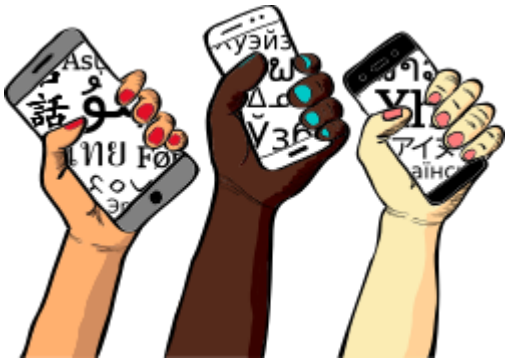

FACT SHEET - STATE OF THE INTERNET'S LANGUAGES REPORT

Most of the world is not served by the internet's languages. This report, created by a collective of three organizations — Whose Knowledge?, Oxford Internet Institute, and the Centre for Internet and Society (India) — maps some of the ways in which languages are represented online. It brings together contributors from 12 countries, from every populated continent, and mostly from the Global South. It has been an extraordinary community-sourced effort, with nearly one hundred people involved, from speakers and writers to translators and community reviewers. This digital report is the first of its kind and brings together contributions in 13 different languages, from Zapotec to Swahili. It is in line with the celebrations of UNESCO's International Decade of Indigenous Languages (2022-2032) and the International Mother Language Day. Through this initiative, we hope to raise awareness about the need to make the internet more multilingual and advance an agenda for action.

In this fact sheet, we highlight the key insights and takeaways from this report.

- **There are more than 7000 (spoken and signed) languages in the world, yet how many of them can we fully experience online? Not many.**

Most people access online content in languages that are not their first: those with European colonial histories or those that are regionally dominant. Over 75% of those who access the

Internet do so in only ten languages — such as English, French, Spanish and Chinese (which is not a single language, but a family of many languages; Mandarin being the most dominant). Read more about it in our summary report.

- **Users in Africa and Asia face multiple limitations in the type of content available in their language.**

In 2021, projections estimate about 7.9 billion humans on our planet, most of whom live in Asia (nearly 4.7 billion) and Africa (nearly 1.4 billion). After analyzing 11 websites, 12 Android apps, and 16 iOS apps, we found out that over 90% of users in these regions still have to switch to a second language in order to use these platforms. Read more about it in "A platform survey: interface language support by widely-used websites and mobile apps".

- **Online content is still biased towards certain languages and often does not include any non-written languages.**

We also know that of over 7000 languages in the world, only about 4000 of them have written systems or scripts. However, most of these languages use scripts that were not developed by the speakers of the languages themselves, but as part of the many colonizing processes across the world. Simply having a script doesn't mean that it is understood or used widely. Most of the languages in the world are transmitted in spoken or signed form, and not through writing mainly through oral traditions, without written text. Yet non-text-based languages — those based on sign, sound, gesture, movement — are completely missing from the publishing industry, and often from digital language technologies. These technologies rely on the automated processing of published materials in different languages in order to improve their language support and content. So when text publishing across the world is itself biased towards certain languages — and cannot include any non-written languages — it deepens the language inequities we experience. Read more about it in our summary report.

- **Even when content is available in certain languages, it might be limited or biased.**

It isn't enough to be able to access information and knowledge created for us in other languages by those who may not understand our contexts and experiences, and worse, be hostile to them. We need to be able to produce meaningful knowledge for ourselves and our communities. For instance, in languages like Bahasa Indonesia and Bengali, it is still difficult to find educational and positive queer content. Read more about these limitations and biases in our Stories section.

- **It is still difficult to create content in languages whose scripts are not easily understood by some technologists working on language support, especially if their form is different from the Latin script of western European languages.**

Users who speak languages like Sinhala – spoken by over 20 million people in Sri Lanka as either a first or second language – face difficulties with standard technology and systems like Unicode. Unicode is the technology standard for coding the text that is expressed in a language's writing system or script.

The Unicode Consortium (a non-profit based in California) also decides on emojis — the symbols that we use every day through different interfaces. For decades after emojis were first introduced to the internet, First Nations or Indigenous peoples had unsuccessfully petitioned for their use to express their oral and visual languages, like Arrernte. The Unicode Consortium deliberates upon public requests for new emojis, and petitions such as an emoji for the Australian Aboriginal Flag were rejected. Read more about these limitations and biases in our Stories section.

- **On Google Maps, we found that coverage in certain major languages like Bengali and Hindi is highly constrained to specific geographic regions, while coverage in other languages is more dispersed — likely relating to the existing language geography.**

There are some indications that Google seeks to address content gaps through the inclusion of foreign-language content when results are not available in the search language. This kind of content substitution happens for some languages more than for others. For example, searches in Arabic, Indonesian, Spanish, and French frequently return some English-language results. Read more about it in "The language geography of Google Maps".

- **On Wikipedia, only around 20 language editions have more than one million articles, and only 70 have more than 100,000 articles. Most Wikipedia language editions have only a small fraction of the content in English Wikipedia.**

When looking at Wikipedia's geographic coverage as a whole, information about places in Europe and North America is highly detailed, while many other regions of the world are relatively underrepresented, particularly places in Africa, parts of Asia, and in other regions of the Global South. In these regions, we found that most of the content about countries is in a foreign language. In other words, in many of these places, people may need to be able to speak a second (possibly foreign) language in order to access Wikipedia information about their own places. Read more about it in "The language geography of Wikipedia".

For interview requests and other media-related issues, please contact Priscila Bellini via languages@whoseknowledge.org.