The State of the Internet's Languages: The Language Geography of Wikipedia

Whose Knowledge?

OXFORD INTERNET INSTITUTE

UNIVERSITY OF OXFORD

THE CENTRE FOR Internet & Society

حالة لغات الإنترنت

LAPORAN KONDISI BAHASA-BAHASA INTERNET

GAKLHE LLUNLELLO LLÍN XTILLO KA LUE DAN NÉ INTERNET

ইন্টারনেটের ভাষাপরিস্থিতি রিপোর্ট

ÉTAT DES LIEUX DES LANGUES D'INTERNET

STATE OF THE INTERNET'S LANGUAGES

RELATÓRIO DO ESTADO DOS IDIOMAS DA INTERNET

INFORME SOBRE EL ESTADO DE LAS LENGUAS EN INTERNET

RIPOTI YA LUGHA YA HALI YA MTANDAO

網路語言狀態報告

internetslanguages.org

Why language?

Language is more than a way of communicating…

We express what we think, believe, and know through our languages
Language is a proxy for knowledge

To be multilingual is to honor and affirm
the full richness and textures of our many selves and our different worlds better

An internet that is not multilingual and multimodal
is inherently unjust

This is the language of friendship, the reconciliation of skies
An ocean of art and knowledge, this world of language
(Latif Siddiqi, Urdu!)

What we asked and aimed for

With over 7000 (spoken and signed) languages in the world...

How many can we fully experience online?

What would a truly multilingual internet look, feel and sound like?

- Map the current status of languages on the internet
- Raise awareness of the challenges and opportunities in making the internet more multilingual
- Advance an agenda for action

What we did

Stories

- 12 countries
- Every continent
- 13 languages (stories + translations)

Numbers

- 11 websites
- 12 Android apps, 16 iOS apps
- Google Maps, Wikipedia

Summary

- Weaving together
- Community review and resource
- Solidarity in action

...and always a work-in-progress!

What we learned

The internet is nowhere near as multilingual as we imagine or need it to be. Only about 500 of our 7000+ languages are represented online in any form of information or knowledge.

Most people have to use their nearest European colonial language (English, Spanish, Portuguese, French…) or regionally dominant language (Mandarin Chinese, Arabic…) to access the internet.

Historical and ongoing structures of power and privilege are intrinsic to the way languages are accessible (or not) online.

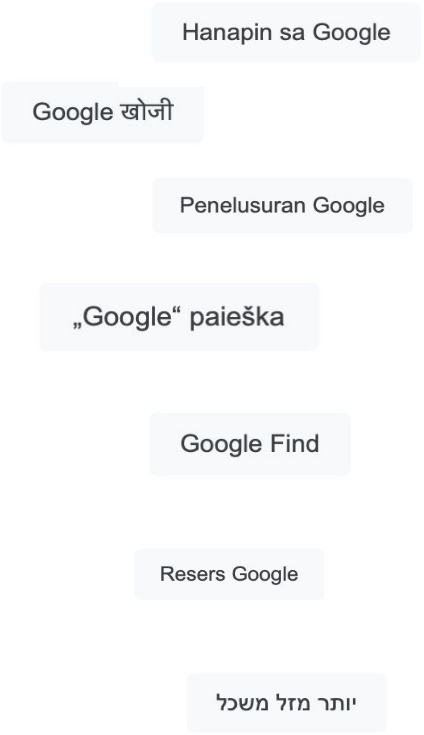
How many are affected by digital language exclusion?

More than 7,000 languages are in use today.

Yet, **digital platforms don’t reflect this** language diversity.

Many of us cannot use our preferred languages to access websites and apps.

How many people are affected by this digital exclusion, on the basis of their language?



A platform survey: languages on major sites & apps

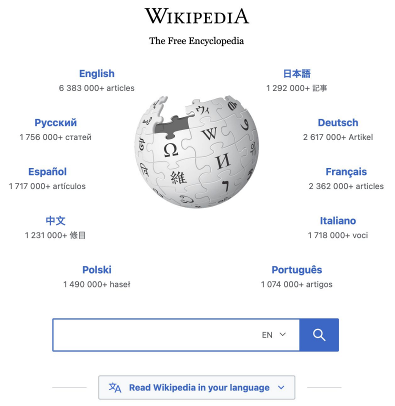
We present the first ever survey of the **interface languages** (written text forms) supported by digital platforms.

Our focus: popular websites and mobile apps for **knowledge access, language learning, and communication.** We reviewed 11 websites, 12 Android apps, and 16 iOS apps.

We asked:

Is the interface for this website or app available in different languages?

Which ones?



WIKIPEDIA

The Free Encyclopedia

English
6 383 000+ articles

日本語
1 292 000+ 記事

Русский
1 756 000+ статей

Deutsch
2 617 000+ Artikel

Español
1 717 000+ artículos

Français
2 362 000+ articles

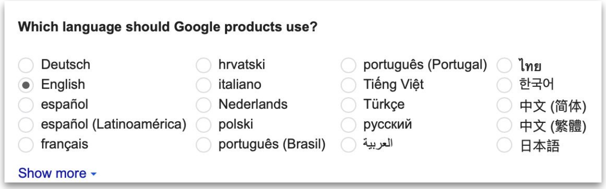
中文
1 231 000+ 條目

Italiano
1 718 000+ voci

Polski
1 490 000+ haseł

Português
1 074 000+ artigos

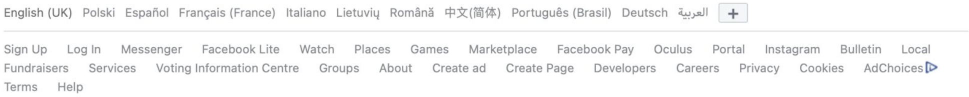
Read Wikipedia in your language



Which language should Google products use?

○ Deutsch
● English
○ español
○ español (Latinoamérica)
○ français
○ hrvatski
○ italiano
○ Nederlands
○ polski
○ português (Brasil)
○ português (Portugal)
○ Tiếng Việt
○ Türkçe
○ русский
○ العربية
○ ไทย
○ 한국어
○ 中文 (简体)
○ 中文 (繁體)
○ 日本語

Show more



English (UK) Polski Español Français (France) Italiano Lietuvių Română 中文(简体) Português (Brasil) Deutsch العربية +

Sign Up Log In Messenger Facebook Lite Watch Places Games Marketplace Facebook Pay Oculus Portal Instagram Bulletin Local Fundraisers Services Voting Information Centre Groups About Create ad Create Page Developers Careers Privacy Cookies AdChoices Terms Help

We find: highly unequal language support

Interface language support is often limited to particular languages:

- Certain **European** (en, fr, es, pt) and certain **Asian** languages (Mandarin Chinese, Indonesian, Japanese, Korean) are **very widely supported**
- However, **thousands of languages are not supported at all**, including languages of the Global South spoken by hundreds of millions

Some platforms do better than others:

- **Wikipedia**, **Google Search**, **Facebook** support more than 100 languages
- Among messaging apps: **Signal Messenger** supports 50-70 languages
- **They are the exception**: many others only support around 10-30

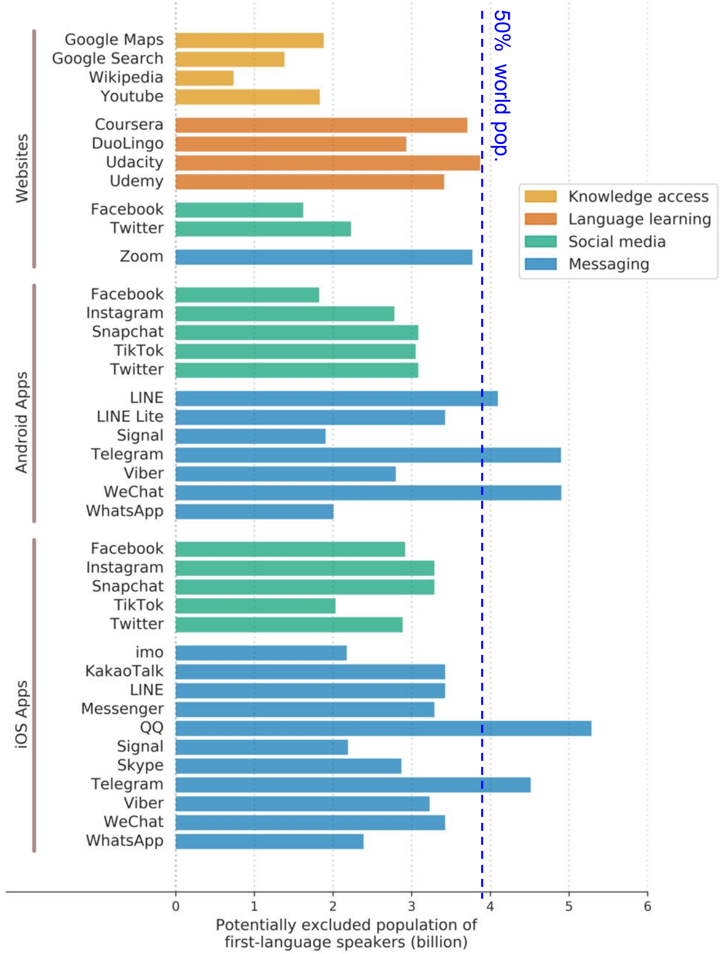
How many people are excluded?

Unfortunately, data about language *literacy* is scarce. Ethnologue has estimates of *oral* language populations.

We add up: how many people are potentially unable to access a platform if a certain language is not supported?

We find potential **digital exclusion of *billions* of people** on the basis of their language preference.

We estimate that at least 3-5 billion people cannot use their preferred language in order to access these widely used websites and apps.

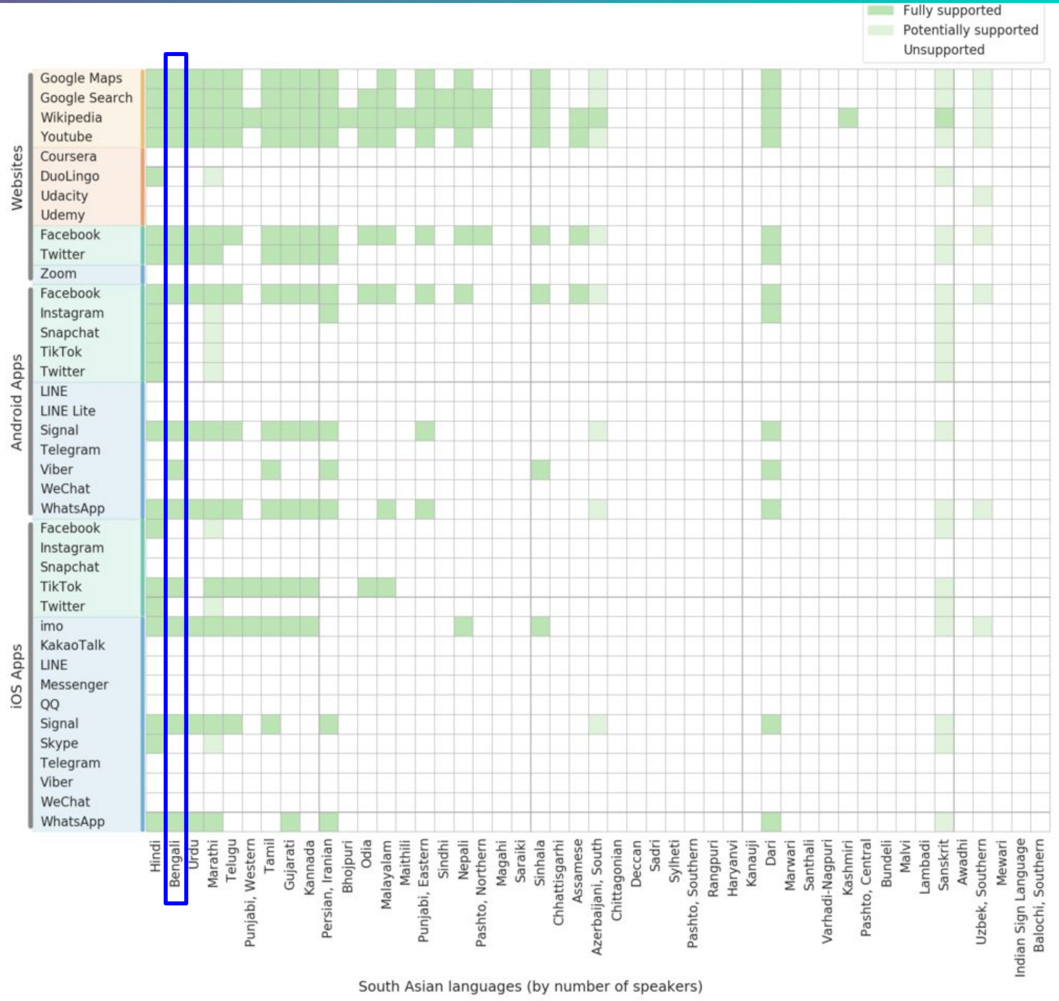


Martin Dittus and Mark Graham, Oxford Internet Institute 2020.
With kind support by Whose Knowledge?

South Asian languages

Only half of the platforms offer interface support for *any* South Asian language.

Hindi and Bengali, spoken by hundreds of millions of people, are not as widely supported as we might expect.

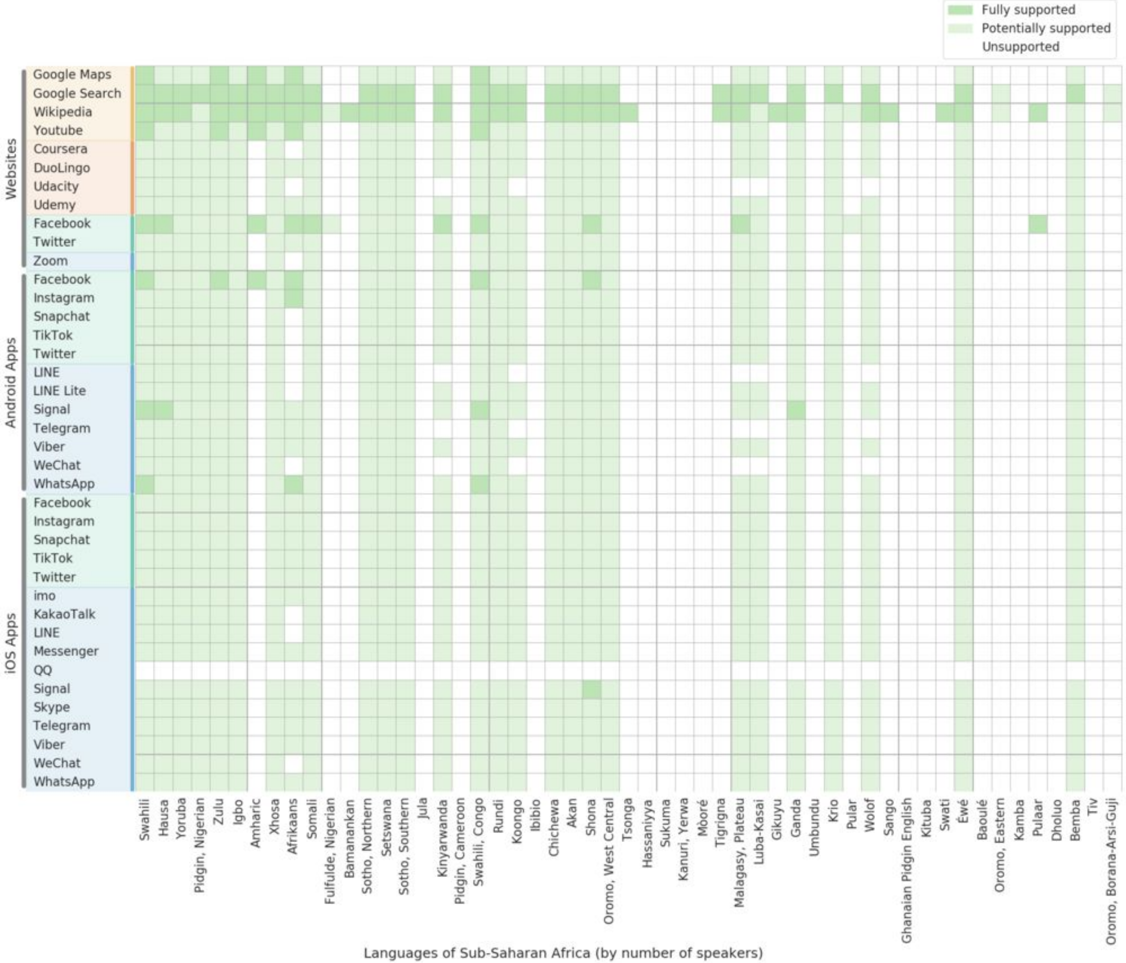


Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

African languages

Most Africans have to use a European-colonial language to access these digital platforms.

Many Africans speak English or French, along other languages. Yet not everyone will agree that these are “African languages”.



Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

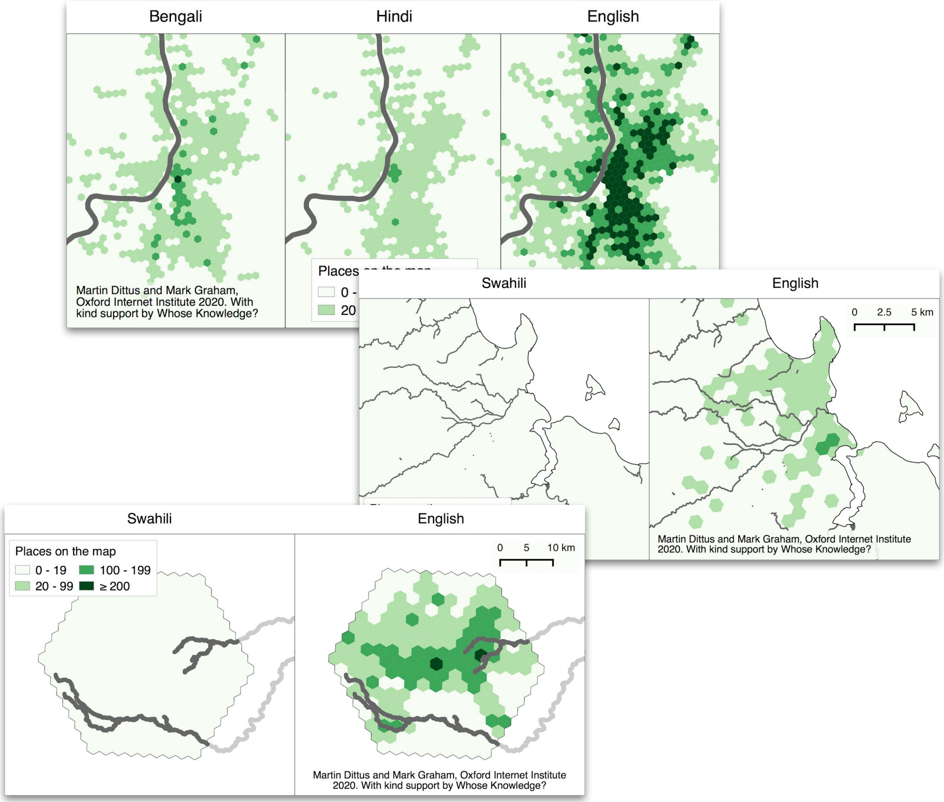
Does **Google Maps** show the world in your language?

Can everyone use Google Maps to navigate the world in their own language?

We ran millions of searches to find out. Again, we find **significant gaps**.

For example, on the map for **Kolkata** (India) we find many English-language results but less in Hindi and Bengali – yet all three languages are spoken there.

On the maps of **Dar es Salaam** (Tanzania) and **Nairobi** (Kenya) we find few results in Swahili, but many in English.

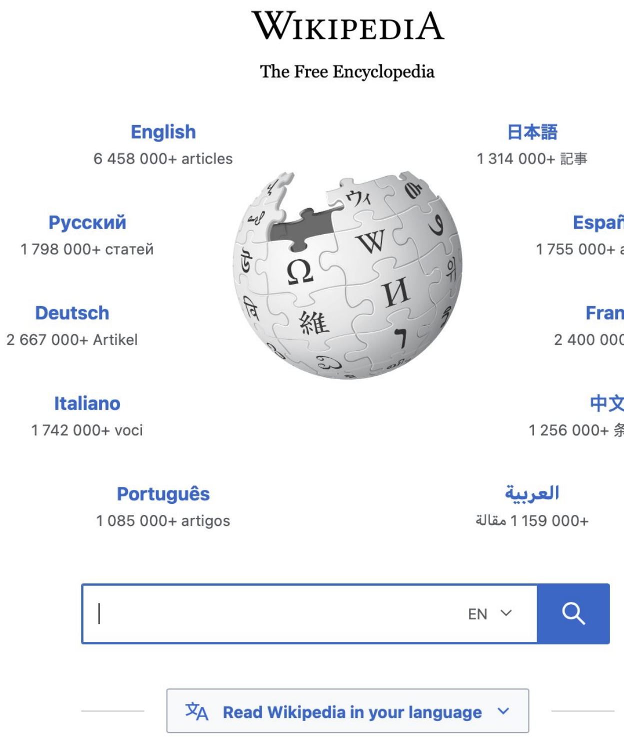


Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

Wikipedia today supports hundreds of languages



WIKIPEDIA

The Free Encyclopedia

English
6 458 000+ articles

日本語
1 314 000+ 記事

Русский
1 798 000+ статей

Español
1 755 000+ artículos

Deutsch
2 667 000+ Artikel

Français
2 400 000+ articles

Italiano
1 742 000+ voci

中文
1 256 000+ 条目 / 條目

Português
1 085 000+ artigos

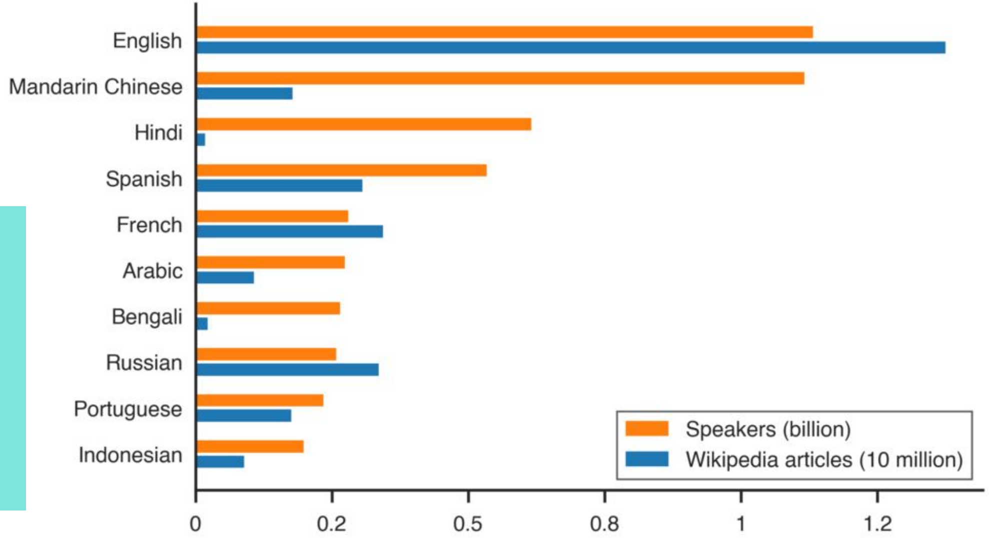
العربية
+000 159 1 مقالة

English Wikipedia has very comprehensive coverage

However, when we compare article counts to the population sizes for the 10 most widely spoken languages in the world…

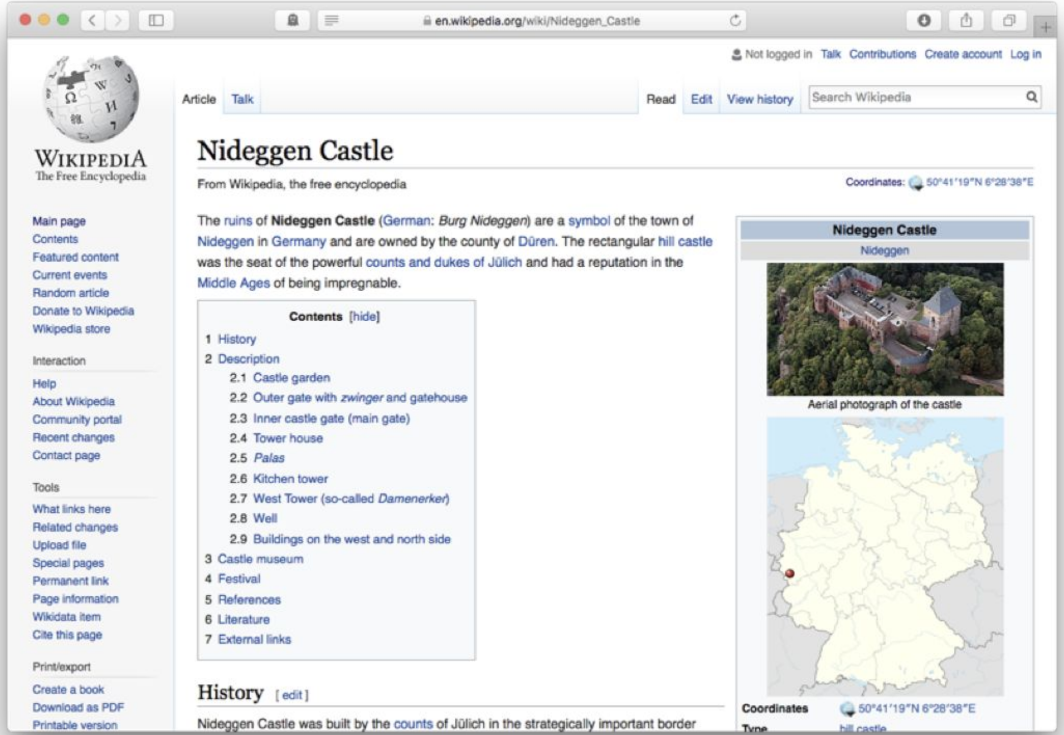
We find there is **comparatively little content in some very widely spoken languages** – for example, Mandarin Chinese, Hindi, Bengali, Indonesian, spoken by billions of people.

(Population estimate: Ethnologue 2019, which includes second-language speakers.)



Martin Dittus and Mark Graham, Oxford Internet Institute 2020.
With kind support by Whose Knowledge?

Wikipedia has descriptions of many of the world’s places



Nideggen Castle

From Wikipedia, the free encyclopedia

Coordinates: 50°41′19″N 6°28′38″E

The ruins of **Nideggen Castle** (German: *Burg Nideggen*) are a symbol of the town of Nideggen in Germany and are owned by the county of Düren. The rectangular hill castle was the seat of the powerful counts and dukes of Jülich and had a reputation in the Middle Ages of being impregnable.

Nideggen Castle

Nideggen

Aerial photograph of the castle

Coordinates	50°41′19″N 6°28′38″E
Type	hill castle

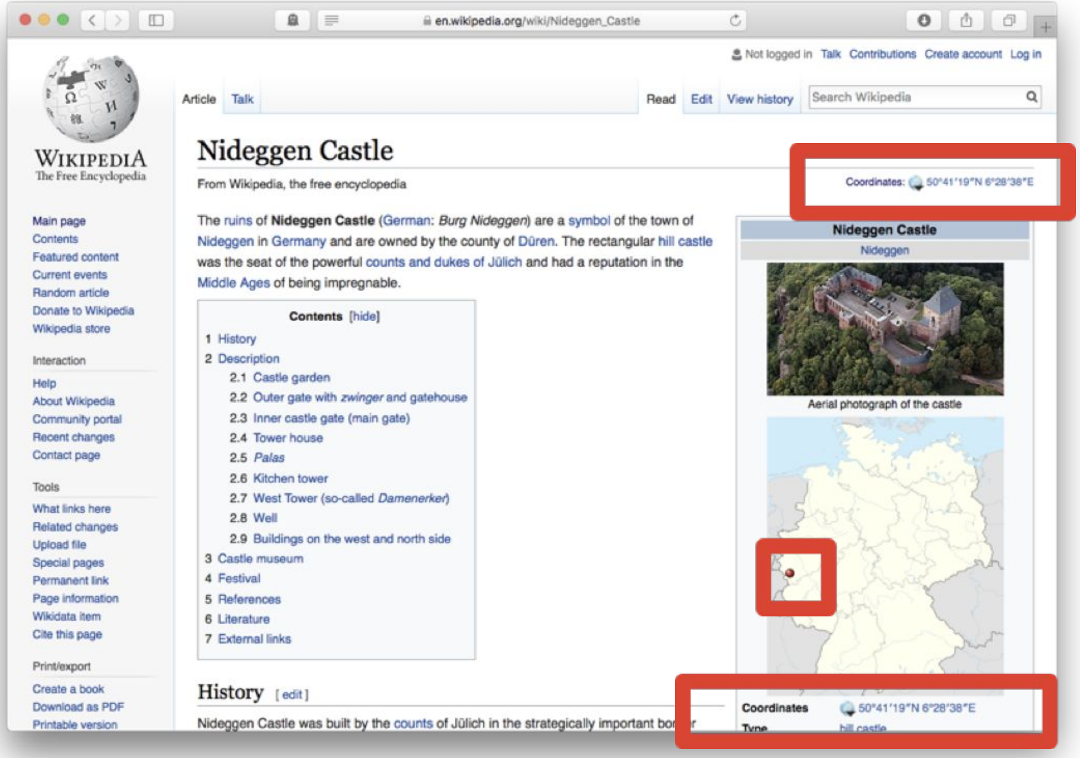
Contents [hide]

1 History
2 Description
2.1 Castle garden
2.2 Outer gate with *zwinger* and gatehouse
2.3 Inner castle gate (main gate)
2.4 Tower house
2.5 *Palas*
2.6 Kitchen tower
2.7 West Tower (so-called *Damenerker*)
2.8 Well
2.9 Buildings on the west and north side
3 Castle museum
4 Festival
5 References
6 Literature
7 External links

History [edit]

Nideggen Castle was built by the counts of Jülich in the strategically important border

Such articles can have geographic coordinates (“geotags”)



Nideggen Castle

From Wikipedia, the free encyclopedia

Coordinates: 50°41′19″N 6°28′38″E

The ruins of **Nideggen Castle** (German: *Burg Nideggen*) are a symbol of the town of Nideggen in Germany and are owned by the county of Düren. The rectangular hill castle was the seat of the powerful counts and dukes of Jülich and had a reputation in the Middle Ages of being impregnable.

Contents [hide]

Nideggen Castle

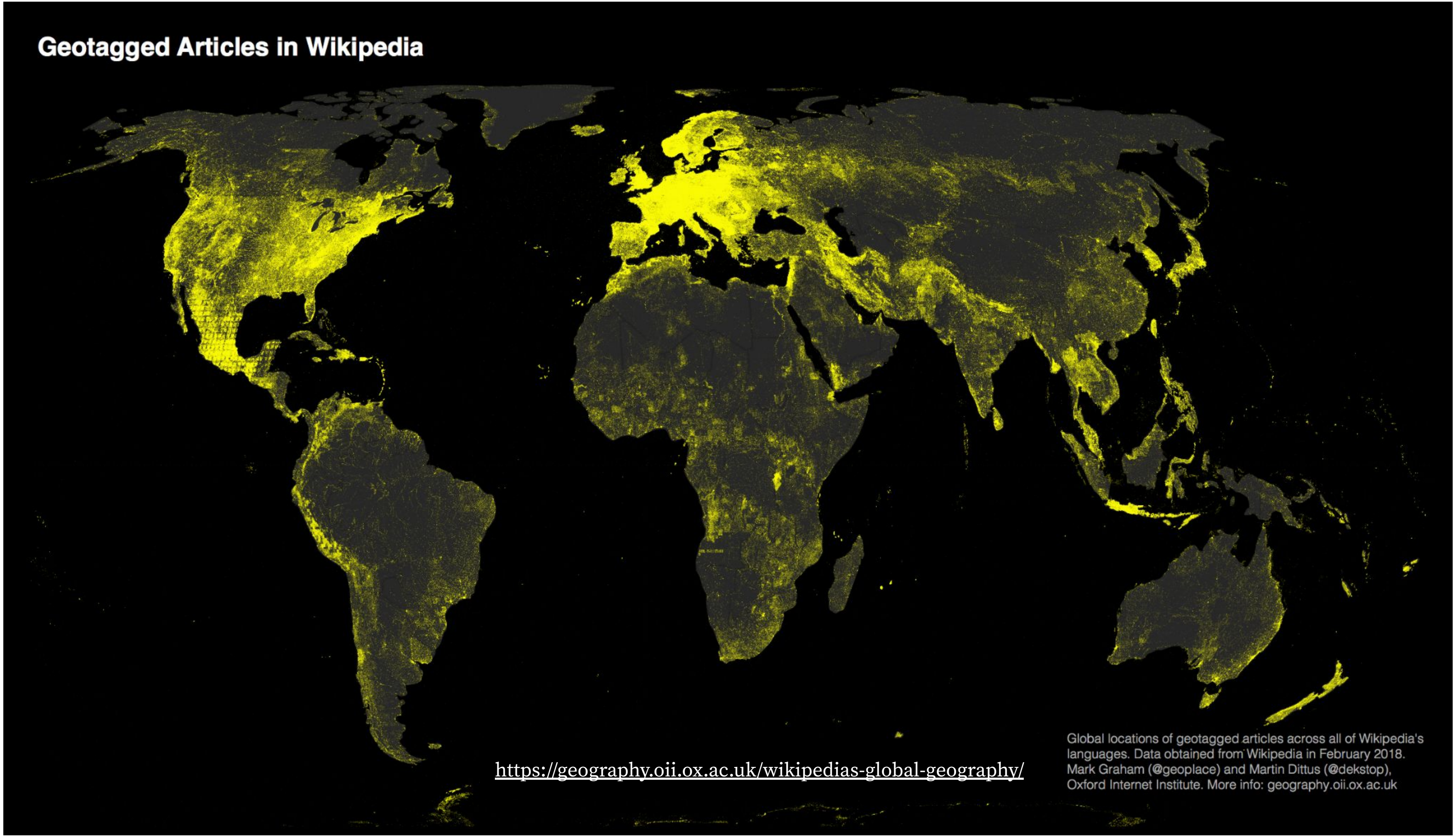
Nideggen

Aerial photograph of the castle

Coordinates	50°41′19″N 6°28′38″E
Type	hill castle

History [edit]

Nideggen Castle was built by the counts of Jülich in the strategically important bor

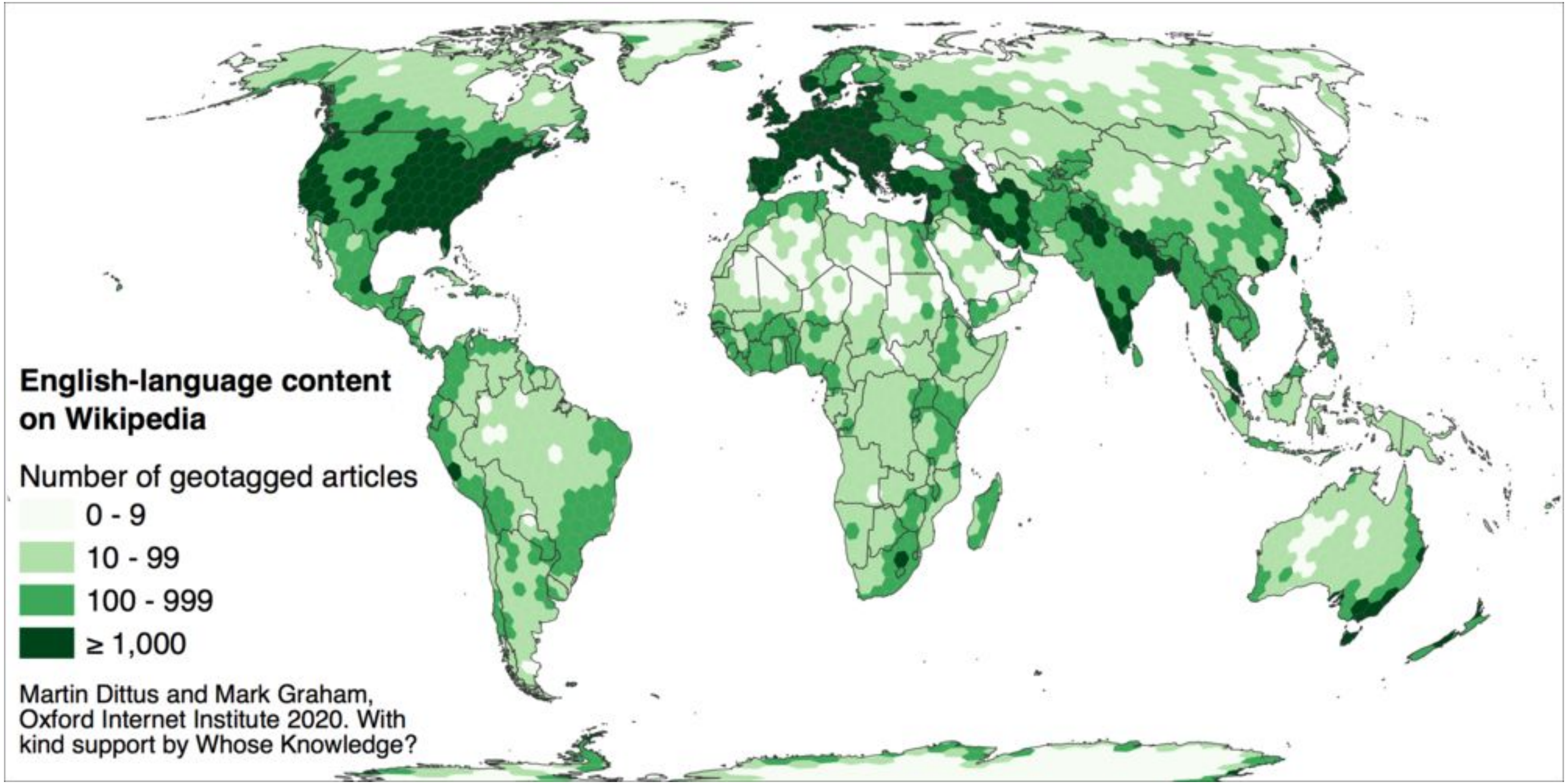


Geotagged Articles in Wikipedia

Global locations of geotagged articles across all of Wikipedia's languages. Data obtained from Wikipedia in February 2018. Mark Graham (@geoplace) and Martin Dittus (@dekstop), Oxford Internet Institute. More info: geography.oii.ox.ac.uk

https://geography.oii.ox.ac.uk/wikipedias-global-geography/

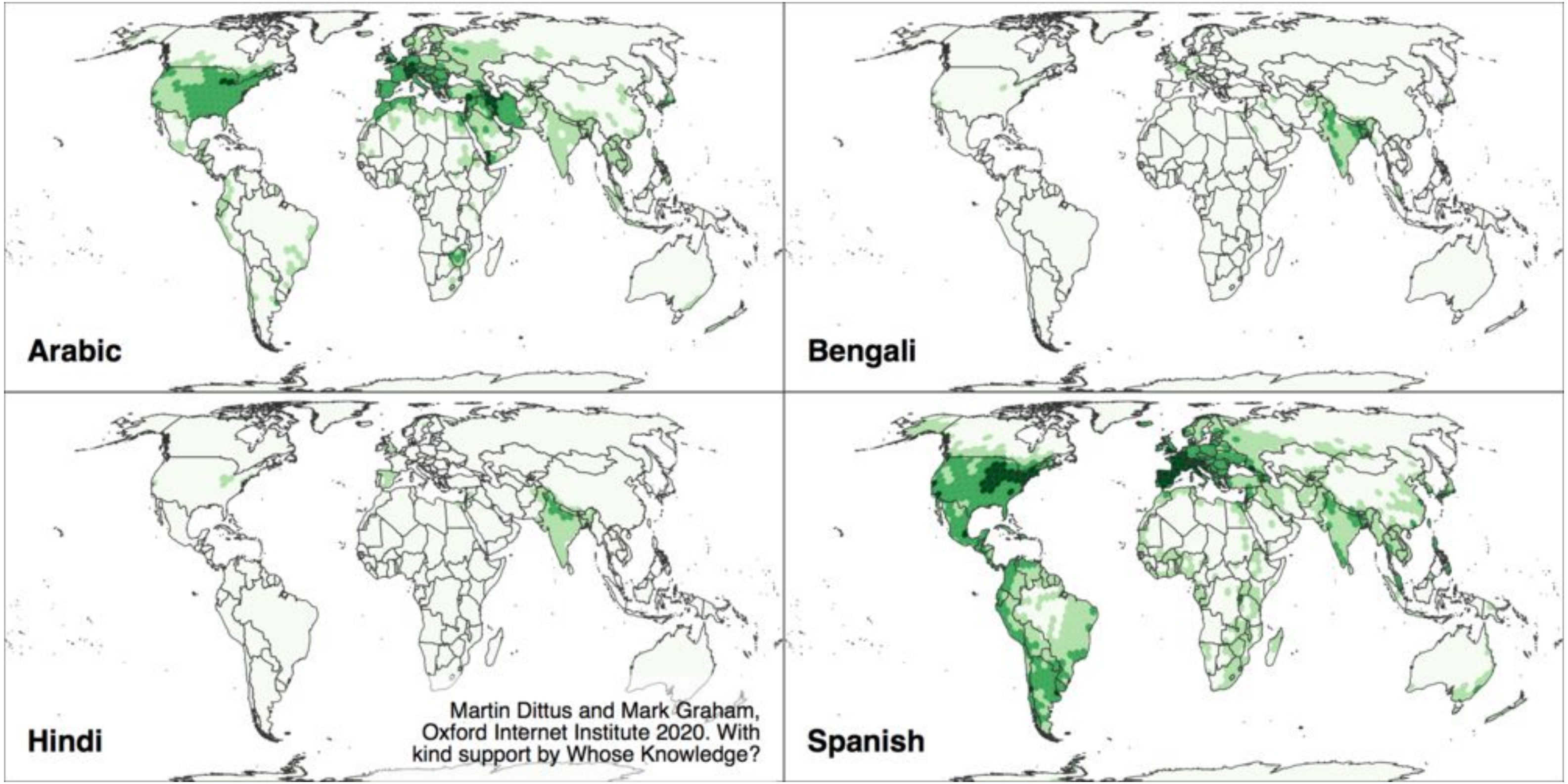
Geographic coverage of *English* Wikipedia



English-language content on Wikipedia

Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

Geographic coverage in other widely spoken languages



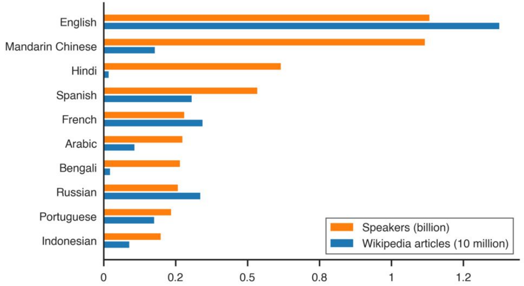
Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

Wikipedia's coverage is highly uneven!

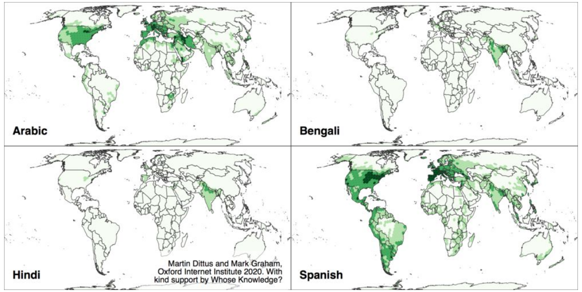
There is **language inequality**: some languages have much more content on Wikipedia than others.

There is **geographic inequality**: some geographic regions are much more well-covered than others.

Taken together, we can say that **Wikipedia has a *highly unequal language geography*.**

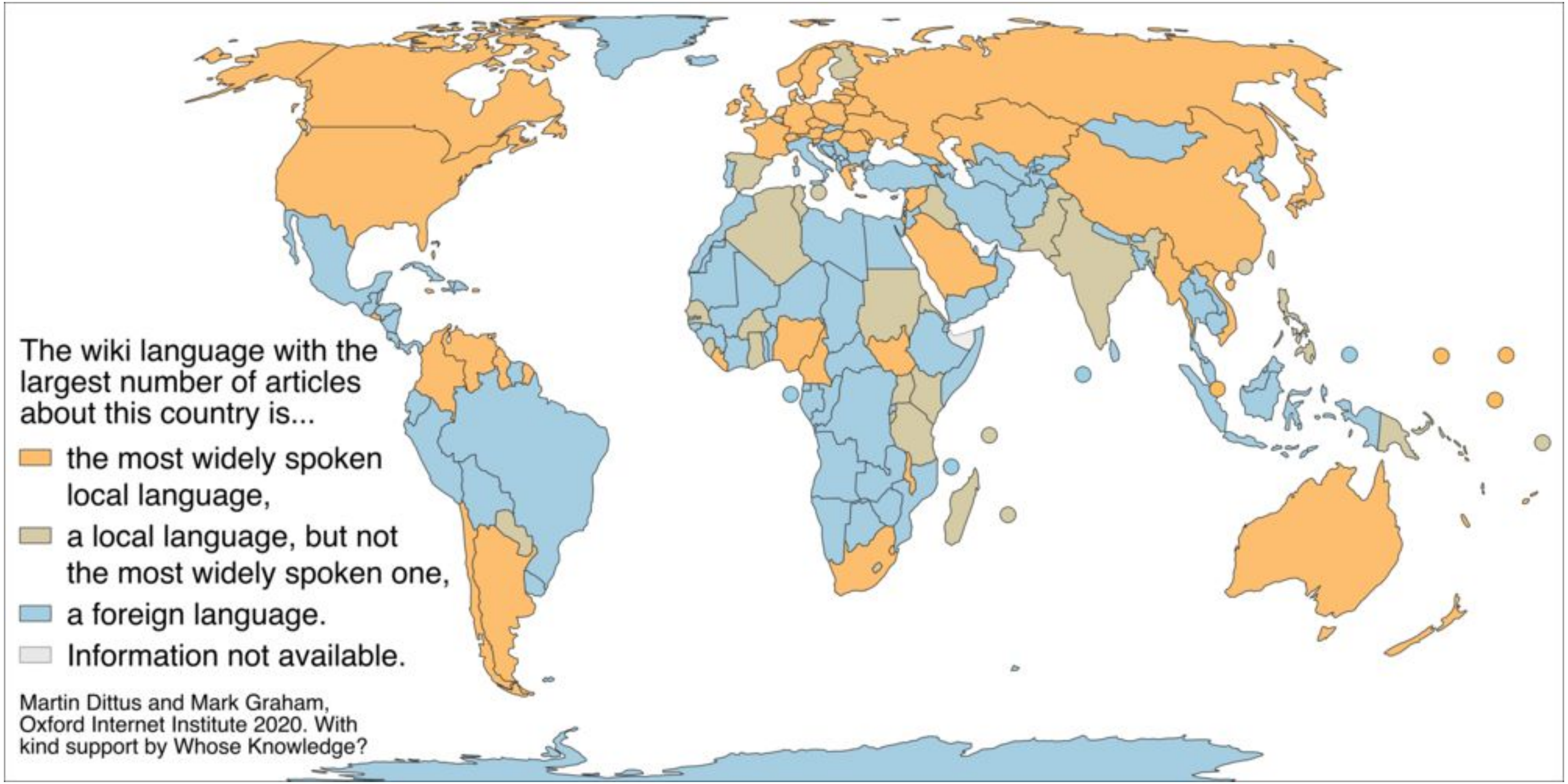


Martin Dittus and Mark Graham, Oxford Internet Institute 2020.
With kind support by Whose Knowledge?



Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

Many places are not well-described in their local languages!


The wiki language with the largest number of articles about this country is...
🟧 the most widely spoken local language,
▨ a local language, but not the most widely spoken one,
🟦 a foreign language.
□ Information not available.

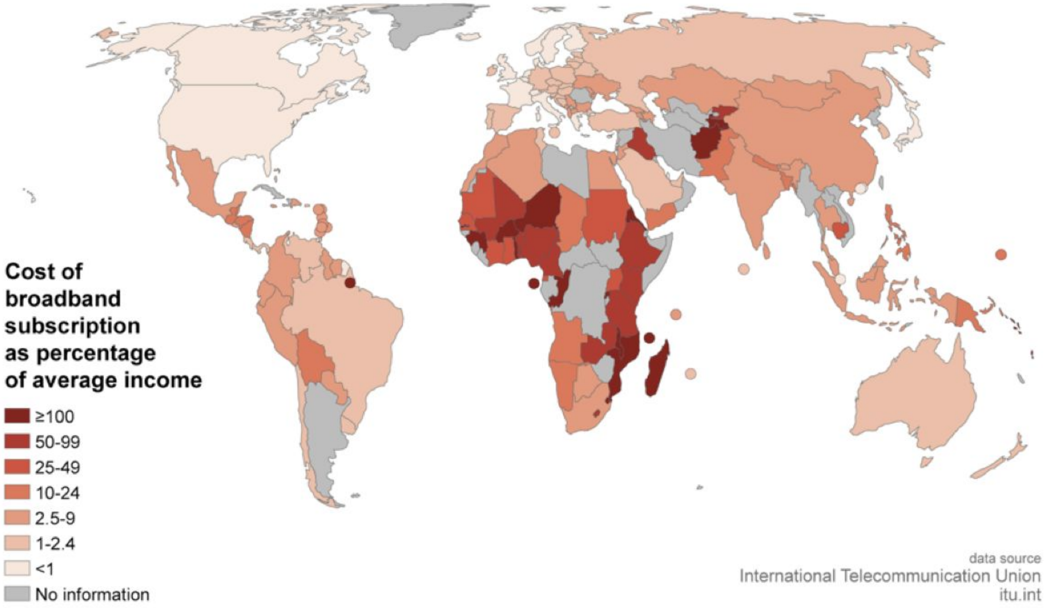
Martin Dittus and Mark Graham, Oxford Internet Institute 2020. With kind support by Whose Knowledge?

Who creates Wikipedia's descriptions of the world?

What perspectives, experiences, histories are left out?

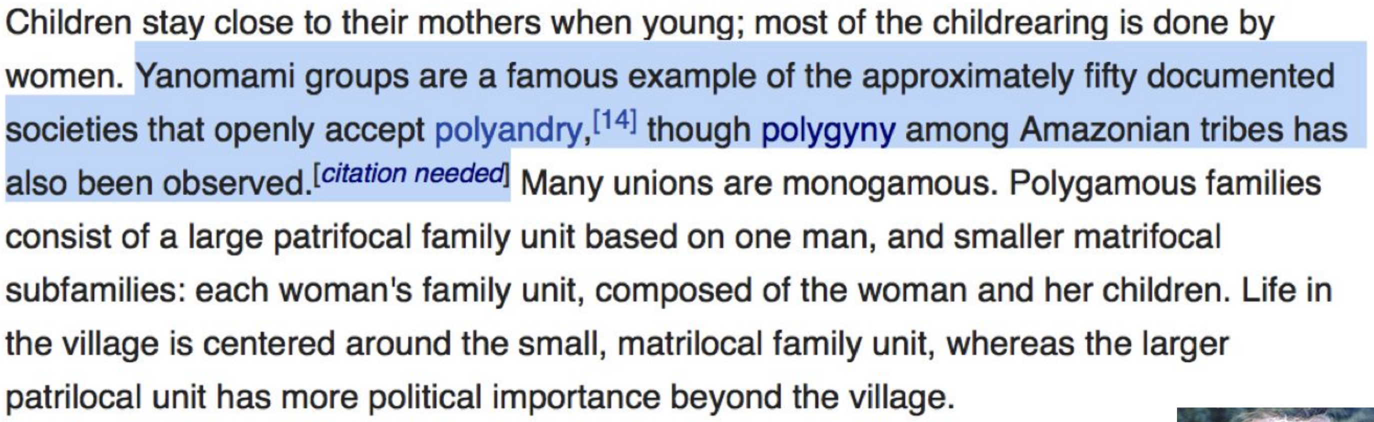
What barriers prevent others from participating?

Local participation requires *connectivity*


Cost of broadband subscription as percentage of average income
≥100
50-99
25-49
10-24
2.5-9
1-2.4
<1
No information
data source
International Telecommunication Union
itu.int

The cost of broadband subscriptions relative to the gross national income per capita, in 2013. Graham et al. (2015), “Towards a Study of Information Geographies”

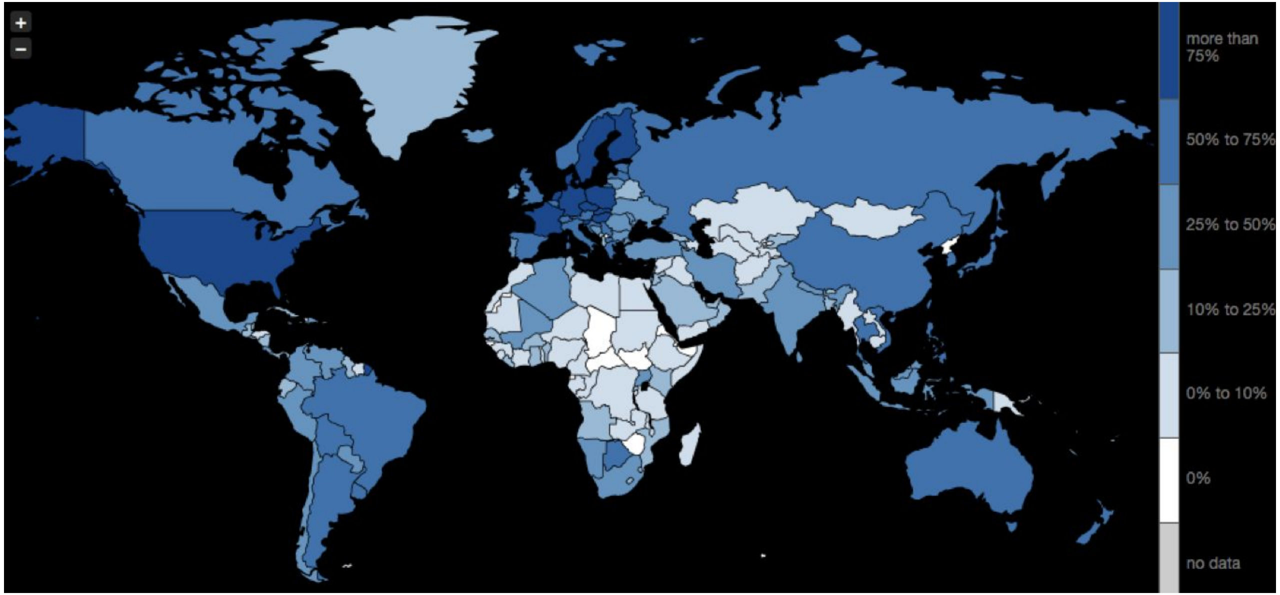
An unintentional barrier: *contribution norms*



Children stay close to their mothers when young; most of the childrearing is done by women. Yanomami groups are a famous example of the approximately fifty documented societies that openly accept polyandry,[14] though polygyny among Amazonian tribes has also been observed.[*citation needed*] Many unions are monogamous. Polygamous families consist of a large patrifocal family unit based on one man, and smaller matrifocal subfamilies: each woman's family unit, composed of the woman and her children. Life in the village is centered around the small, matrilocal family unit, whereas the larger patrilocal unit has more political importance beyond the village.

https://en.wikipedia.org/wiki/Yanomami

Local contributors need *local-language references*… which are not always available.

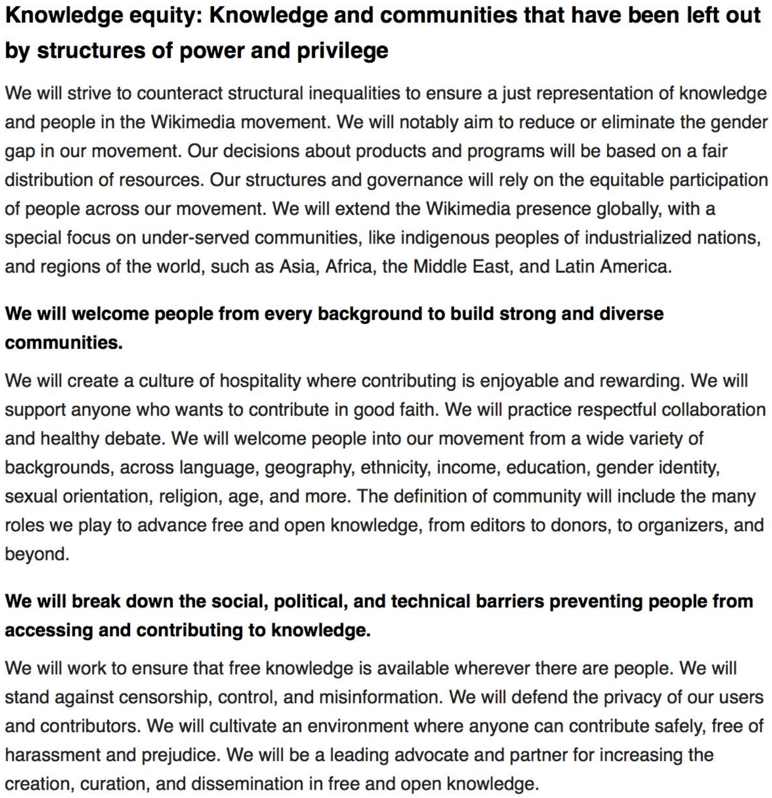


Source localness for Wikipedia articles. Interactives at http://www.shilad.com/localness/
Sen et al. (2015), “Barriers to the Localness of Volunteered Geographic Information”

The Wikipedia response: **knowledge equity**

“We will strive to **counteract structural inequalities** to ensure a **just representation** of knowledge and people in the Wikimedia movement.”

https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction



Knowledge equity: Knowledge and communities that have been left out by structures of power and privilege

We will strive to counteract structural inequalities to ensure a just representation of knowledge and people in the Wikimedia movement. We will notably aim to reduce or eliminate the gender gap in our movement. Our decisions about products and programs will be based on a fair distribution of resources. Our structures and governance will rely on the equitable participation of people across our movement. We will extend the Wikimedia presence globally, with a special focus on under-served communities, like indigenous peoples of industrialized nations, and regions of the world, such as Asia, Africa, the Middle East, and Latin America.

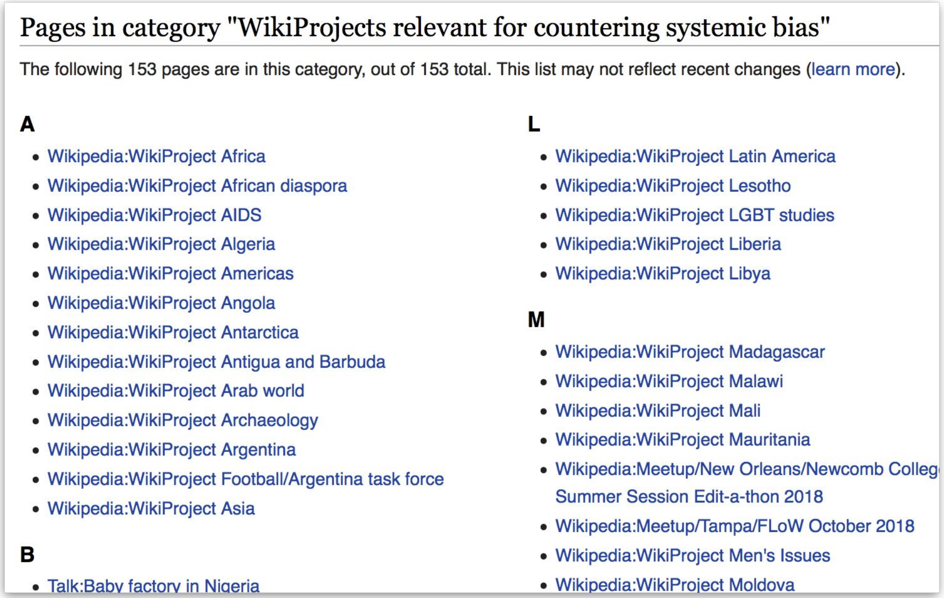
We will welcome people from every background to build strong and diverse communities.

We will create a culture of hospitality where contributing is enjoyable and rewarding. We will support anyone who wants to contribute in good faith. We will practice respectful collaboration and healthy debate. We will welcome people into our movement from a wide variety of backgrounds, across language, geography, ethnicity, income, education, gender identity, sexual orientation, religion, age, and more. The definition of community will include the many roles we play to advance free and open knowledge, from editors to donors, to organizers, and beyond.

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

We will work to ensure that free knowledge is available wherever there are people. We will stand against censorship, control, and misinformation. We will defend the privacy of our users and contributors. We will cultivate an environment where anyone can contribute safely, free of harassment and prejudice. We will be a leading advocate and partner for increasing the creation, curation, and dissemination in free and open knowledge.

Hundreds of Wikipedia projects are now working hard to address *systemic bias*



Pages in category "WikiProjects relevant for countering systemic bias"

The following 153 pages are in this category, out of 153 total. This list may not reflect recent changes (learn more).

A

- Wikipedia:WikiProject Africa
- Wikipedia:WikiProject African diaspora
- Wikipedia:WikiProject AIDS
- Wikipedia:WikiProject Algeria
- Wikipedia:WikiProject Americas
- Wikipedia:WikiProject Angola
- Wikipedia:WikiProject Antarctica
- Wikipedia:WikiProject Antigua and Barbuda
- Wikipedia:WikiProject Arab world
- Wikipedia:WikiProject Archaeology
- Wikipedia:WikiProject Argentina
- Wikipedia:WikiProject Football/Argentina task force
- Wikipedia:WikiProject Asia

B

- Talk:Baby factory in Nigeria

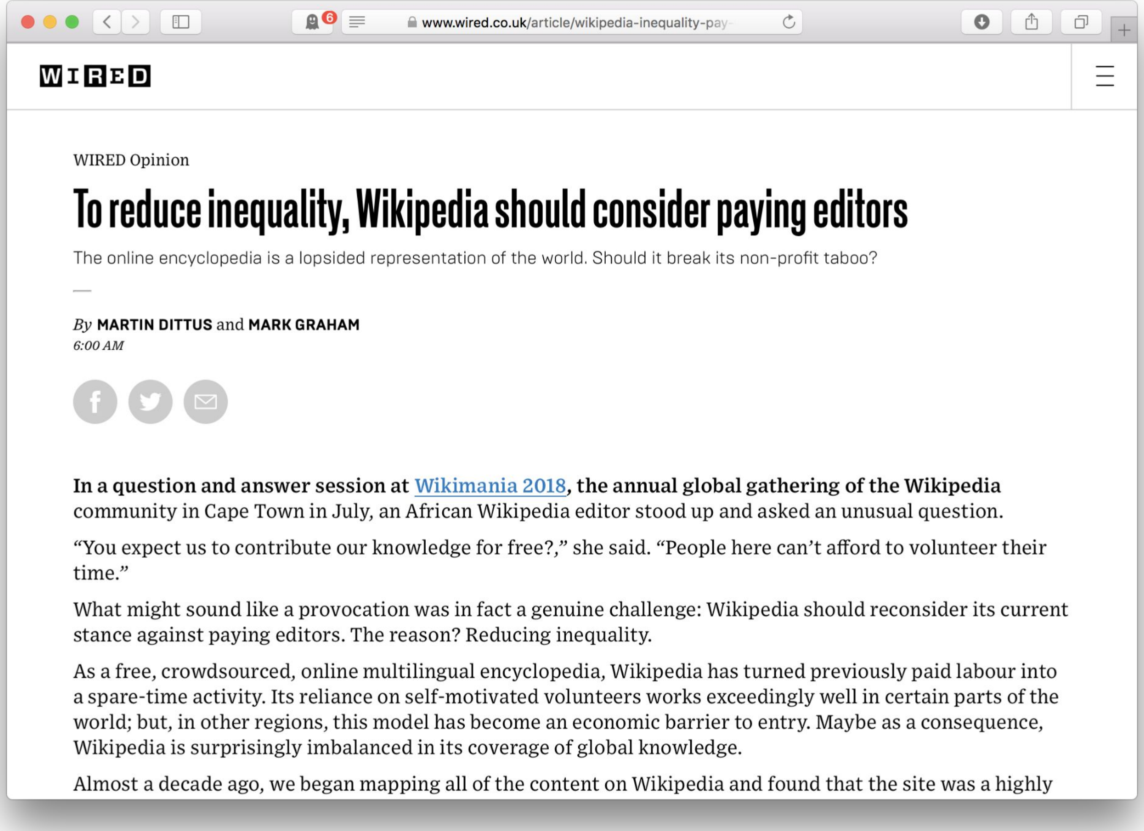
L

- Wikipedia:WikiProject Latin America
- Wikipedia:WikiProject Lesotho
- Wikipedia:WikiProject LGBT studies
- Wikipedia:WikiProject Liberia
- Wikipedia:WikiProject Libya

M

- Wikipedia:WikiProject Madagascar
- Wikipedia:WikiProject Malawi
- Wikipedia:WikiProject Mali
- Wikipedia:WikiProject Mauritania
- Wikipedia:Meetup/New Orleans/Newcomb Colleg Summer Session Edit-a-thon 2018
- Wikipedia:Meetup/Tampa/FLoW October 2018
- Wikipedia:WikiProject Men's Issues
- Wikipedia:WikiProject Moldova

https://en.wikipedia.org/wiki/Category:WikiProjects_relevant_for_countering_systemic_bias



WIRED Opinion

To reduce inequality, Wikipedia should consider paying editors

The online encyclopedia is a lopsided representation of the world. Should it break its non-profit taboo?

By **MARTIN DITTUS** and **MARK GRAHAM**
6:00 AM

In a question and answer session at Wikimania 2018, the annual global gathering of the Wikipedia community in Cape Town in July, an African Wikipedia editor stood up and asked an unusual question.

“You expect us to contribute our knowledge for free?,” she said. “People here can’t afford to volunteer their time.”

What might sound like a provocation was in fact a genuine challenge: Wikipedia should reconsider its current stance against paying editors. The reason? Reducing inequality.

As a free, crowdsourced, online multilingual encyclopedia, Wikipedia has turned previously paid labour into a spare-time activity. Its reliance on self-motivated volunteers works exceedingly well in certain parts of the world; but, in other regions, this model has become an economic barrier to entry. Maybe as a consequence, Wikipedia is surprisingly imbalanced in its coverage of global knowledge.

Almost a decade ago, we began mapping all of the content on Wikipedia and found that the site was a highly

https://www.wired.co.uk/article/wikipedia-inequality-pay-editors

In closing...

Many, many more examples, stories, and lived experiences in the report!

We want to reset expectations. Wikipedia is doing really well, however the remaining gaps might require fundamentally different approaches.

Most of all, **we look forward to shared creative approaches** to explore these issues!

Whose Knowledge?

The Centre for Internet & Society

Oxford Internet Institute

University of Oxford

internetslanguages.org