observatory Approach

The Wikipedia
Diversity Observatory

Content diversity in Wikipedia is about representing and sharing concepts across language editions

Religion

Sexual Orientation

Geography

Gender

Ethnicity

To achieve more content diversity in content and fight for knowledge equity, it is necessary to represent all the different:

1) places (geographical entities),

2) peoples (gender, sexual orientation, religious groups, ethnic groups, and indigenous group),

3) cultural context concepts for each group of people and place, and

4) languages (national, indigenous and marginalized) of the world in Wikipedia.

Religion

Sexual Orientation

Geography

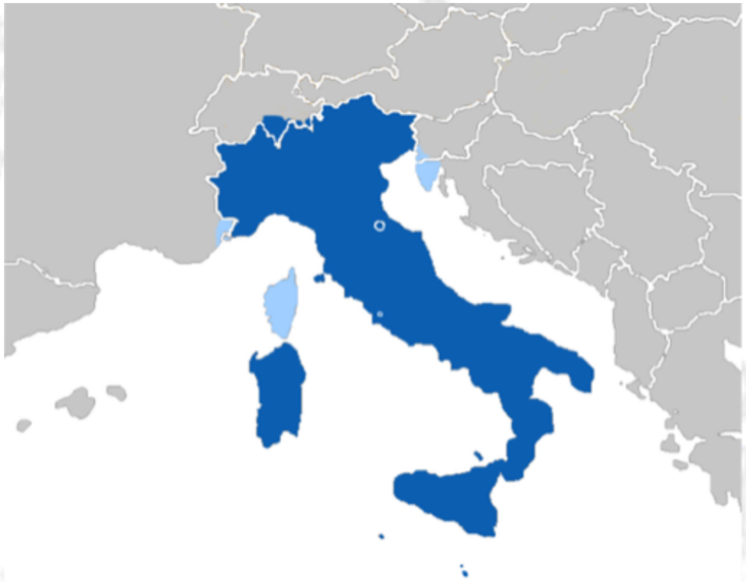
Gender

Ethnicity

Culture Gap
Gender Gap
Ethnicity Gap
Geography Gap
...

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, etc.)

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.



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).

We define each category relevant to overall diversity

Concepts can relate to categories such as:

- Geography
- Events
- People
    - Gender
    - Sexual Orientation
    - Ethnic Group
    - Religious Group

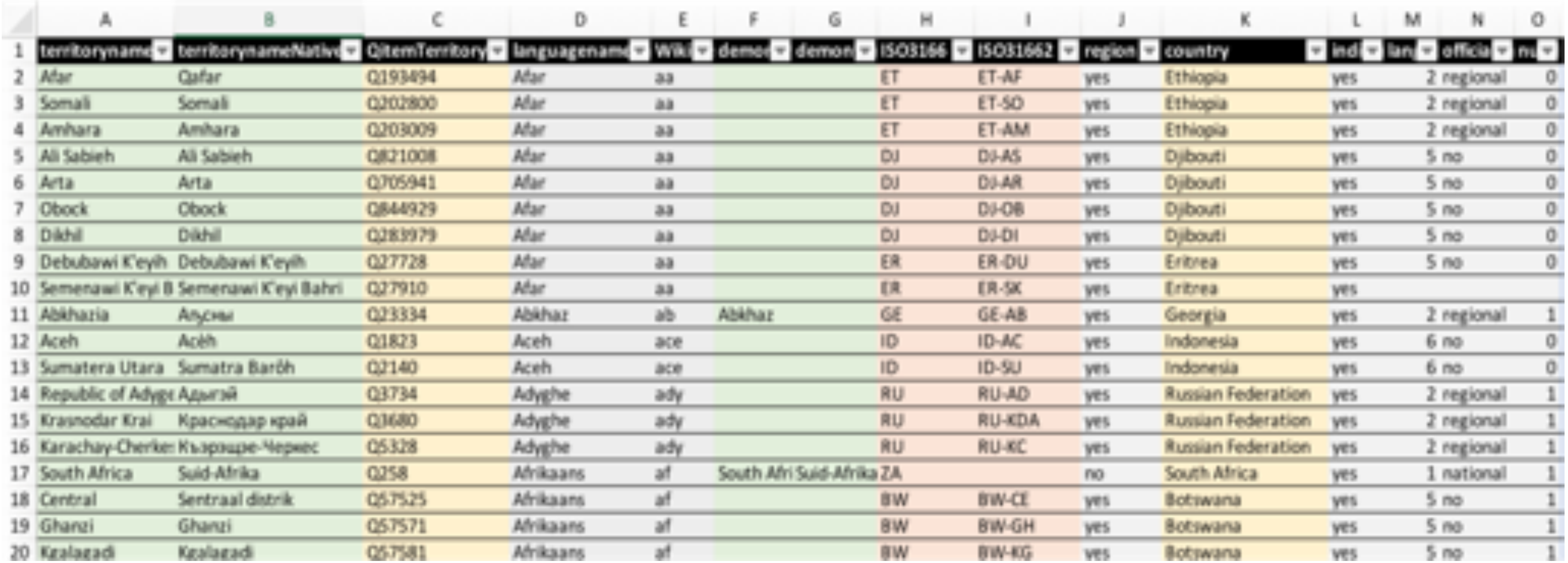
- Cultural Context Content
- Topics
- Quality Articles

We store the “root” data in a database named diversity groups. It contains all the values for each category.

https://wcdo.wmflabs.org/databases/diversity_groups_production.db

https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory/Glossary

Language-Territories Mapping is essential to obtain CCC



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	territoryname	territorynameNative	QitemTerritory	languagename	Wiki	demo	demon	ISO3166	ISO31662	region	country	ind	lan	officia	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	Аҧсны	Q23334	Abkhaz	ab	Abkhaz		GE	GE-AB	yes	Georgia	yes	2	regional	1
12	Aceh	Acèh	Q1823	Aceh	ace			ID	ID-AC	yes	Indonesia	yes	6	no	0
13	Sumatera Utara	Sumatra Barôh	Q2140	Aceh	ace			ID	ID-SU	yes	Indonesia	yes	6	no	0
14	Republic of Adyge	Адыгэй	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	Къэрэщэ-Черкес	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

(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.

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**
- **Keywords to retrieve articlese and categories**
- **Machine Learning for “Local Content”**

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

Wikipedia Diversity Database

One table for each Wikipedia language edition and one row per article.

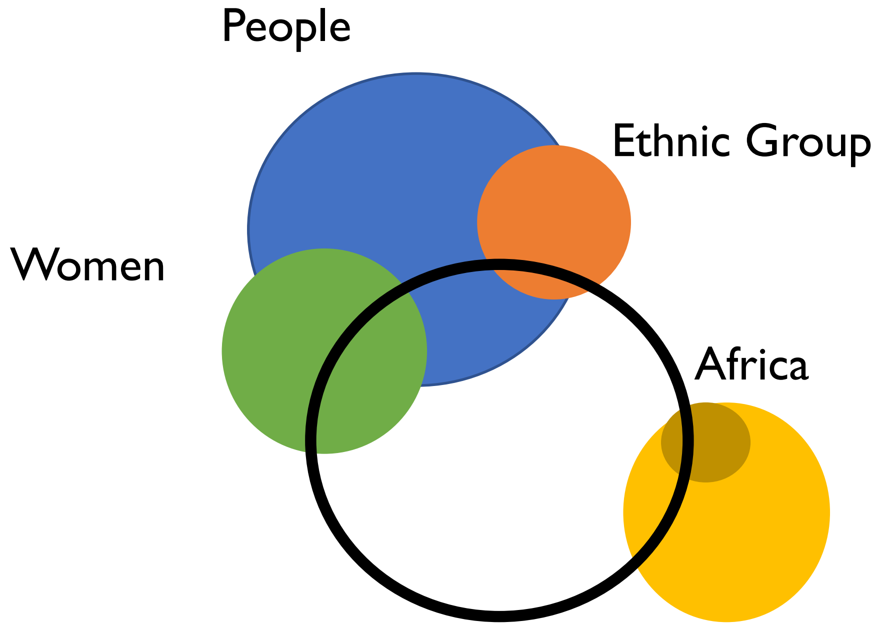
- General features (page_id, qitem, date created, first language created,…)
- Geography (geocoordinates, ISO3166, ISO31662, continent,…)
- People (gender, ethnic_group, supra_ethnic_group, sexual_orientation…)
- Cultural Context (binary, ccc,…)
- CCC features (country_wd, location_wd, language_strong_wd,…)
- General Topics (folk, earth, monuments, music, sports and teams, industry, books, glam,…)
- Ethnic Group Topics
- LGBT Topics
- Relevance Characteristics (num_inlinks, num_outlinks, num_bytes, num_refernces, …)
- Quality (featured_article, wikirank)

https://wcdo.wmflabs.org/databases/wikipedia_diversity_production.db

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.

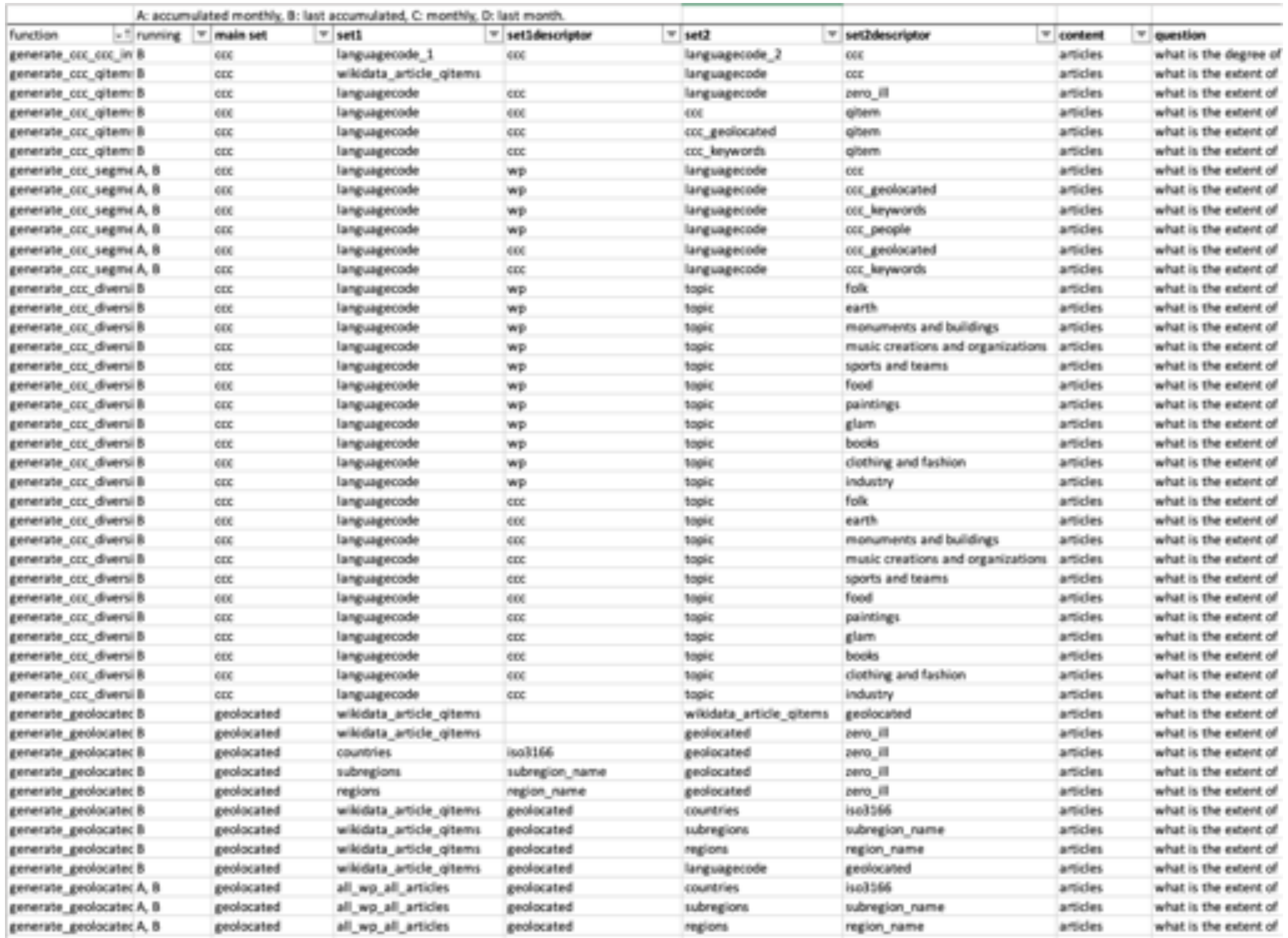


German Wikipedia
Accumulated articles May 2020

Stats Database

History log of content diversity and gaps in Wikipedia language editions.

This Excel file shows almost 200 types of intersections.
The result in relative (percentage) and absolute number of X (articles).



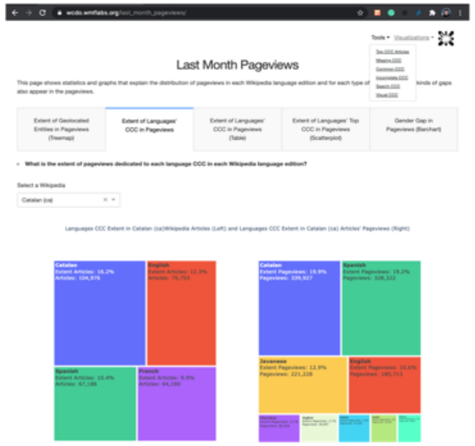
A: accumulated monthly, B: last accumulated, C: monthly, D: last month.

function	running	main set	set1	set1descriptor	set2	set2descriptor	content	question
generate_ccc_ccc_in	B	ccc	languagecode_1	ccc	languagecode_2	ccc	articles	what is the degree of
generate_ccc_qitem	B	ccc	wikidata_article_qitems		languagecode	ccc	articles	what is the extent of
generate_ccc_qitem	B	ccc	languagecode	ccc	languagecode	zero_ill	articles	what is the extent of
generate_ccc_qitem	B	ccc	languagecode	ccc	ccc	qitem	articles	what is the extent of
generate_ccc_qitem	B	ccc	languagecode	ccc	ccc_geolocated	qitem	articles	what is the extent of
generate_ccc_qitem	B	ccc	languagecode	ccc	ccc_keywords	qitem	articles	what is the extent of
generate_ccc_segme	A, B	ccc	languagecode	wp	languagecode	ccc	articles	what is the extent of
generate_ccc_segme	A, B	ccc	languagecode	wp	languagecode	ccc_geolocated	articles	what is the extent of
generate_ccc_segme	A, B	ccc	languagecode	wp	languagecode	ccc_keywords	articles	what is the extent of
generate_ccc_segme	A, B	ccc	languagecode	wp	languagecode	ccc_people	articles	what is the extent of
generate_ccc_segme	A, B	ccc	languagecode	ccc	languagecode	ccc_geolocated	articles	what is the extent of
generate_ccc_segme	A, B	ccc	languagecode	ccc	languagecode	ccc_keywords	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	folk	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	earth	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	monuments and buildings	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	music creations and organizations	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	sports and teams	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	food	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	paintings	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	glam	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	books	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	clothing and fashion	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	wp	topic	industry	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	folk	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	earth	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	monuments and buildings	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	music creations and organizations	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	sports and teams	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	food	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	paintings	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	glam	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	books	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	clothing and fashion	articles	what is the extent of
generate_ccc_diversi	B	ccc	languagecode	ccc	topic	industry	articles	what is the extent of
generate_geolocatec	B	geolocated	wikidata_article_qitems		wikidata_article_qitems	geolocated	articles	what is the extent of
generate_geolocatec	B	geolocated	wikidata_article_qitems		geolocated	zero_ill	articles	what is the extent of
generate_geolocatec	B	geolocated	countries	iso3166	geolocated	zero_ill	articles	what is the extent of
generate_geolocatec	B	geolocated	subregions	subregion_name	geolocated	zero_ill	articles	what is the extent of
generate_geolocatec	B	geolocated	regions	region_name	geolocated	zero_ill	articles	what is the extent of
generate_geolocatec	B	geolocated	wikidata_article_qitems	geolocated	countries	iso3166	articles	what is the extent of
generate_geolocatec	B	geolocated	wikidata_article_qitems	geolocated	subregions	subregion_name	articles	what is the extent of
generate_geolocatec	B	geolocated	wikidata_article_qitems	geolocated	regions	region_name	articles	what is the extent of
generate_geolocatec	B	geolocated	wikidata_article_qitems	geolocated	languagecode	geolocated	articles	what is the extent of
generate_geolocatec	A, B	geolocated	all_wp_all_articles	geolocated	countries	iso3166	articles	what is the extent of
generate_geolocatec	A, B	geolocated	all_wp_all_articles	geolocated	subregions	subregion_name	articles	what is the extent of
generate_geolocatec	A, B	geolocated	all_wp_all_articles	geolocated	regions	region_name	articles	what is the extent of

https://github.com/marcmiquel/WCDO/blob/wcdo/docs/sets_intersections.xlsx
https://wcdo.wmflabs.org/databases/stats_production.db

3. We create the dashboards with visualizations and tools.

Website


wcdo.wmflabs.org/last_month_pageviews/
Tools
Visualizations
Top CCC Articles
Missing CCC
Common CCC
Incomplete CCC
Search CCC
Visual CCC

Last Month Pageviews

This page shows statistics and graphs that explain the distribution of pageviews in each Wikipedia language edition and for each type of [illegible] kinds of gaps also appear in the pageviews.

Extent of Geolocated Entities in Pageviews (Treemap)
Extent of Languages' CCC in Pageviews
Extent of Languages' CCC in Pageviews (Table)
Extent of Languages' Top CCC in Pageviews (Scatterplot)
Gender Gap in Pageviews (Barchart)

- What is the extent of pageviews dedicated to each language CCC in each Wikipedia language edition?

Select a Wikipedia: Catalan (ca)

Languages CCC Extent in Catalan (ca)Wikipedia Articles (Left) and Languages CCC Extent in Catalan (ca) Articles' Pageviews (Right)

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

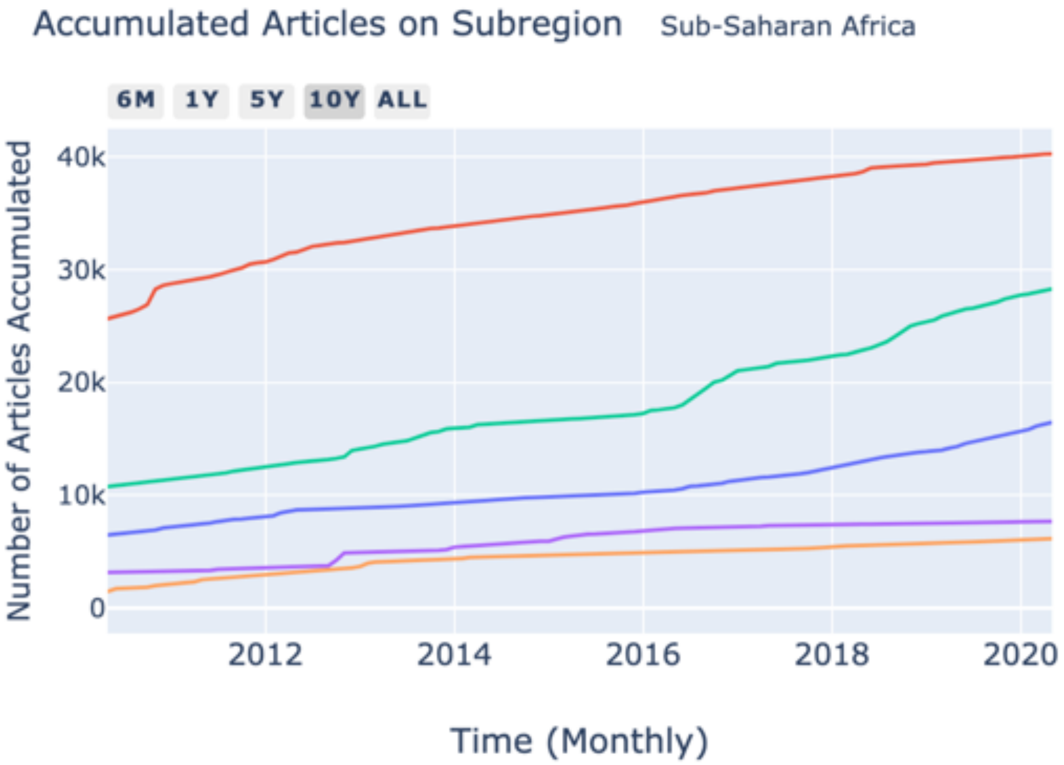
Dashboards: Visualizations and Tools

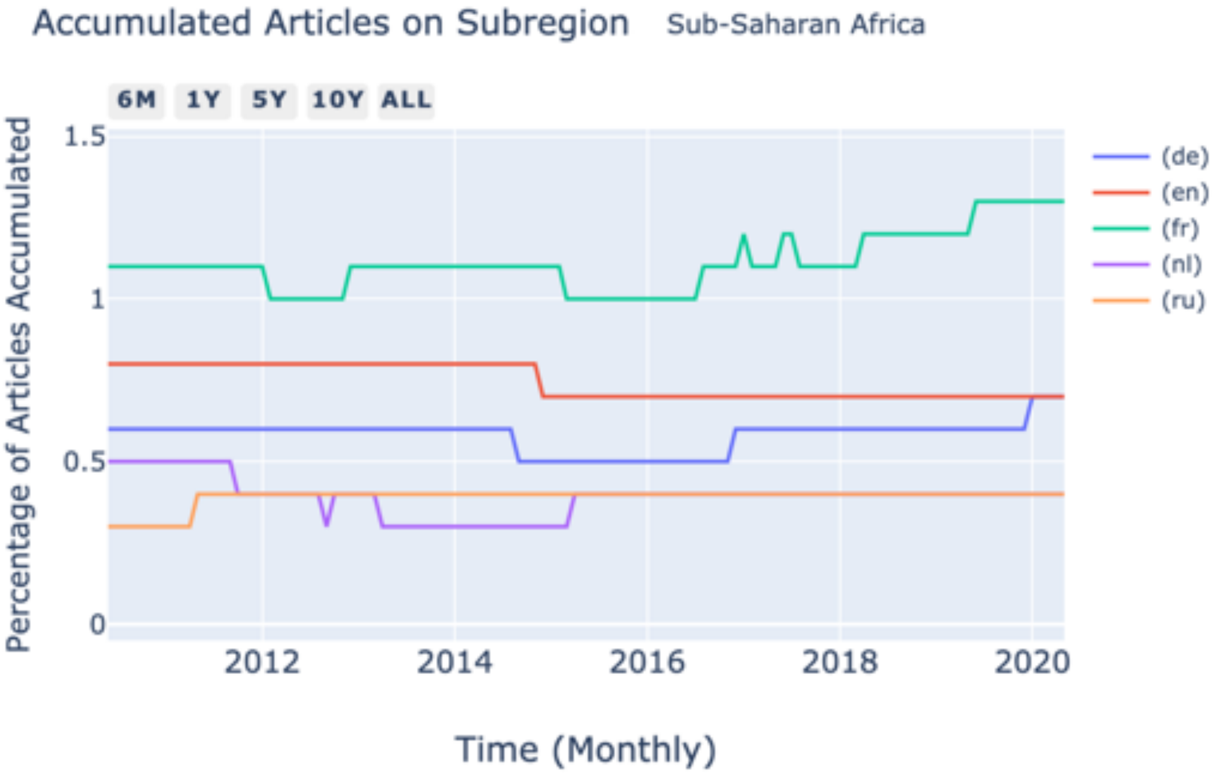
The Wikipedia
Diversity Observatory

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

(Geography Gap)

Diversity Over Time





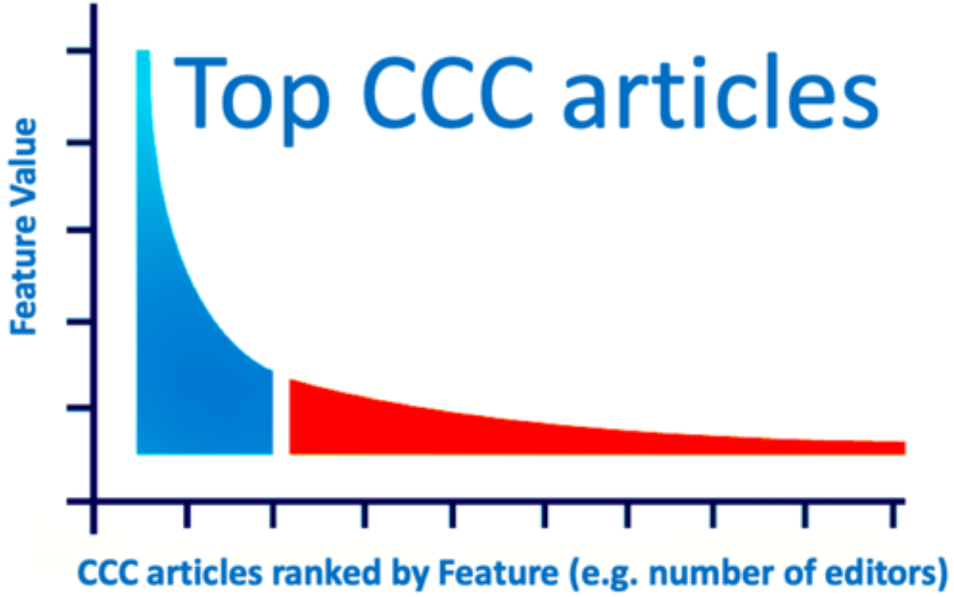
Has Wikimania been useful to encourage the creation of articles geolocated in Subsaharan Africa?

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

Dashboards:
Visualizations and Tools

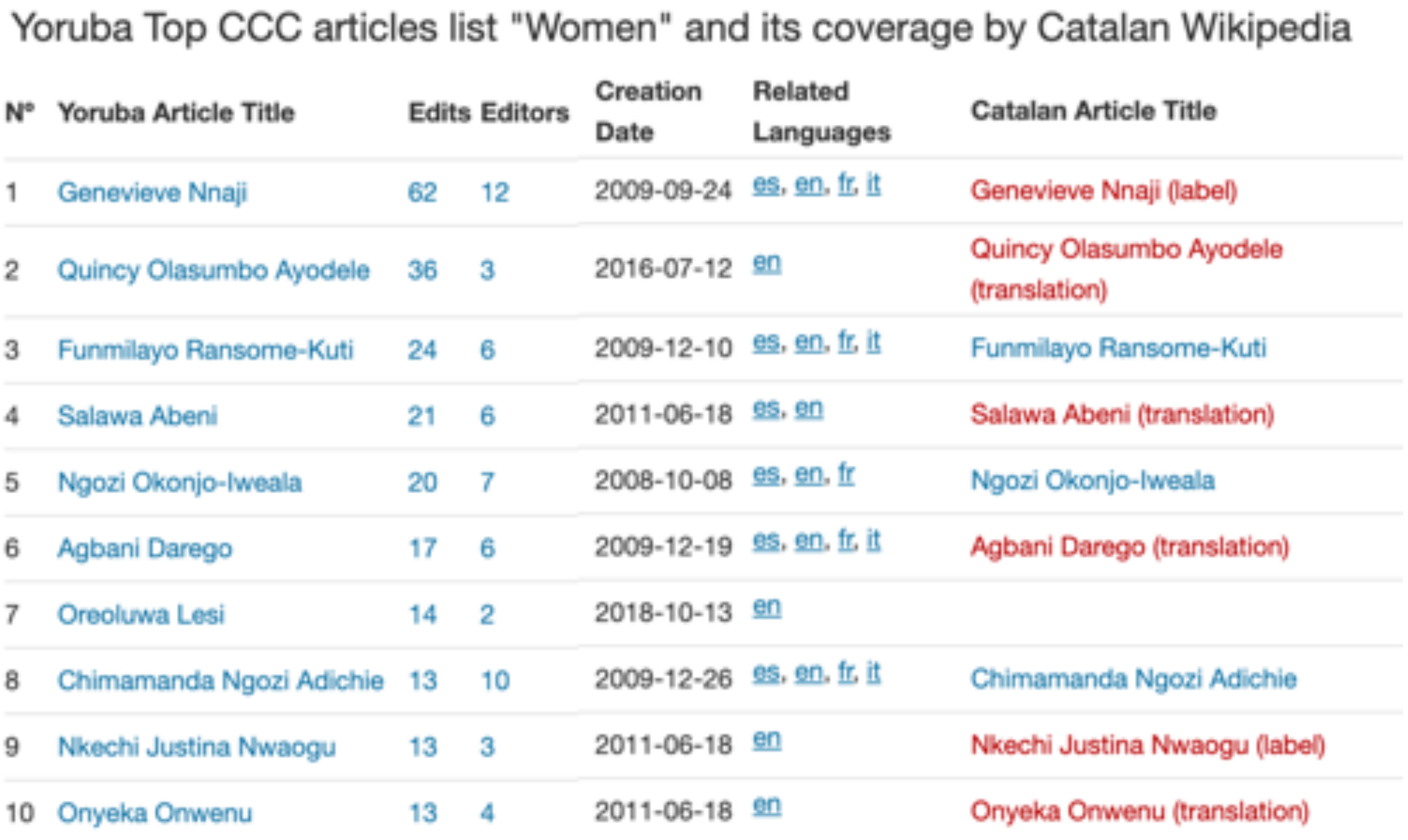
The Wikipedia
Diversity Observatory

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

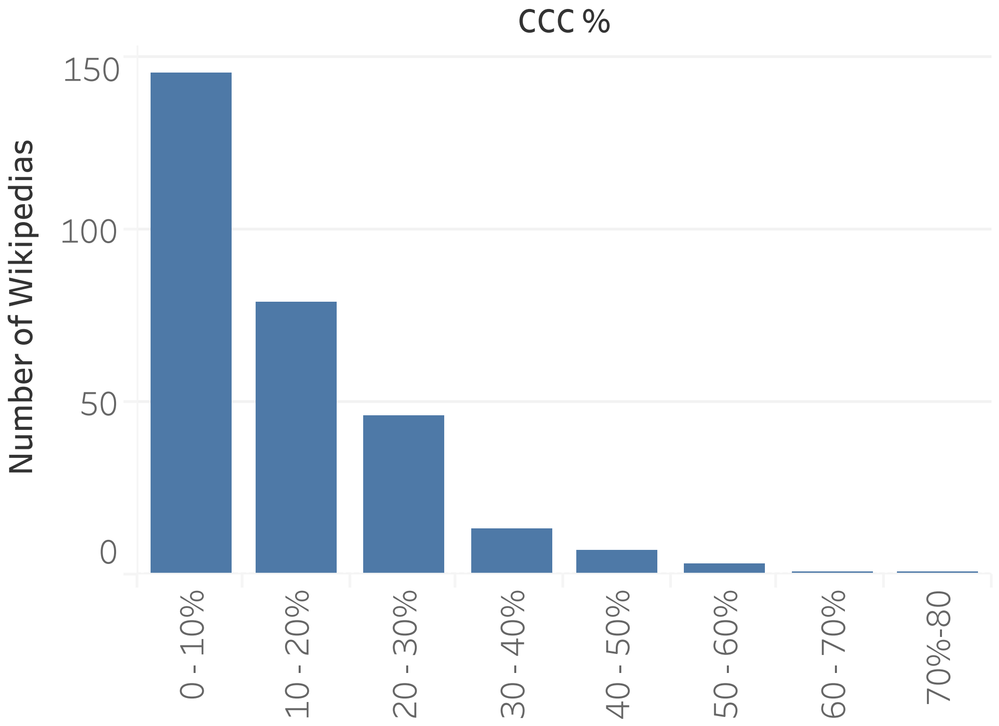


Yoruba Top CCC articles list "Women" and its coverage by Catalan Wikipedia

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

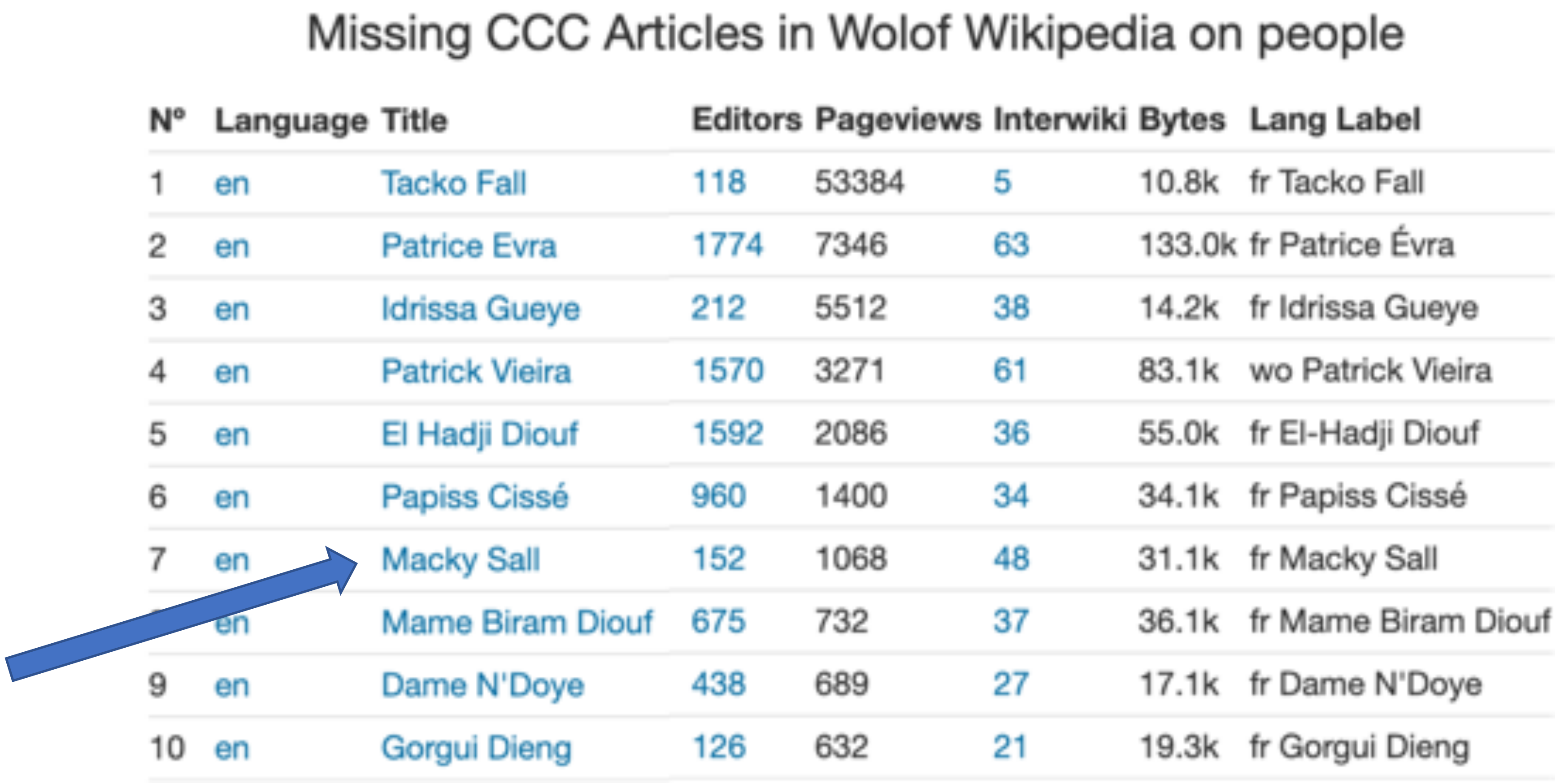
Anna Rita del Piano,
Italian theatre actress

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

One Language Case Studies

The State of Cultural Diversity in Arabic Wikipedia: Challenges and Opportunities

Marc Miquel, PhD

(marcmiquel@gmail.com)

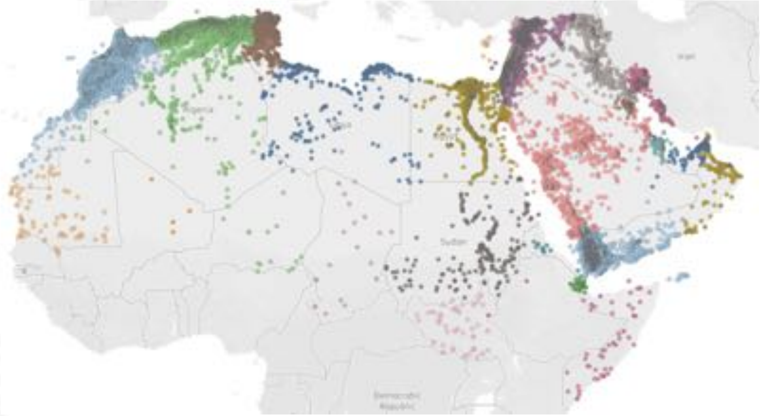
Username:marcmiquel

Pompeu Fabra University, Barcelona, **Catalonia**

Catalan Wikipedia, **Amical Wikimedia**

Marrakesh, 4-6 October 2019

In-depth Arabic content diversity 45 min presentation in Wikiarabia 2019



https://commons.wikimedia.org/wiki/File:The_State_of_Cultural_Diversity_in_Arabic_Wikipedia_2019.pdf

Solutions to improve coverage and spread

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

For every category relevant to diversity.

👥 ♀ ✡☯✝☪ 🌍 🌈

Reflections

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.

Current and future steps

Research and Tools

- Dashboards for the “time gap”, “diversity essentials lists”, etc.
- Recent Changes Diversity dashboard (monitoring articles by diversity category)
- Contextualized Article Notability
- Editor Participation in Content Diversity across languages
- Editor diversity (surveys?)

Dissemination

- Changes in the Observatory Meta page to include many other tools.
- Contact communities and affiliates to explain the tools.

The Wikipedia Diversity Observatory

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

References (if you want to know more)

Miquel-Ribé, M., & Laniado, D. **(2020)**. The Wikipedia Diversity Observatory: A Project to Identify and Bridge Content Gaps in Wikipedia. *Proceedings of the International Symposium on Open Collaboration. ACM.*

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.

The Knowledge Gaps Project

Systems that identify, measure and address gaps across Wikimedia projects.

Motivation

Taxonomy

It is a comprehensive theoretical framework that can provide “ground reference” to see the evolution of the Movement for the next years.

Great opportunity to settle down a research and data-driven culture in the Wikimedia to grow larger and more diverse.

Alignment with the WDO

The Diversity Observatory encompasses research, tools and dissemination on content diversity, so far.

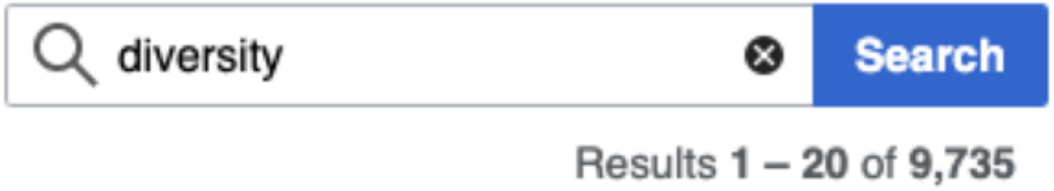
The Knowledge Gaps project aims at identifying and measuring both content, contributors and readers gaps.

Storytelling: Diversity / Gaps

- Diversity is "the condition of having or being composed of differing elements".
- Gap is a "difference between two things".
- Vision of "Sum of human knowledge"

Some people connect with "working towards diversity" and others with "bridging the X gap", even though they are doing the same thing.

Use both! Title, introductions, section, etc.



diversity

Results 1 – 20 of 9,735

Searching ”diversity” in meta you get 9,735 results

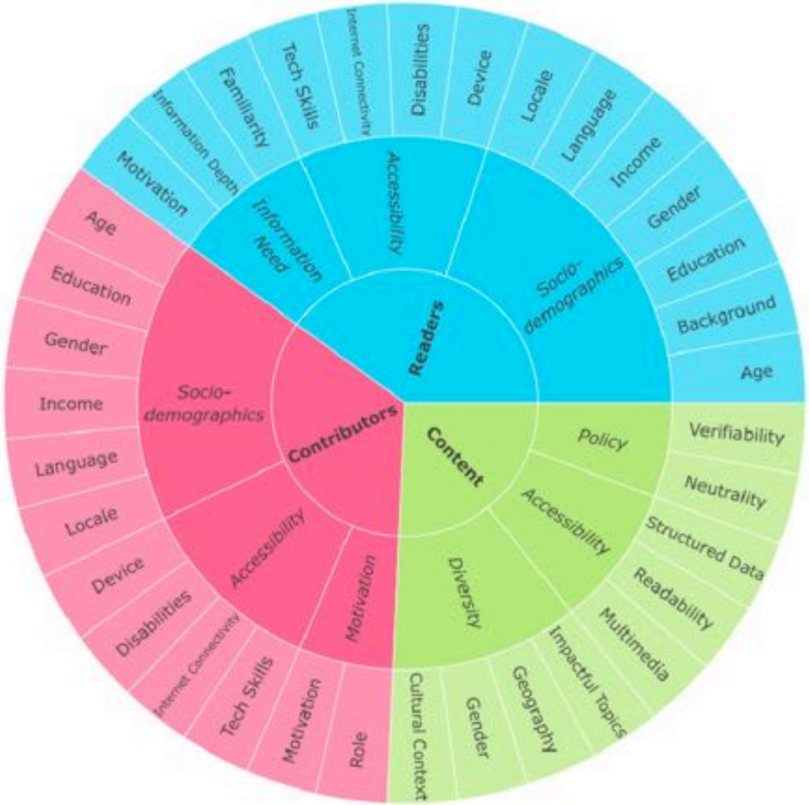
Searching “gap” you get 9,582 results

Searching ”barriers” you get 3,484 results

Use them all! Title, introductions, section, etc.

Theoretical Framework

Dimensions



Fantastic approach! I love it!

- **Readers**
- **Contributors**
- **Content**

Dimension > Facet > Gap

The more we bridge these gaps, the more diversity in each dimension.

Facets and Gaps

- The categories for the gaps must reflect the sources but also the intent.
- Use categories that communities or affiliates are working on (e.g. LGBT).
- Be consistent with the categories across dimensions to facilitate the readers' understanding and the analyses.

Special Gaps

- Language is in readers and contributors, but not in content (6,000 languages).
- Geography is in content, but not in readers and contributors.

Language is a special characteristics because projects are usually organized by it.

Geography is a special characteristic because people are usually related to space.

I suggest use in the every dimension because they are very useful to segment data.

Special Gaps

I suggest introducing Time as a gap in every dimension

- Content we can see the “recency bias” or the lack of articles about some centuries or ages.
- Contributors we can see the imbalances in contributors’ lifetime in the project (also the lack of retention).
- Readers we can see that population has not known or used Wikipedia or free knowledge for the same period of time (e.g. related to brand awareness).

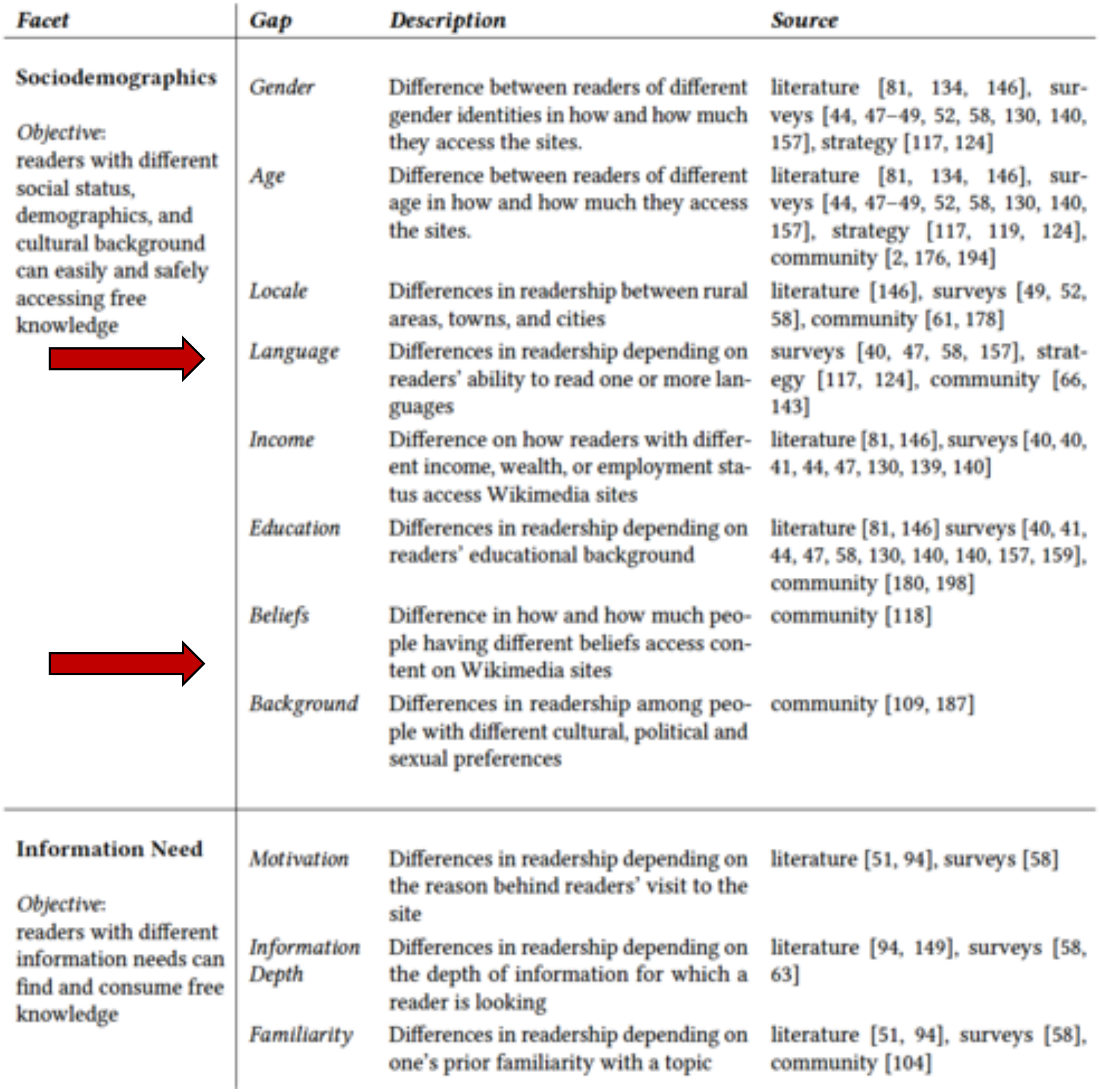
Readers

Language as the first gap. Because it is the one that relates to the project.

We have the facets “Background” (Readers) and “Additional Characteristics” (Contributors). They are similar with a different name.

Suggestion: Use three subcategories in the three dimensions.

- Cultural Groups
- Sexual Orientation
- Ethnicity



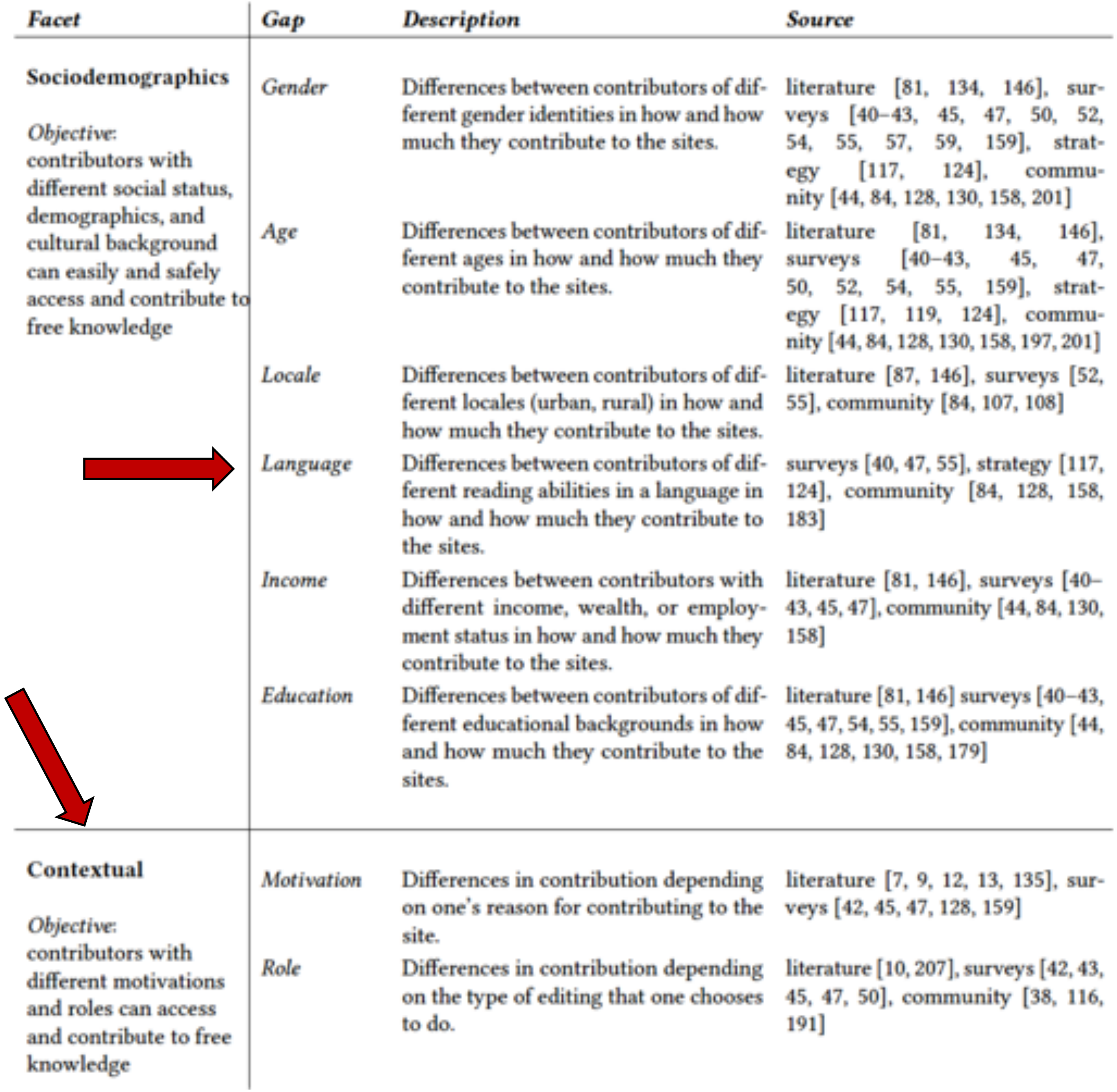
Facet	*Gap*	*Description*	*Source*
Sociodemographics *Objective:* readers with different social status, demographics, and cultural background can easily and safely accessing free knowledge	*Gender*	Difference between readers of different gender identities in how and how much they access the sites.	literature [81, 134, 146], surveys [44, 47–49, 52, 58, 130, 140, 157], strategy [117, 124]
	Age	Difference between readers of different age in how and how much they access the sites.	literature [81, 134, 146], surveys [44, 47–49, 52, 58, 130, 140, 157], strategy [117, 119, 124], community [2, 176, 194]
	Locale	Differences in readership between rural areas, towns, and cities	literature [146], surveys [49, 52, 58], community [61, 178]
	Language	Differences in readership depending on readers’ ability to read one or more languages	surveys [40, 47, 58, 157], strategy [117, 124], community [66, 143]
	Income	Difference on how readers with different income, wealth, or employment status access Wikimedia sites	literature [81, 146], surveys [40, 40, 41, 44, 47, 130, 139, 140]
	Education	Differences in readership depending on readers’ educational background	literature [81, 146] surveys [40, 41, 44, 47, 58, 130, 140, 140, 157, 159], community [180, 198]
	Beliefs	Difference in how and how much people having different beliefs access content on Wikimedia sites	community [118]
	Background	Differences in readership among people with different cultural, political and sexual preferences	community [109, 187]
Information Need *Objective:* readers with different information needs can find and consume free knowledge	*Motivation*	Differences in readership depending on the reason behind readers’ visit to the site	literature [51, 94], surveys [58]
	Information Depth	Differences in readership depending on the depth of information for which a reader is looking	literature [94, 149], surveys [58, 63]
	Familiarity	Differences in readership depending on one’s prior familiarity with a topic	literature [51, 94], surveys [58], community [104]

Contributors

Language as the first gap. Because it is the one that relates to the project.

I would change "contextual" for **Engagement**, and include more gaps that relate to:

- Motivation
- Local community role
- Participation in projects
- Affiliate / board member
- Multiproject / Meta
- Time gap



Facet	*Gap*	*Description*	*Source*
Sociodemographics *Objective:* contributors with different social status, demographics, and cultural background can easily and safely access and contribute to free knowledge	*Gender*	Differences between contributors of different gender identities in how and how much they contribute to the sites.	literature [81, 134, 146], surveys [40–43, 45, 47, 50, 52, 54, 55, 57, 59, 159], strategy [117, 124], community [44, 84, 128, 130, 158, 201]
	Age	Differences between contributors of different ages in how and how much they contribute to the sites.	literature [81, 134, 146], surveys [40–43, 45, 47, 50, 52, 54, 55, 159], strategy [117, 119, 124], community [44, 84, 128, 130, 158, 197, 201]
	Locale	Differences between contributors of different locales (urban, rural) in how and how much they contribute to the sites.	literature [87, 146], surveys [52, 55], community [84, 107, 108]
	Language	Differences between contributors of different reading abilities in a language in how and how much they contribute to the sites.	surveys [40, 47, 55], strategy [117, 124], community [84, 128, 158, 183]
	Income	Differences between contributors with different income, wealth, or employment status in how and how much they contribute to the sites.	literature [81, 146], surveys [40–43, 45, 47], community [44, 84, 130, 158]
	Education	Differences between contributors of different educational backgrounds in how and how much they contribute to the sites.	literature [81, 146] surveys [40–43, 45, 47, 54, 55, 159], community [44, 84, 128, 130, 158, 179]
Contextual *Objective:* contributors with different motivations and roles can access and contribute to free knowledge	*Motivation*	Differences in contribution depending on one's reason for contributing to the site.	literature [7, 9, 12, 13, 135], surveys [42, 45, 47, 128, 159]
	Role	Differences in contribution depending on the type of editing that one chooses to do.	literature [10, 207], surveys [42, 43, 45, 47, 50], community [38, 116, 191]

Content

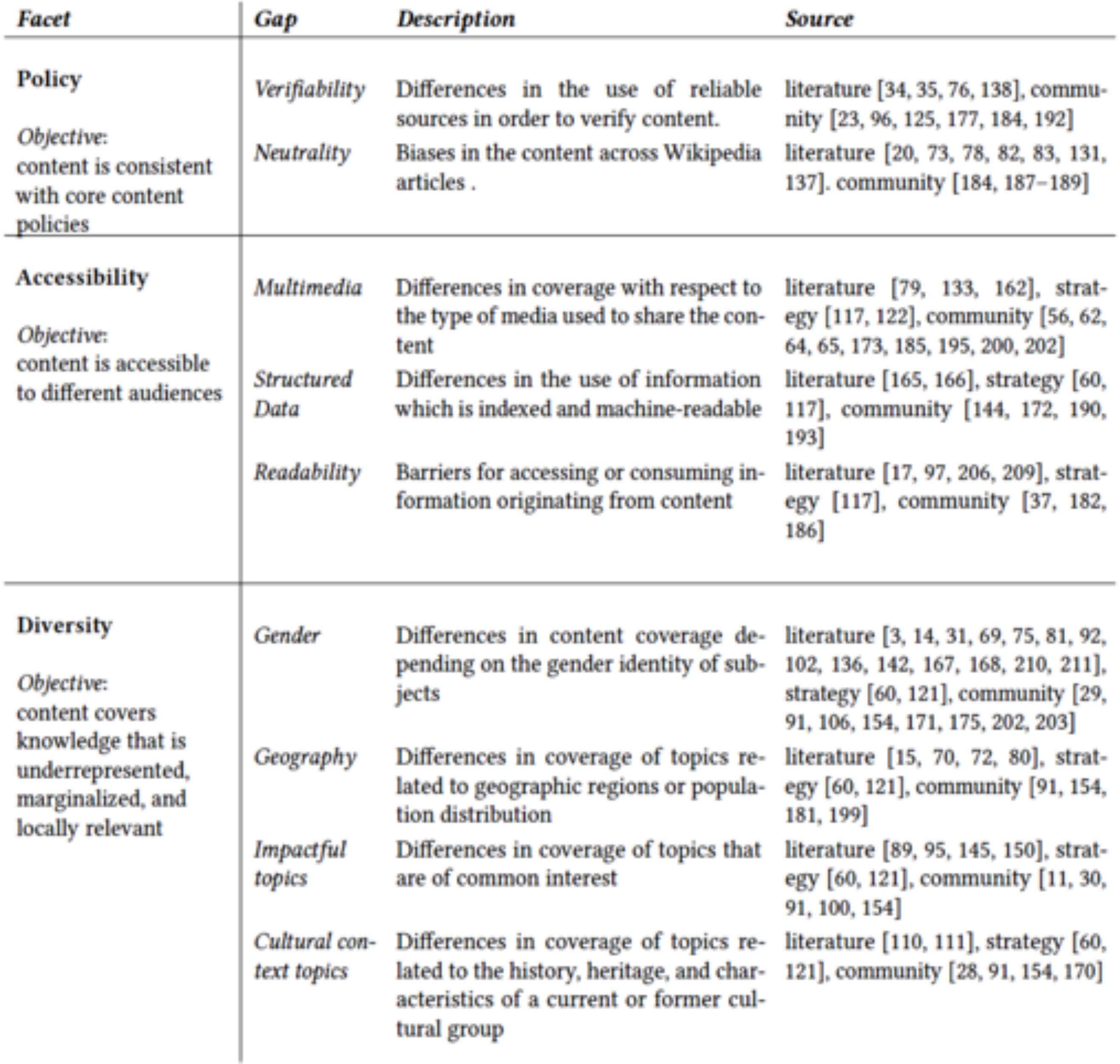
Suggest to change diversity and use **Topic** and put it as the first facet. Include the three people gaps mentioned earlier.

Impactful topics first or last, as it is a category multitopic.

Suggest to change accessibility and call it **Content characteristics**. Accessibility is used in a different way.

In policy, maybe we could introduce Notability as a gap.

In the content section, I would mention that topic applies to both article and in-article levels.



Facet	***Gap***	***Description***	***Source***
Policy *Objective*: content is consistent with core content policies	*Verifiability*	Differences in the use of reliable sources in order to verify content.	literature [34, 35, 76, 138], community [23, 96, 125, 177, 184, 192]
	Neutrality	Biases in the content across Wikipedia articles .	literature [20, 73, 78, 82, 83, 131, 137]. community [184, 187–189]
Accessibility *Objective*: content is accessible to different audiences	*Multimedia*	Differences in coverage with respect to the type of media used to share the content	literature [79, 133, 162], strategy [117, 122], community [56, 62, 64, 65, 173, 185, 195, 200, 202]
	Structured Data	Differences in the use of information which is indexed and machine-readable	literature [165, 166], strategy [60, 117], community [144, 172, 190, 193]
	Readability	Barriers for accessing or consuming information originating from content	literature [17, 97, 206, 209], strategy [117], community [37, 182, 186]
Diversity *Objective*: content covers knowledge that is underrepresented, marginalized, and locally relevant	*Gender*	Differences in content coverage depending on the gender identity of subjects	literature [3, 14, 31, 69, 75, 81, 92, 102, 136, 142, 167, 168, 210, 211], strategy [60, 121], community [29, 91, 106, 154, 171, 175, 202, 203]
	Geography	Differences in coverage of topics related to geographic regions or population distribution	literature [15, 70, 72, 80], strategy [60, 121], community [91, 154, 181, 199]
	Impactful topics	Differences in coverage of topics that are of common interest	literature [89, 95, 145, 150], strategy [60, 121], community [11, 30, 91, 100, 154]
	Cultural context topics	Differences in coverage of topics related to the history, heritage, and characteristics of a current or former cultural group	literature [110, 111], strategy [60, 121], community [28, 91, 154, 170]

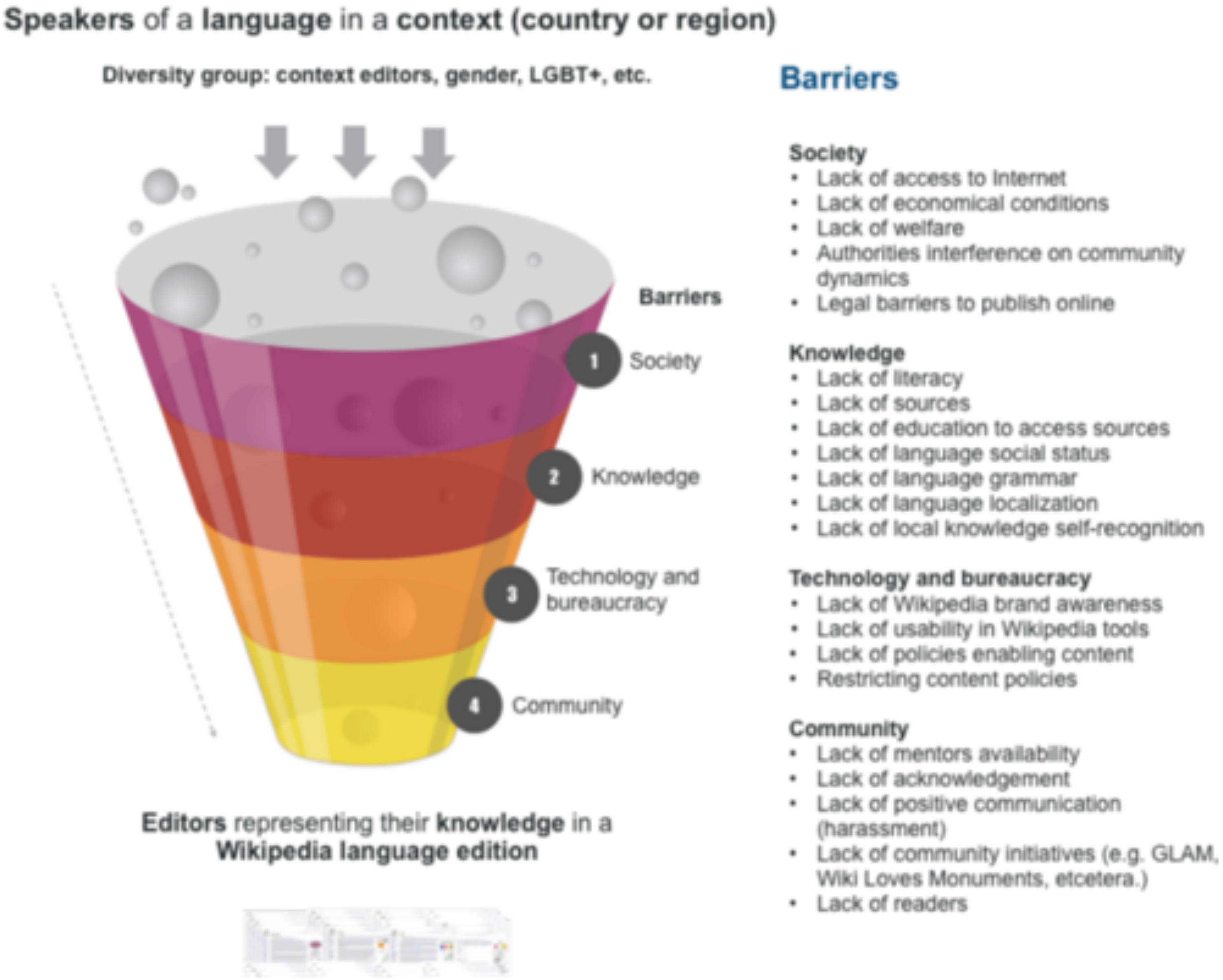
Gaps or Barriers?

The Knowledge
Gaps Project

Idea to play with (Conversion Funnel)

Lightning talk 2019

WIKIMANIA
STOCKHOLM
AUGUST 14-18, 2019



Speakers of a language in a context (country or region)

Diversity group: context editors, gender, LGBT+, etc.

Editors representing their knowledge in a Wikipedia language edition

Barriers

Society
- Lack of access to Internet
- Lack of economical conditions
- Lack of welfare
- Authorities interference on community dynamics
- Legal barriers to publish online

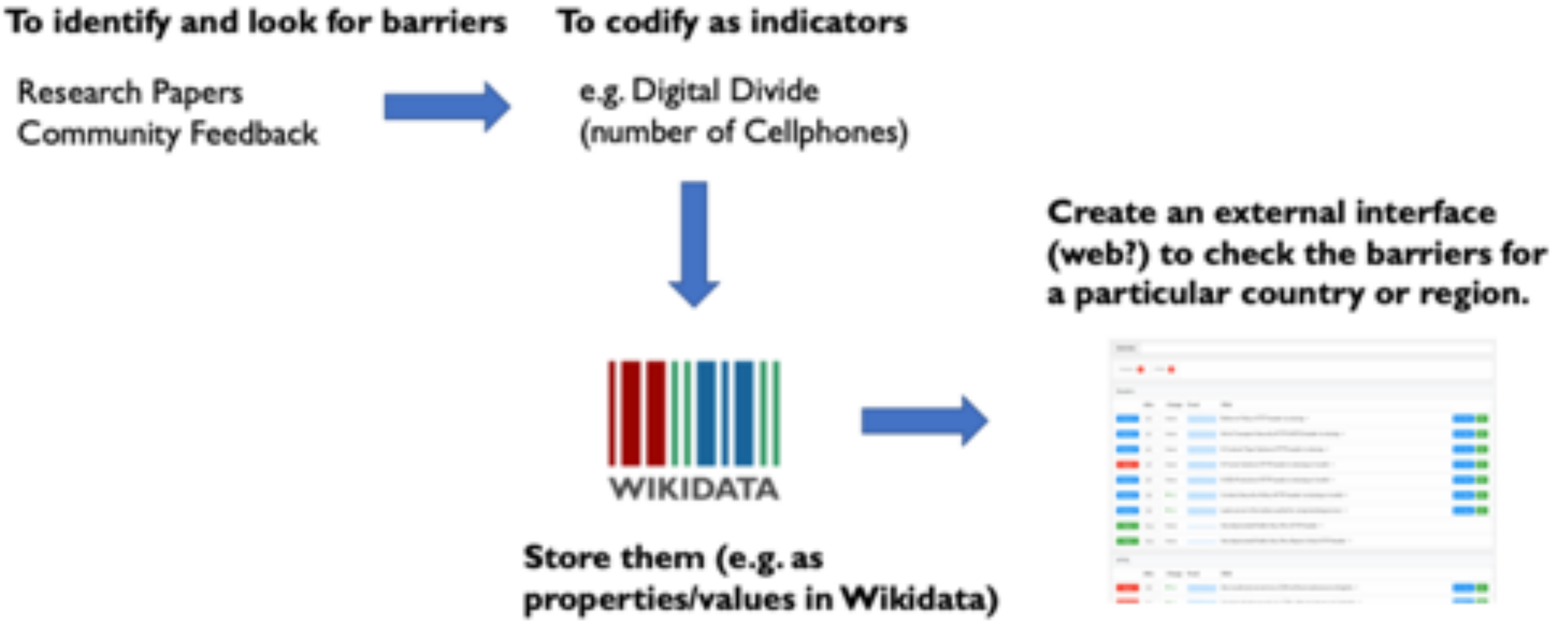
Knowledge
- Lack of literacy
- Lack of sources
- Lack of education to access sources
- Lack of language social status
- Lack of language grammar
- Lack of language localization
- Lack of local knowledge self-recognition

Technology and bureaucracy
- Lack of Wikipedia brand awareness
- Lack of usability in Wikipedia tools
- Lack of policies enabling content
- Restricting content policies

Community
- Lack of mentors availability
- Lack of acknowledgement
- Lack of positive communication (harassment)
- Lack of community initiatives (e.g. GLAM, Wiki Loves Monuments, etcetera.)
- Lack of readers

https://commons.wikimedia.org/wiki/File:Mmiquel_wikimania2019_research_wcdo.pdf

Identifying the Wikimedia Editing and Community Diversity Barriers for Users in Each Country and Introduce Them in Wikidata



To identify and look for barriers

Research Papers
Community Feedback

To codify as indicators

e.g. Digital Divide
(number of Cellphones)

Store them (e.g. as properties/values in Wikidata)

Create an external interface (web?) to check the barriers for a particular country or region.

“We also have to directly **address the barriers** and circumstances that prevent people from utilizing or participating in our Movement.”

Introduction

“For everyone to feel welcome, we will embrace diversity and actively work to **remove barriers** to improve the user experience for consumption and contribution of free knowledge.”

Narrative of the recommendations

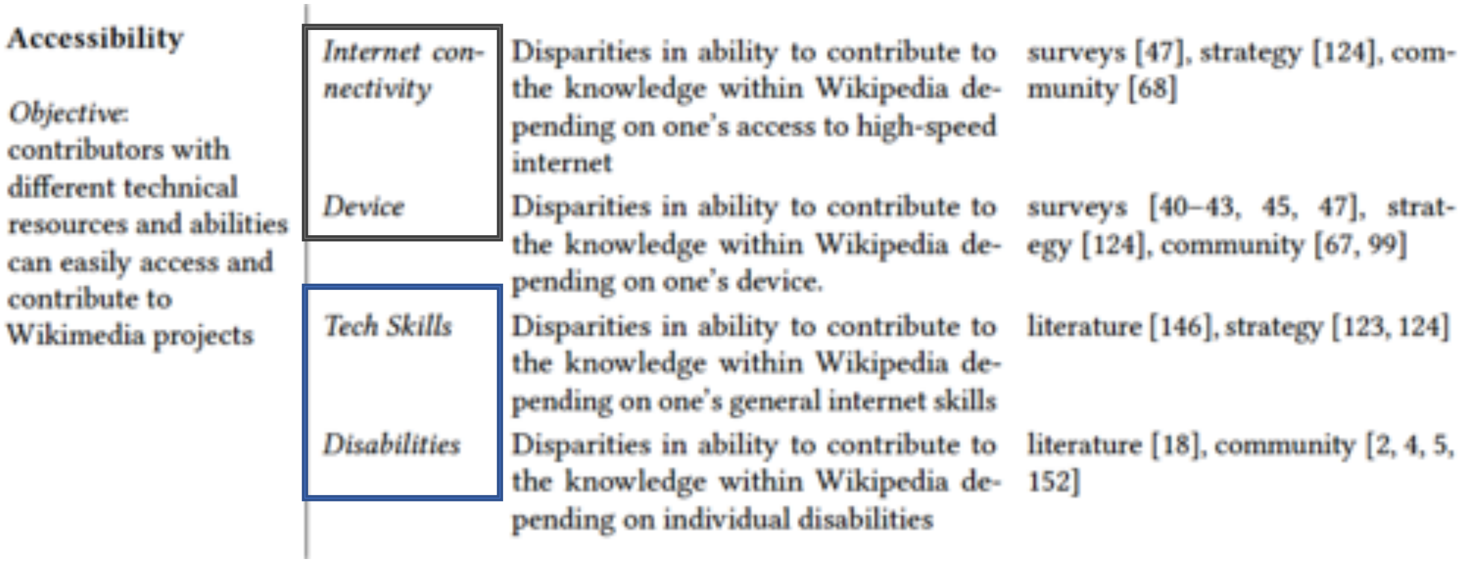
“We will break down the social, political, and technical barriers preventing people from accessing and contributing to free knowledge.”

Strategic direction

https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20/Recommendations

Rethinking the Accessibility Facet

Accessibility gaps contain gaps that belong to the person and others that relate to the context.



Accessibility *Objective:* contributors with different technical resources and abilities can easily access and contribute to Wikimedia projects	*Internet connectivity*	Disparities in ability to contribute to the knowledge within Wikipedia depending on one's access to high-speed internet	surveys [47], strategy [124], community [68]
	Device	Disparities in ability to contribute to the knowledge within Wikipedia depending on one's device.	surveys [40–43, 45, 47], strategy [124], community [67, 99]
	Tech Skills	Disparities in ability to contribute to the knowledge within Wikipedia depending on one's general internet skills	literature [146], strategy [123, 124]
	Disabilities	Disparities in ability to contribute to the knowledge within Wikipedia depending on individual disabilities	literature [18], community [2, 4, 5, 152]

Gaps that relate to a person are **"outcome gaps",** because we want to include everyone. Instead, gaps relate to the context are **"medium gaps"** usually known as barriers.

Suggestion: organize some of the accessibility gaps in some other way and call this facet "human-computer".

Structural and Context Gaps

Accessibility for both contributors and readers

# Wikimedia Structure

- Usability Gap
- Mentorship Gap
- Safety gap
- …

# Geographical context

- Security gap
- Freedom of speech gap
- GDP gap
- Language status gap
- Internet connectivity gap
- …

We may have an overlapping between “Wikimedia structure” and “Contributors”

Fourth Dimension “Medium” with Context and Structure Facets

Medium

Reasons for a Fourth dimension on barriers

- **Outcome gaps set the direction.** These other barriers are the "cause". It is easy to work on them if you highlight them as a group together.

- **Wording that connects.** People call them barrier. Gaps in Accessibility are barriers.

- **Measure the context when you cannot measure the gap.** When you cannot find data about contributors/readers, you need to understand those who are not in.

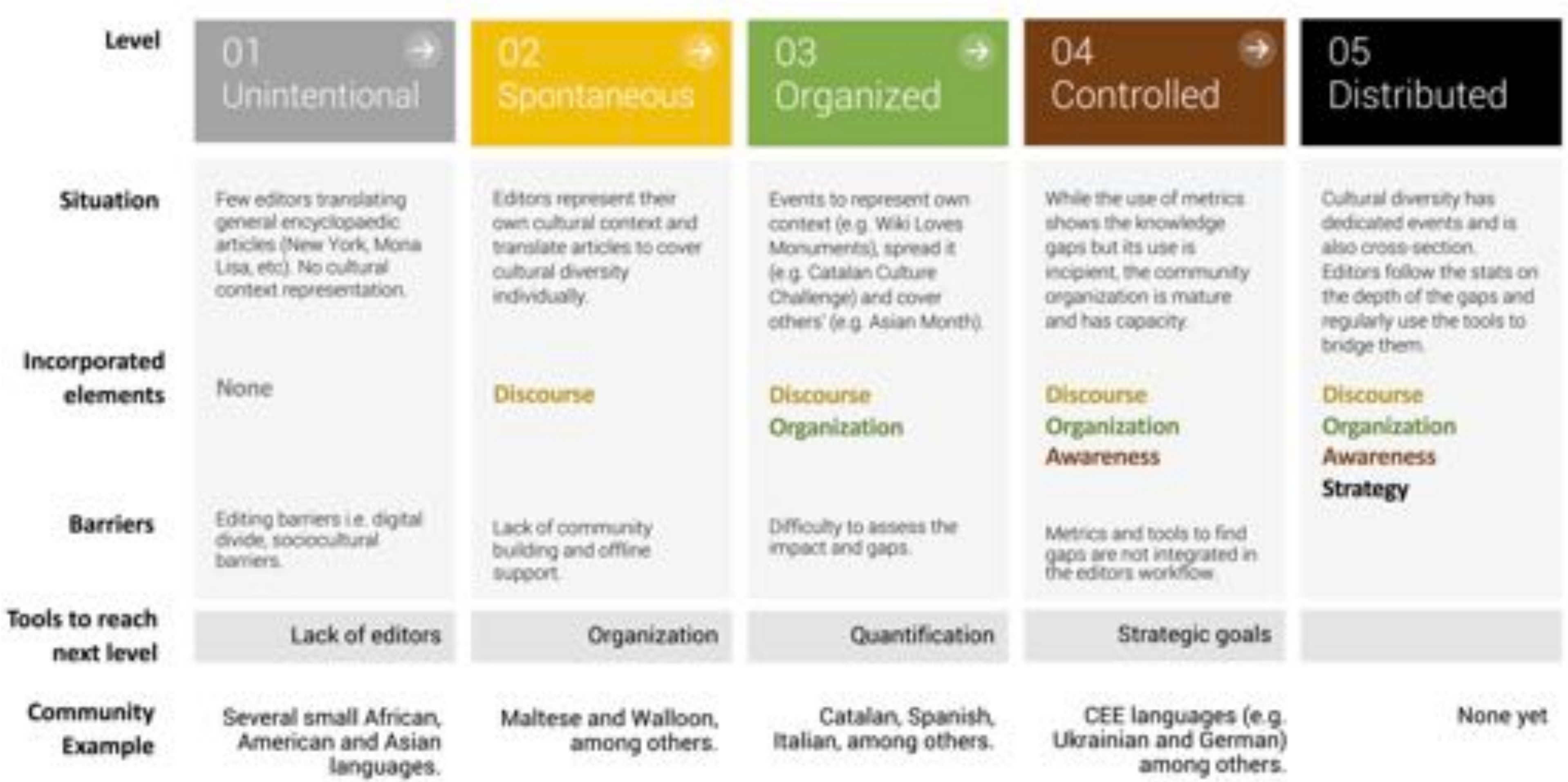
"You need to know the gaps in the world before you measure them in Wikimedia".

- **Engage people across Wikimedia.** You can engage the whole movement and ask them to complete the data.

Engage WMF "product" in creating usability metrics over time; engage communities in getting indicators for context (e.g. upload UN education levels by country in Wikidata).

- **Types of data.** Context data is generated by external agents. Internal barriers data is generated by WMF depts. or communities.

Example: Content Diversity Maturity Levels (Content/Contributors/Medium)

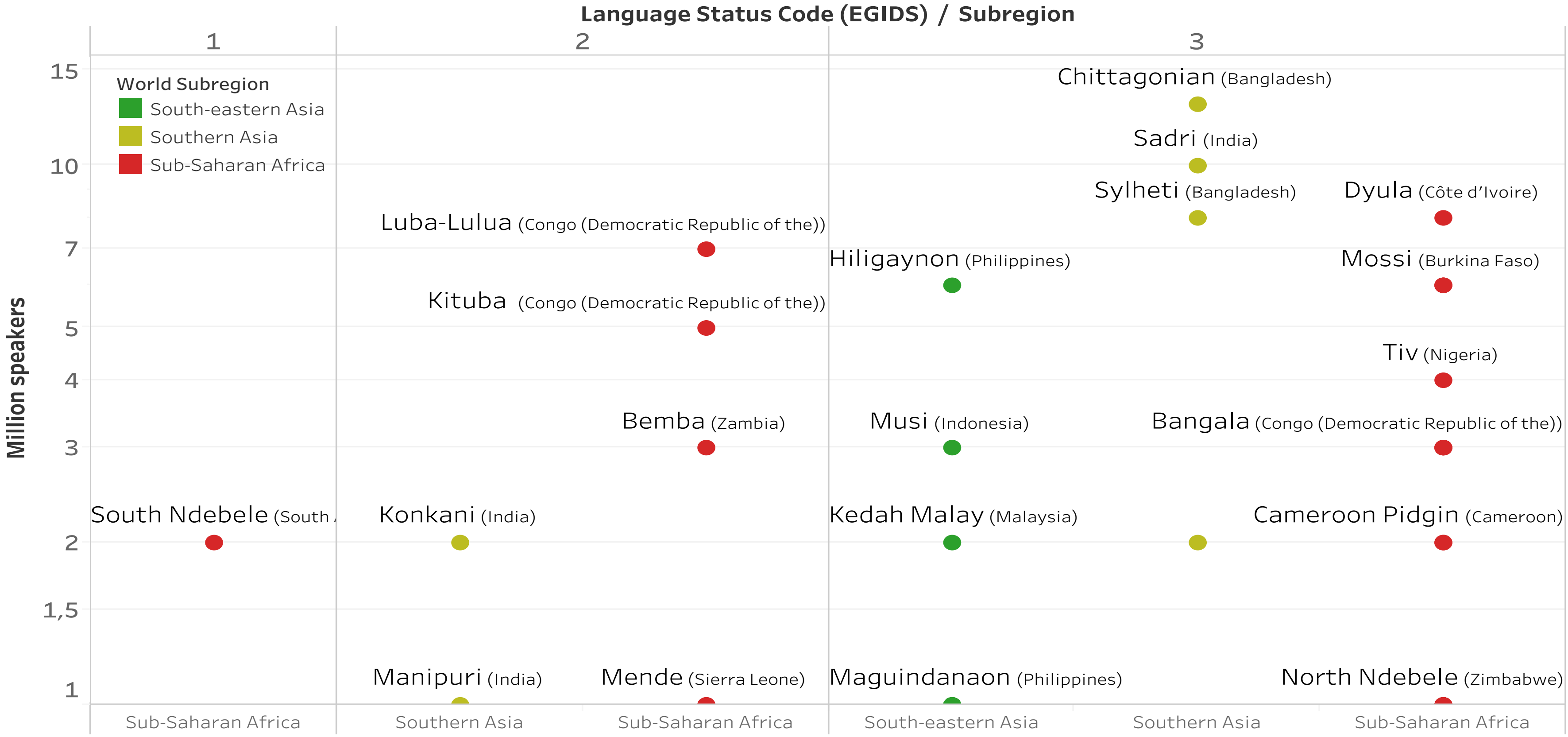


Level	01 Unintentional	02 Spontaneous	03 Organized	04 Controlled	05 Distributed
Situation	Few editors translating general encyclopaedic articles (New York, Mona Lisa, etc). No cultural context representation.	Editors represent their own cultural context and translate articles to cover cultural diversity 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	Several small African, American and Asian languages.	Maltese and Walloon, among others.	Catalan, Spanish, Italian, among others.	CEE languages (e.g. Ukrainian and German) among others.	None yet

https://wikipedia20.pubpub.org/pub/26ke5md7/release/15
https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2020-05-31/Special_report

Example: Potential Languages (Content/Medium)

We could detect potential languages that can become Wikipedias by observing the number of speakers and the language status code (EGIDS).

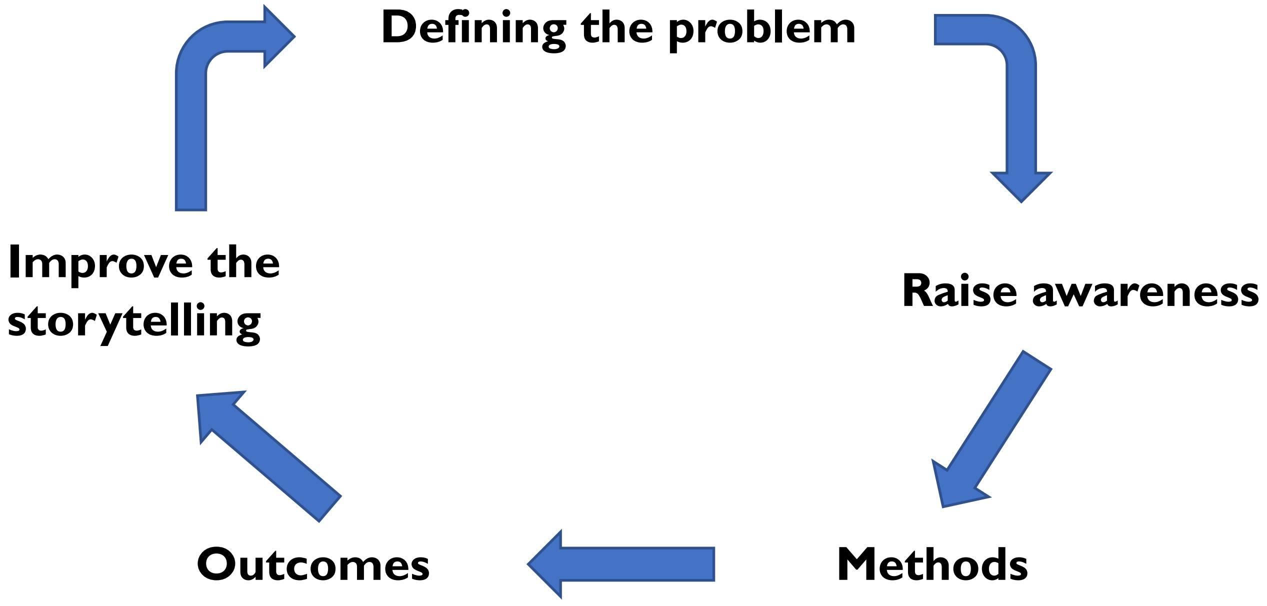


https://commons.wikimedia.org/wiki/File:Languages_Matter_to_Cultural_Diversity_Finding_Missing_Languages_and_Bridging_the_Gaps_in_Minority_Languages.pdf

Use-Cases for the Taxonomy

Any stakeholder is a potential user of the taxonomy and its derived indexes, visualizations and tools

Improving every part of the process



Defining the problem

Raise awareness

Methods

Outcomes

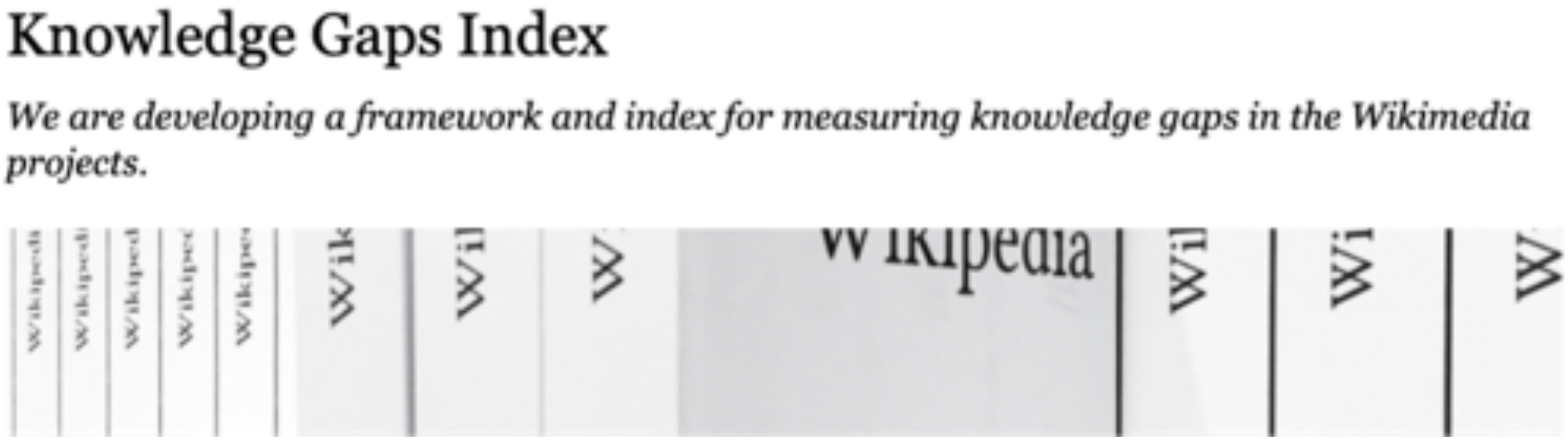
Improve the storytelling

Usability of the outcomes:

- Paper
- Video
- Database / Dataset
- Index or metric
- Visualization
- Tool with list of actions
- Wikitext table

Some outcomes raise the awareness, others are actionable

An Index



Knowledge Gaps Index

We are developing a framework and index for measuring knowledge gaps in the Wikimedia projects.

The diversity of stakeholders and interests is so high that it is reasonable to have an index for every facet, and if possible, an index for each project dimension (contributors, readers and content).

The Knowledge Gaps Project

Systems that identify, measure and address
gaps across Wikimedia projects.

Conclusions

The Wikipedia
Diversity Observatory

The Knowledge
Gaps Project

Two sides of the same coin

Defining categories, retrieving data, building indicators, visualizations and tools

Disseminating the results, showing the tools, understanding each stakeholder processes

Opportunities

- Great opportunity to make Wikimedia more data-driven, self-aware and dynamic to overcome its challenges.

- Totally aligned with Wikimedia Strategy recommendations – especially number 10, "Evaluate, Iterate, and Adapt".

- The taxonomy is the right theoretical tool that encompasses all the dimensions that relate to Wikimedia, internal and external.

- I have been bold at proposing changes, as this is how I would "frame it". My intent is to open it so it connects better with the different stakeholders. Bridging.

Thank you very much!

Marc Miquel i Ribé

marcmiquel@gmail.com

Universitat Pompeu Fabra, Barcelona