Digital Development

Working Paper Series

The Digital Development working paper series discusses the broad issues surrounding digital data, information, knowledge, information systems, and information and communication technologies in the process of socio-economic development

Paper No. 112

Knowledge Divides in the Era of Big Data: Using Wikipedia Data to Identify and Measure Divides across Countries and Languages

ALFONSO RIVERA-ILLINGWORTH, RICHARD HEEKS & JACO RENKEN

2025

Publisher: ***Centre for Digital Development***
Global Development Institute, SEED
University of Manchester, Arthur Lewis Building, Manchester, M13 9PL, UK
Email: cdd@manchester.ac.uk Web: https://www.cdd.manchester.ac.uk/

View/Download from:
http://www.gdi.manchester.ac.uk/research/publications/di/

Table of Contents

Knowledge Divides in the Era of Big Data: Using Wikipedia Data to Identify and Measure Divides across Countries and Languages

Alfonso Rivera-Illingworth[1,2], Richard Heeks[1] & Jaco Renken[1]
[1]Centre for Digital Development, University of Manchester, UK
[2]Tecnologico de Monterrey, Mexico

2025

Abstract

Knowledge is fundamental to societal development and individual wellbeing, with digital technologies becoming crucial for its production and consumption. This study explores the use of Wikipedia log files as a big data source to measure knowledge divides, focusing on knowledge consumption. Using the Wikipedia API to extract billions of pageviews, this research measures consumption volumes, per capita estimates, and technological divides, broken down by income level and regions. It also addresses language divides in knowledge consumption by combining Wikipedia data with traditional language data. In addition a knowledge divides index is estimated using a Wikipedia consumption indicator and an existing production indicator.

This study analyses Wikipedia big data qualities using a conceptual framework. Compared to traditional datasets, Wikipedia data offers increased geographical availability, cost-effectiveness, improved accuracy in some aspects, timeliness, and accessibility. However, it is less complete for measuring sociodemographic dimensions and reflects Western knowledge representation, with potential biases in content production and consumption profiles.

The research identifies implications for development and policy, suggesting Wikipedia data should complement other measures in multidimensional knowledge indicators. These can help policymakers track countries' progress and identify underrepresented languages. Despite some limitations Wikipedia big data proves valuable for measuring knowledge divides when used alongside traditional sources, offering insights into disparities previously difficult to measure.

A. Introduction

Knowledge is recognised as a fundamental element for the development of societies and the wellbeing of individuals (UNESCO-UIS, 2003; WB, 2007; Castells, 2009; Russell, 2009; UNESCO/IFAP, 2016). The relationship between certain forms of knowledge and digital technologies is very close as the former is required to make proper use of the latter; moreover, digital technologies have become an increasingly widespread means to acquire and share knowledge (Segev, 2010; UNESCO/IFAP, 2016; WB, 2016; ITU, 2018a). The term "digital divides" is used to designate three types of disparities: technological, value chain, and socioeconomical (Heeks, 2018). Knowledge divides underpin technology-related divides, and are included as part of the socioeconomical dimension. The divides that were identified and measured before the era of big data are referred to here as "traditional digital divides".

Knowledge divides, at least those related to the Western conceptualisation of knowledge, have been measured using multiple traditional indicators such as educational attainment, literacy rates, research articles produced, patents, and indices of the knowledge economy at the country level, among others (UNESCO-UIS, 2003; WB, 2012; Cornell University et al., 2019). These traditional indicators present some of the limitations related to the use of traditional data sources. Big data sources have been used as part of the indices designed to measure knowledge divides with some benefits compared to traditional data (Cornell University et al., 2019; Ojanperä et al., 2019). Wikipedia, a widely used online encyclopaedia that has been around for almost 20 years, appears as a valuable big data source that can be used to measure knowledge divides. This paper focuses on responding to the following question: how can Wikipedia be used to measure knowledge divides?

This case explores the use of Wikipedia data to measure divides in the consumption of digital knowledge. With the use of log files about page views generated as a byproduct of the consumption of Wikipedia contents, this case focuses on measuring divides between countries in the consumption of digital knowledge. Using descriptive analyses this case estimates knowledge divides by measuring consumption volumes, page views per connected capita, and technological divides, all these broken down by income level and regions. Also, language divides in the consumption of knowledge are addressed combining big data with traditional data about languages (speakers, families, country of origin) and income to identify additional disparities. In addition, an existing knowledge production indicator is explored, along with an estimated knowledge consumption indicator, in order to suggest the use of a Wikipedia production-consumption index to measure inter-country divides. The results of these measurements allow the identification of multiple knowledge divides that cannot be identified using traditional data and present the possibilities of using these data as a complement to traditional indicators.

This working paper is structured in the following manner. Next, the notion of knowledge divides is presented alongside the way these divides have been traditionally measured. The research strategy section describes the data collection and preparation, along with the descriptive and exploratory methodology used to analyse the consumption of knowledge. The fourth section shows the results of the measures of global divides in the consumption of

knowledge in terms of volumes and per capita, and a technological subdimension of knowledge divides. The fifth section presents the discussion of results of the exploration of language dimensions of knowledge divides, including volumes consumed, page views per speaker, and income divides. The sixth section includes the estimation of a Wikipedia consumption indicator, the analysis of an existing production indicator, and the estimation of a knowledge divides index at the country level combining production and consumption. Then, a discussion following a conceptual framework focuses on measures of knowledge divides identified, the value of these data compared to traditional data and the data-related divides. The last section presents some concluding remarks.

B. Knowledge Divides

This paper begins with a discussion of knowledge and knowledge divides. This section addresses the concept of knowledge divides and the way these divides have been traditionally measured.

B1. Knowledge and Digital Divides

First of all, it is important to mention that this document takes a particular view of knowledge. This case assumes a view of knowledge that is related to the understanding or awareness that individuals have about data, information, facts, or skills, that are acquired through a process of learning or discovering (Castells, 2009; Russell, 2009; UNESCO, 2010; UNESCO/IFAP, 2016). This form of knowledge has been shown to be an important basis for human wellbeing and for particular development goals. In addition this case utilises the term knowledge economy to refer to the capacity that countries have to get value from this type of knowledge and related technology for economic and social growth. Given this view of knowledge, and as noted above, it is possible to identify that digital information and communication technologies (ICTs) and knowledge are intertwined: such knowledge is essential for utilisation of these digital technologies, and digital ICTs are valuable tools to acquire and share this form of knowledge. Formal knowledge is thus integrally bound up with data and with Western notions of development, and it is equally bound up with digital technologies and their divides. However, we acknowledge that there are other views on knowledge that do not follow the dominant Western conceptualisation and positivist views.

The digital production and consumption of this knowledge presents disparities that might be impacted by the levels of access and use of ICTs. As a result of these disparities knowledge can be represented as a continuum – less or more knowledge – for instance, some individuals might know some facts or have scientific knowledge around a topic, while others do not (Latour, 2007; Castells, 2009; Lundvall, 2016). For instance, ICTs can help in the production and consumption of knowledge; however, those without access to ICTs, or to the analogue foundations required to take advantage of this knowledge – education, literacy, among others – might face divides (Segev, 2010; Hilbert, 2011; ITU, 2018a). Also, these knowledge divides might be impacted by any of the traditional divides.

This study of knowledge divides focuses on the disparities in the digital production and consumption of multiple forms of knowledge between countries. Global knowledge divides

are present as it is expected that countries produce and consume knowledge in a different manner. For instance, while some countries produce and consume knowledge and benefit from participating in the knowledge economy, other countries produce and consume less and get only some of the benefits, thus negatively impacting their wellbeing (Dutta, 2012; WB, 2012; Mishra, 2015; Maaroof, 2016). The means to produce and consume knowledge – which are also unevenly distributed – are, for example, related to the access and use of ICTs, combined with elements such as educational attainment, economic conditions, human capital, etc. (Cornell University et al., 2015; Mishra, 2015; WB, 2016; ITU, 2018a). The state of the knowledge divides and the circumstances of countries in the knowledge economy are not easy to measure – this will be addressed in the next section.

Disparities in the consumption and production of knowledge, in addition to inequalities in access and use of ICTs, can also impact existing divides. For instance, countries with a higher number of individuals and organisations connected to the Internet, and with the knowledge and skills to participate in the knowledge economy, are expected to obtain the benefits (Cornell University et al., 2015; UNCTAD, 2019; Unger, 2019; WB, 2019c). One more example is that of countries whose workers have the knowledge to develop digital skills, as these can participate or adjust to the new opportunities and challenges brought by growing digitalisation within labour markets (Heeks, 2017; EC, 2019; UNCTAD, 2019; WB, 2019c). On the contrary, the existing divides will amplify for those countries that cannot properly participate in the knowledge economy – due to the lack of connectivity and ICT skills, among others. ICTs are facilitating the production and consumption of knowledge and are vital to participation in the knowledge economy; however, there is no clear evidence about the magnitude of the divides (Mishra, 2015; WB, 2016). The fact that knowledge and ICTs are intertwined can amplify existing divides, thus it is important to have clear measures.

In sum, the term knowledge divides here specifically refers to disparities in the digital production and consumption of particular forms of knowledge. Disparities in access and use of ICTs and knowledge can impact existing divides; yet, there is no complete evidence about the state of the global knowledge divides.

B2. Measuring Knowledge and Knowledge Divides

The lack of a unique definition and complete evidence about knowledge poses some challenges when measuring these knowledge divides. This sub-section describes the way these knowledge divides have been traditionally measured at the country level, the data sources that have been used – emphasising the limitations of these – and the new measurements of knowledge divides using big data sources.

Considering the lack of a sole definition of knowledge, multiple measures have been used to measure divides at the country level. This case recognises three types of measures: traditional indicators, indicators of the analogue foundations of knowledge, and indices of the knowledge economy and society.[1] Usually these measures have been defined by

[1] A series of knowledge tasks that have been identified from early attempts to measure the knowledge economy provide ancillary elements to quantify knowledge and the related divides. These tasks are related to measuring knowledge inputs; knowledge stocks and flows; knowledge outputs; knowledge networks; and knowledge and learning (Machlup, 1980; OECD, 1996).

international, intergovernmental or multilateral organisations. In the case of traditional indicators, measures of research and development (R&D) such as scientific articles published, as well as indicators of intellectual property like patents have been widely used to measure knowledge in a country (OECD, 2015; IPA-WIPO, 2016; OECD/Eurostat, 2018). In the second group, indicators of the analogue foundations of knowledge such as countries' literacy rates and educational attainment have been frequently used to measure knowledge (Mishra, 2015; WB, 2019a). The third group includes indices of the knowledge economy such as the Knowledge Economy Index (KEI) (WB, 2012), the Index of Knowledge Societies (IKS) (UNESCWA, 2005), and the Global Innovation Index (Cornell University et al., 2018). These multidimensional indices aggregate some traditional knowledge indicators with measures of the analogue foundations of knowledge, ICT measures, economic and employment indicators, innovation indicators, among others (Chen and Dahlmann, 2006; WB, 2012; EBRD, 2019; Ojanperä et al., 2019). The data sources used to feed these three types of measurements are usually traditional data coming from surveys, industry, administrative, and statistical records. Overall, these data sources follow a model of measuring usually defined by high-income countries.

In addition to the lack of a unique definition of knowledge and the difficulty to quantify its measures, multiple limitations are identified due to the nature of the data sources used. The use of traditional data sources presents some limitations, and a selection of those related to the aforementioned traditional measures of knowledge divides can be summarised using ACARTA, a model used to identify data qualities: availability, completeness, accuracy, relevance, timeliness and accessibility (adapted from Heeks (2018)). When looking at the indicators used to measure knowledge there are usually some availability issues. For instance, data about educational attainment, books and scientific articles published/read, or patents registered are not available for some lower-income countries. In addition, some of the indices created are not available any more as these have been discontinued, like the World Bank KEI and the UN IKS (EBRD, 2019). Looking at the accuracy of these data, measures of intellectual property or patents use data from industry records whose accuracy it is not possible to verify. Indicators such as literacy and enrolment rates also have issues related to their accuracy as some countries have weak or absent National Statistical Offices (NSOs) to collect these data (WB, 2019b). The relevance of traditional indicators is uncertain, such as the case of some R&D measures that focus only on the role of the public sector. The timeliness of traditional data is a big issue as it takes at least a year – and sometimes many years when collected via censuses – to gather and finally release these data. Finally, regarding accessibility there are also challenges as some of these measures are not open-access, and some others have to be compiled from multiple sources, such as scientific articles published. These limitations highlight the importance of finding new sources of data that can be used to measure knowledge divides.

The measures of knowledge that incorporate the use of big data sources are presented as a separate group due to their importance in this study. At least two indices were identified measuring the knowledge economy using big data sources.[2] First, the Global Innovation Index included measures of knowledge around 'creative outputs', using big data about Wikipedia edits, and Internet domains data to rank countries (Cornell University et al.,

[2] For a comprehensive list of knowledge economy indices using traditional data see Ojanperä et al. (2019).

2018). Second, the Digital Knowledge Economy Index is based on the defunct KEI and makes use of big data from Wikipedia, GitHub, and Internet domain registrations to measure the knowledge economy (Ojanperä et al., 2019). In addition, work has been conducted using big data about Wikipedia edits to identify geographic disparities in the production of knowledge, representation, participation, and language divides – among others (Graham, 2011; Graham et al., 2014; Graham et al., 2015; Dutta et al., 2019). This prior research has focused on measuring knowledge stocks and flows, knowledge outputs and networks. However, all these big data measures have focused mostly on the production of knowledge without measuring consumption as a fundamental component of knowledge divides and to generate measures of the state of the knowledge economy. Previous research using Wikipedia production data highlights the value of using these big data to measure divides and opens the opportunity to measure the consumption of knowledge and its related divides focusing on knowledge flows, and exploring the related knowledge stocks and networks.

Overall, it is possible to identify that knowledge generally is fundamental for both national development and for the wellbeing of individuals, including formal knowledge, and that ICTs have become a fundamental means to produce and consume this type of knowledge. Knowledge is necessary to benefit from the use of ICTs: knowledge and ICTs are hence intertwined. Yet, disparities between countries in the production and consumption of multiple digital forms of knowledge can be identified along a continuum, and referred to as knowledge divides. These divides have been measured using traditional indicators, such as R&D indicators, educational attainment, and basic ICT indicators defined by international organisations. These indicators present multiple limitations. A few big data sources – including Wikipedia data – have been used to measure knowledge divides, mostly from the production side. This case will focus on using Wikipedia consumption data to explore how big data can be used to measure knowledge divides.

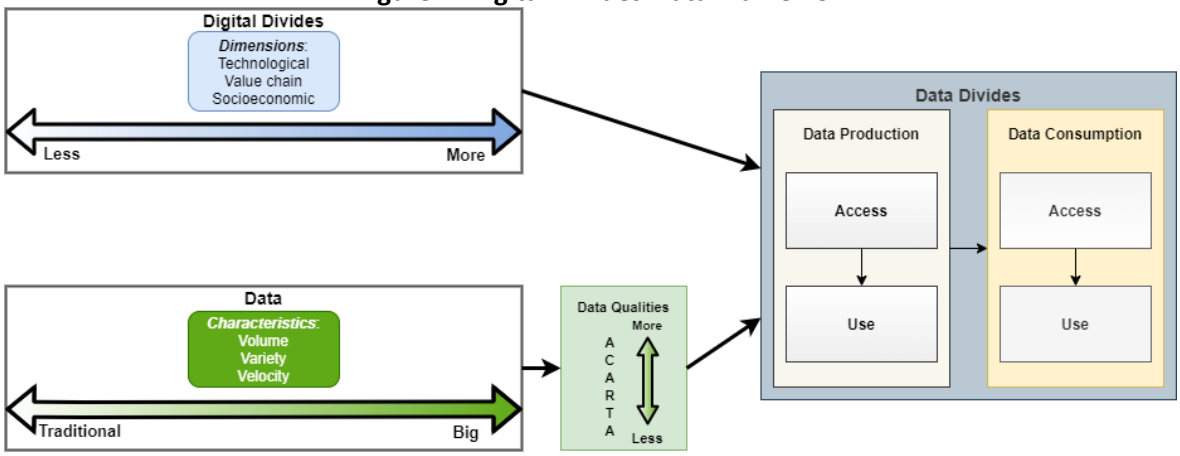
C. Research Strategy

A research strategy following a Social Data Science Model (SDSM) (see Appendix 1) and the conceptual framework was defined to answer the research question. This section describes the conceptual framework, the data collection and preparation stages and the research strategy followed to analyse divides in the consumption of knowledge using Wikipedia log data.

C1. Conceptual Framework

A conceptual and a methodological foundation are used to identify how big data can be used to measure digital divides. Figure 1 presents the elements of a Digital Divides-Data framework.

Figure 1: Digital Divides-Data Framework



Firstly, the digital divides are incorporated in the model considering the interactions between multiple dimensions, including technological and value chain dimensions – the former underpinned by socioeconomic dimensions. These dimensions are interacting with each other, confirming the multidimensional character of these divides in the era of big data. Secondly, data can be also placed on a continuum from the traditional data that has less volume, variety and velocity, to the opposite side where data with more volume, variety and velocity is found: big data. In addition to the 3 Vs, the qualities of any data can be analysed using the ACARTA model, with each one of these qualities also located on a more—less continuum. Thirdly, big data-related divides are found at all the stages of this section of the model – from access to use and from production to consumption. The model is not limited to the identification of big data-related divides but it can also be used to identify divides related to traditional data and related to aggregated big data or "less-big data".

Based on the conceptual framework and its operationalisation via the social data science methodology, this case focuses on identifying how Wikipedia big data can be used to measure divides around the consumption of knowledge.

C2. Data Collection and Preparation

The prior section presented some of the traditional data sources used to measure knowledge divides, their limitations, and introduced Wikipedia as a big data source that has been used to measure knowledge divides. Wikipedia – an online encyclopaedia – works as a collaborative project which allows individuals to freely produce and consume digital knowledge (Wikimedia, 2019e).[3,4] This online encyclopaedia is available in more than 300 languages (referred here as 'Wikipedias') that can be accessed in the majority of the countries in the world: at the time of this analysis China and Turkey were blocking access to Wikipedia (Wikimedia, 2019d; Wikimedia, 2019c). Despite the fact that there are other big

[3] "[A]n encyclopedia is a written compendium of knowledge" and a stock of recorded knowledge (Wikimedia, 2019e).

[4] During 2018 Wikipedia was among the top 5 visited sites globally; a little more than half of the traffic came from search engines (52%) (Alexa, 2019).

data sources – such as data and metadata from online search engines – that might be better sources to measure knowledge divides, these are not easily accessible to researchers. On the contrary Wikipedia is an open-access and cost-free data source that is valuable to conduct this type of research. The analysis of strengths and limitations of these data are addressed in Section G using the ACARTA model.

In contrast to those studies that have used Wikipedia data to measure the production of knowledge (edits), this case focuses on the consumption side at the country level. Specifically, this case uses the Wikipedia log files generated when users access online contents. These log files are categorised as data exhaust, a byproduct of the activities conducted by users of the Wikipedias – data exhaust is considered one of the big data sources useful for development.[5] These data were not generated with the aim of measuring knowledge divides, yet they are repurposed here to explore their value. This case assumes Wikipedia page views as a measure (proxy) of consumption of formal knowledge around the world. It is important to mention that this case does not analyse the content of the articles and does not take into consideration the quality of these, as content quality has been extensively analysed by other researchers (Okoli et al., 2012; Mesgari et al., 2015). Samples of these log files were obtained and analysed.

Table 1 presents the data points collected. For more details about the data collection and preparation processes see Appendices 1 and 2.

Table 1: Wikipedia data points collected

Characteristic	Description
Page views	Monthly page views in increments of 1,000
Country	Page views aggregated by country: monthly data for 249 countries and territories
Type of access	Page views by device used: mobile, desktop, and totals
Project	Language project page: 247 “Wikipedias”
Years	2016, 2017, 2018

The data presented in Table 1 were collected using the Wikimedia REST API to get page views between January 2016 and December 2018, with the aim of having annual measurements that could be compared – as detailed in the following sections.[6] The API cannot be used to access geo-located data for previous years and as described in Appendix 1 the extraction process is labelled as “experimental”; this means, that the extraction tools might be modified without prior notice before a stable version is released (or not) – for more limitations see Section G.2.

A total of 560 billion page views were aggregated for all the countries and languages in this analysis; only around 0.3 percent of all the page views could not be geolocated. Data were obtained for 249 countries and territories – the whole of the countries in the ISO 3166

[5] Even if the Wikipedia articles might not be considered per se as a big data source – the English version is about 15 GB when compressed –, the aggregated log files generated through the consumption of contents have the characteristics of a big data source (Wikimedia, 2019b).

[6] The Wikimedia REST API is a tool to process and extract data from the Wikimedia servers (Wikimedia, 2019c).

standard catalogue – categorised by income level using the OECD Development Assistance Committee (DAC) 2018 list of Official Development Assistance recipients and World Bank (WB) income data, and by region using World Bank data (OECD, 2018). Table 2 presents the count of countries and territories by region and income level.[7]

Table 2: List of countries by region and income level

Regions – Income level	Count
East Asia & Pacific	**38**
High-Income	13
Middle-Income	16
Low-Income	9
Europe & Central Asia	**56**
High-Income	36
Middle-Income	20
Latin America & Caribbean	**42**
High-Income	15
Middle-Income	26
Low-Income	1
Middle East & North Africa	**21**
High-Income	8
Middle-Income	11
Low-Income	2
North America	**3**
High-Income	3
South Asia	**8**
Middle-Income	4
Low-Income	4
Sub-Saharan Africa	**48**
High-Income	1
Middle-Income	14
Low-Income	33
Not categorised (WB)	**34**
Middle-Income	6
No level	27
Total	**249**

Some countries (34) have not been categorised by the World Bank under a specific region due to the fact that these are small countries, islands, and territories that are not sovereign states; also, the majority of these have not been classified by income level by the OECD. A total of 221 countries were categorised by income level and 216 by income level and region.[8]

Page views were extracted for 247 Wikipedias (languages) out of 259 that at the end of 2018 had at least 1,000 articles and had views during 2016-2018 (Wikimedia, 2019d). Then,

[7] Data for Kosovo (xk) were dropped from the analyses as it is not recognised as a UN member.

[8] The low-income category in Latin America and high-income in Sub-Saharan Africa representing Haiti and the Seychelles, respectively, are not shown in the charts as these include only one country.

the 247 Wikipedia projects were classified by size based on the number of articles available for each language at the end of 2018. Due to the high variance in the number of articles available (e.g. 5.8 million articles in English vs. 1,200 in Tahitian) six "ad hoc" categories were created. Table 3 presents the categorisation.

Table 3: Wikipedias size categorisation

Category Size	Wikipedias Size (articles)	Articles (%)	Wikipedias (count)	Wikipedias (%)
A	1 million or more	65.9	15	6.1
B	250,000 – less than 1 million	17.8	19	7.7
C	100,000 – less than 250,000	9.1	27	10.9
D	50,000 – less than 100,000	3.8	26	10.5
E	10,000 – less than 50,000	2.5	54	21.9
F	1,000 – less than 10,000	0.9	106	42.9
	Total	**100**	**247**	**100**

This stock of 247 Wikipedias comprise around four percent of the 7,867 languages in the ISO 639-3 catalogue.[9] Log data were augmented using information from Ethnologue (Eberhard et al., 2020) about the estimated total number of speakers (first- [L1] and second-language [L2] users). The 247 languages represented by these Wikipedias include more than 10 billion L1 and L2 estimated speakers (ISO, 2013; BSI, 2020; Eberhard et al., 2020; Wikimedia, 2020b). These data also included information about language families that were used to categorise the results. Table 4 presents the cross tabulation of Wikipedias by language family and size utilised.

[9] This is a rough estimate considering that some "macro" languages have only one Wikipedia page despite having multiple local variations of the language. Also, some of these almost 8,000 languages are not written languages.

Table 4: Wikipedias by language family and size categorisation (counts)

	Wiki Size Category					
Family	A	B	C	D	E	F
Abkhaz-Adyghe						2
Afro-Asiatic*		1	1		2	5
Austro-Asiatic	1					1
Austronesian*	2	2	1	3	4	12
Aymaran						1
Constructed language		1	1		2	3
Creole				1		5
Dravidian			1	2	1	
Eskimo-Aleut						1
Eyak-Athabaskan						1
Indo-European*	10	10	14	16	32	34
Japonic	1					
Kartvelian			1		1	
Koreanic		1				
Kra-Dai			1			2
Language isolate		1				
Mongolic					1	2
Nakh-Daghestanian			1			3
Niger-Congo*					2	10
Quechuan					1	
Sino-Tibetan*	1		1	2	3	3
Tupian						1
Turkic		1	4	2	3	7
Uralic		2	1		2	8
Uto-Aztecan						1

*Major language families

Out of a total of 142 language families, 25 (17.6%) had at least one language represented in the Ethnologue dataset. Indo-European languages were the ones with more Wikipedias covering around a quarter of all the languages in this family. None of the languages in the Trans-New Guinea family – one of the six major language families – were included in this dataset. This is because there are no Wikipedias for the majority of these languages, or the Wikipedias for these have less than 1,000 articles.

Finally, the data from Ethnologue allowed the identification of the primary country of each language. With these data and the income level data about countries (OECD, 2017; OECD, 2018; WB, 2020b), the income level of the primary country of each language was matched. Table 5 shows the count of languages broken down by income level of the primary country of language, the region of the language, and the Wikipedia size category.

Table 5: Wikipedias by income level of primary country of language, region, and size categorisation (counts)

Primary Country of Language	Language-Region	Size Category A	B	C	D	E	F
High-Income	East Asia & Pacific	1	1				2
	Europe & Central Asia	9	6	11	8	17	26
	Latin America & Caribbean						1
	Middle East & North Africa			1		1	2
	North America						3
Middle-Income	East Asia & Pacific	4	2	3	3	7	19
	Europe & Central Asia	1	7	5	7	7	22
	Latin America & Caribbean					1	3
	Middle East & North Africa		2	2		4	4
	South Asia			3	3	10	6
	Sub-Saharan Africa				1	1	5
Low-Income	East Asia & Pacific					1	3
	Latin America & Caribbean				1		
	South Asia				2	1	
	Sub-Saharan Africa				1	2	7
Not Applicable	Constructed International		1	1		2	3
	International (Latin)			1			

Note: Constructed international languages cannot be related to a single country or region; these languages include Esperanto, Interlingua, Ido, among others.

The majority of the Wikipedias are related to languages whose primary country is considered a middle-income one (53.0%); only a small fraction of the Wikipedias are related to languages whose primary country is a low-income one (7.3%). When looking at region, languages from high- and middle-income countries from Europe & Central Asia have the highest number of Wikipedias, with a good number of smaller ones. Overall, for all income levels the majority of Wikipedias are characterised by those with fewer articles available (categories D-F); and in the case of low-income countries these are the only ones available.

Considering the availability of big data for all countries and the fact that a high number of speakers were represented by Wikipedias included in the dataset, this source appeared as a good option to explore divides in the consumption of knowledge. When compared to traditional data sources the availability and completeness of these data is higher. In addition, data were available for a period of three years which was useful to identify initial trends in the divides. Yet, limitations are also acknowledged and discussed throughout the sections below. The following section describes the methodology used to analyse these data.

C3. Analysis Preparation

The analyses follow a quantitative research strategy focused mainly on measuring the consumption of knowledge by country and by language. As mentioned in the previous section this analysis comprised the years 2016, 2017, and 2018, for all countries available that viewed the 247 Wikipedias. The strategy comprises three stages: analysis of knowledge

divides by country; analysis by language; and the estimation of a consumption indicator and a production-consumption index.

I. The first stage focuses on exploring knowledge divides by country to identify global divides which includes three sub-stages: the overall descriptive analyses, estimations of per capita consumption, and the identification of technological divides related to knowledge divides.

a) The volume of Wikipedia page views by country is explored for the year 2018 to have a preliminary understanding of divides between countries.

b) The per capita consumption of articles per month for individuals connected to the Internet is estimated for each country to be able to compare between countries and years. The results are aggregated by regions and income level (year=2018). Then comparisons for the years 2016-2018 are presented. Kruskal-Wallis tests are used to test divides between income levels in 2018, to identify whether the magnitude of the divides is statistically significant or not (Scheff, 2016).

c) The type of access (mobile or desktop) is explored to understand if there are technological disparities in the consumption of knowledge. The results are aggregated by income level and region (year 2018) and then analysed for the period 2016-2018. Income level divides are tested for statistical significance as in b).

II. The second stage focuses on analysing knowledge divides by language and comprises four sub-stages:

a) The volume of Wikipedia pages consumed is explored by language for the year 2018 to get a preliminary view of divides between languages.

b) The per capita consumption of articles per month in 2018 is estimated for the total number of speakers of each language (first and second language) to allow comparison between languages and size categories.

c) Language divides are explored using the size categorisation of Wikipedias presented in Table 3 to identify the share of volume of page views consumed by income level in 2018 and the trends for 2016-2018. This allows us to identify if certain groups of languages are consumed more by a group of countries.

d) Language divides are also explored by income level to identify the proportion of page views consumed by Wikipedia size in 2018 and the trends for 2016-2018. This allows us to identify if certain groups of countries are consuming more pages from a category of languages (size).

III. The third stage comprises the estimation of a standardised indicator to explore the consumption of contents for all individuals in a country, and contrast it with an existing Wikipedia knowledge production indicator included in the Global Innovation Index (GII) (Cornell University et al., 2018). The indicators allow the identification of production-

consumption divides at the country, region, and income levels. The two indicators are estimated in the following manner:

a) The production indicator (wikiprod) is extracted from the GII and includes Wikipedia yearly page edits (per million population 15-69 years old) for a total of 125 countries available in the GII (Cornell University et al., 2018; Dutta et al., 2019). In this case the indicator estimates the production of contents per capita for all individuals in a country and not only for connected individuals (as in stage I, b), the implications of these estimates are discussed in Section F. Data for 2017 is used in this case as it was the latest available at the time of this analysis.

b) The consumption indicator (wikiconsum) uses the data extracted in this analysis to estimate the Wikipedia yearly page views in 2017 (per million population 15-69 years old) for the same 125 countries, following the GII approach. For the estimation of this indicator the consumption variable in b) is rescaled to a value of 0 – 100 in a similar manner as done for a) in the GII and following the re-escalation formula:

$$resc = \frac{x - min}{max - min}$$

This is done to normalise the distributions and correct the effect of outliers (OECD, 2008).

c) *Wikiprod* and *wikiconsum* take values between 0 and 100.

d) Finally, this stage includes also the estimation of an index that aggregates knowledge production and consumption, which allows the creation of a Wikipedia score and ranking based on these two dimensions. This index estimated a score for each country by giving equal weight to *wikiprod* and *wikiconsum* assuming similar relevance for the production and consumption of knowledge:

- *wikindex* = (.5 * *wikiprod*) + (.5 * *wikiconsum*)

Wikindex can take values between 0 and 100 within each country.

D. Data Analysis, Consumption

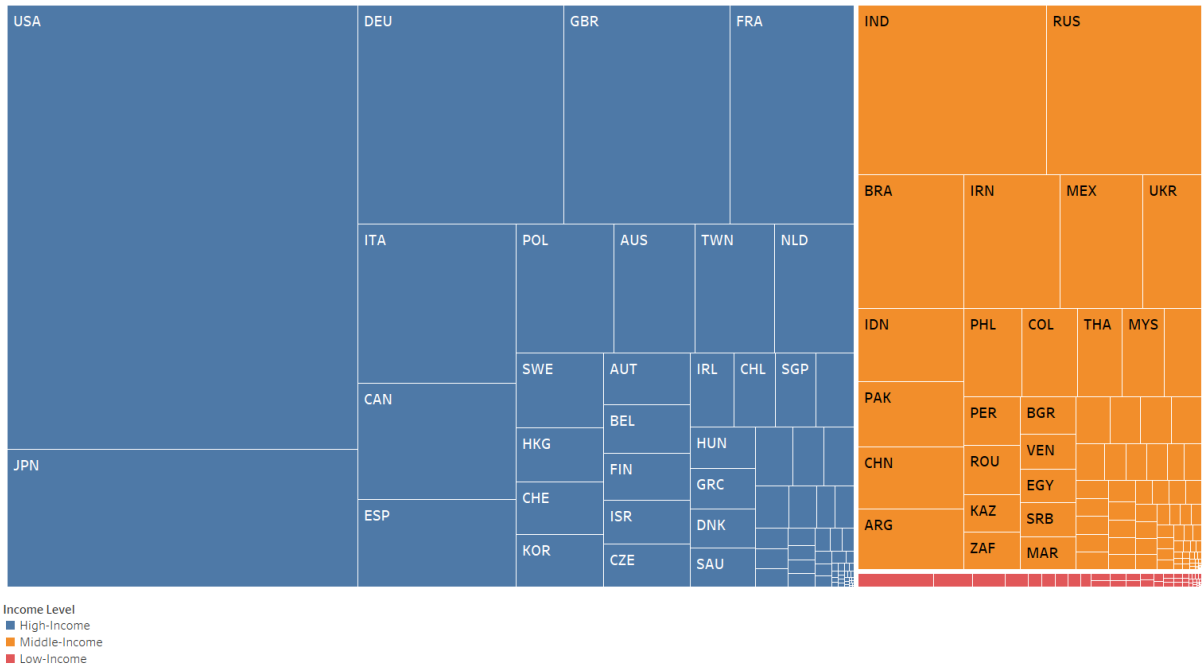
The analyses start by presenting the results of the Wikipedia log data related to the consumption of contents. The strategy focused on the analysis of page views per country in 2018 aggregated by income level and region, and by income level for 2016-2018. Three sub-stages allow the identification of the following divides: knowledge consumption divides; a per capita consumption of knowledge; and the identification of technological divides. Summary statistics for the variables in this section can be found in Appendix 3.

D1. Knowledge Volumes

With the aim of exploring divides in the consumption of knowledge, the volume of contents consumed per country were analysed in 2018. As mentioned before the integrated dataset includes data for all of the 249 countries, territories, and small islands recognised by the UN

and included in the ISO 3166 catalogue. Figure 2 shows the average monthly page views by country (size of shape) and by income level (colour). In this subsection the use of averages is an adequate measurement to analyse monthly data with seasonal trends.

Figure 2: Monthly average page views by country and income level (2018)



Labels ISO 3166-3: High-Income> USA=United States; JPN=Japan; DEU=Germany; ITA=Italy; CAN=Canada; ESP=Spain; GBR=United Kingdom; POL=Poland; SWE=Sweden; HKG=Hong Kong; CHE=Switzerland; KOR= South Korea; AUS=Australia; BEL=Belgium; FIN=Finland; ISR=Israel; CZE= Czech Republic; FRA=France; TWN=Taiwan; IRL=Ireland; HUN=Hungary; GRC=Greece; DNK=Denmark; SAU=Saudi Arabia; NLD=Netherlands; CHL=Chile; SGP=Singapore. Middle-Income> IND=India; BRA=Brazil; IDN=Indonesia; PAK=Pakistan; CHN=China; ARG=Argentina; RUS=Russia; IRN=Iran; PHL=Philippines; PER=Peru; ROU=Romania; KAZ= Kazakhstan; ZAF= South Africa; MEX=Mexico; COL=Colombia; BGR=Bulgaria; VEN=Venezuela; EGY=Egypt; SRB=Serbia; MAR=Morocco; UKR=Ukraine; THA=Thailand; MYS= Malaysia.

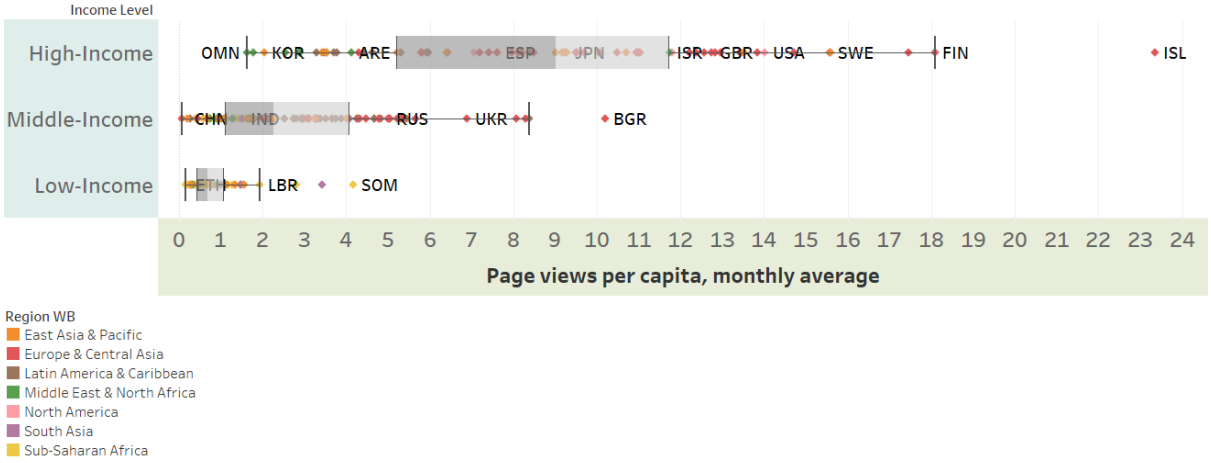
Two types of divides can be identified using Figure 2: income level divides and inter-country (global) divides in the volume of contents consumed by country. Firstly, divides are evident when looking at the volume consumed by income level. The blue area (left) shows that high-income countries consumed 71.0 percent of all page views in 2018, followed by middle-income countries with 28.1 percent, and low-income countries with only 0.8 percent. Secondly, the USA appears as an outlier in the average monthly page views of any language with around 3.5 billion monthly views (22% of the total), followed by Japan (1 billion), Germany (1 billion), the UK (800 million), and India (700 million). Consumption divides appeared as evident; however, this is addressed in more detail in the following section using per capita estimates for connected individuals, and aggregating the results by income level and regions.[10] This first insight is only possible with the use of big data as there were no traditional data about consumption volumes available.

[10]As mentioned before, the Wikipedia REST API connects to the analytical files that have already been processed to identify only what appears to be Internet human traffic.

D2. Page Views per Connected Capita

To generate fair comparisons, this section presents the estimates of the per capita consumption of Wikipedia page views at the country level for individuals connected to the Internet. The estimates are based on population and individuals' data with access to the Internet within each country for 2018 (ITU, 2018b; WB, 2020a).[11] In this case, data about individuals connected to the Internet was only available for 205 countries in the sample; for the rest of this section this will be the sample size. Using population and the number of connected individuals the monthly average of page views per connected capita (*pvpcc*) was estimated at the country level and is shown in the figures aggregated by income level (Figs. 3 and 4) and region (Fig. 5). Figure 3a presents a box plot of page views per capita (x-axis) by country and income level (y-axis); regions are shown using colours. Figure 3b presents the map showing *pvpcc*.

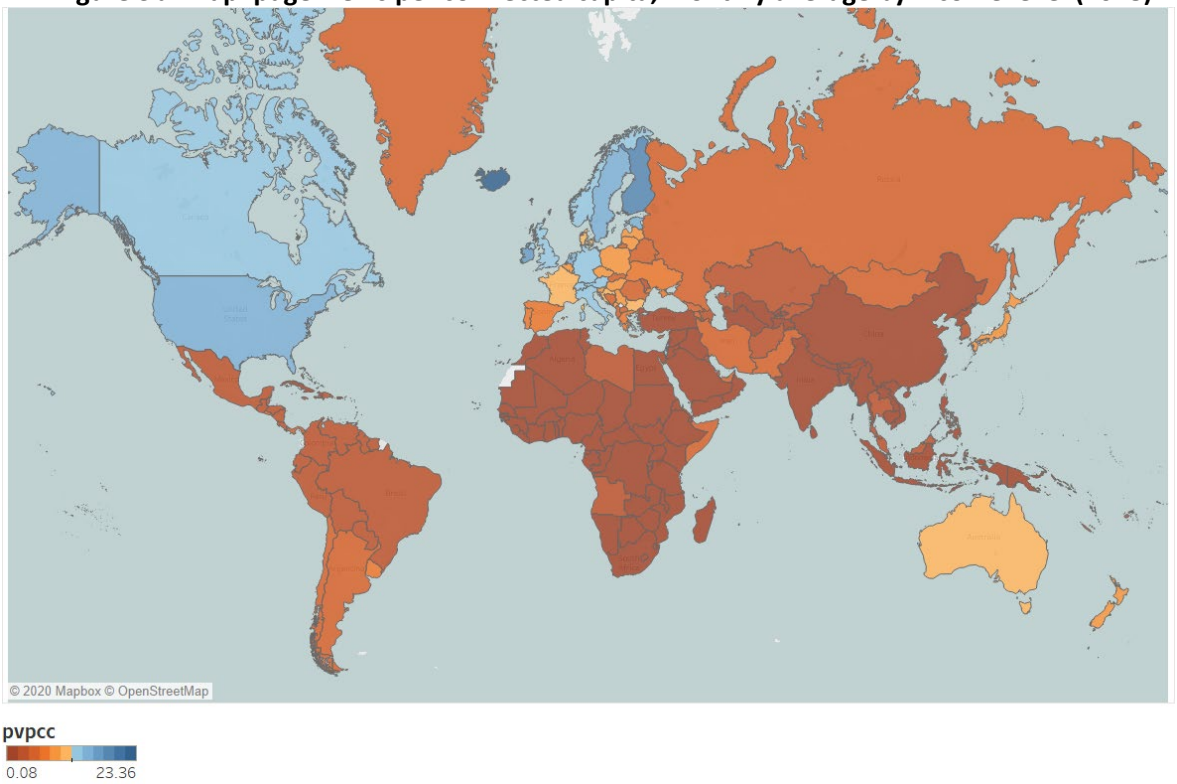
Figure: 3a. Box plot: page views per connected capita, monthly average by income level (2018)



Labels: ISL=Iceland; FIN=Finland; SWE=Sweden; USA=United States; GBR=United Kingdom; ISR=Israel; JPN=Japan; ESP=Spain; ARE= United Arab Emirates; KOR= South Korea; OMN=Oman; BGR=Bulgaria; UKR=Ukraine; RUS=Russia; IND=India; CHN= China; SOM=Somalia; LBR=Liberia; ETH=Ethiopia.

[11] Considering global internet divides where individuals in high-income countries have higher internet penetration rates compared to the rest of the countries, the decision was to estimate "per connected capita" to allow fair comparisons. Section F presents estimations based on all individuals in a country.

Figure 3b: Map: page views per connected capita, monthly average by income level (2018)

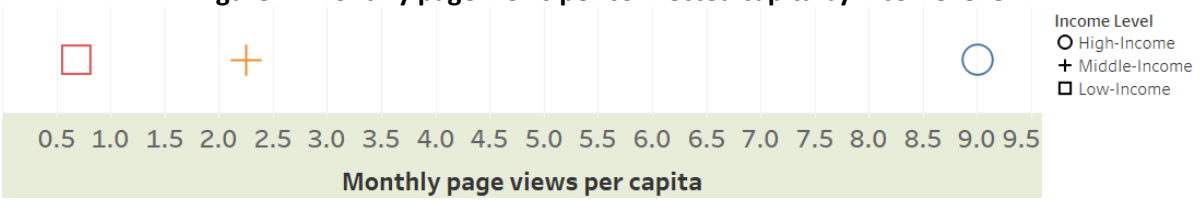


The boxplot in Figure 3a is useful to identify two divides: the first thing that stands out is the disparities in the consumption of knowledge between countries; and second, the disparities between and within income levels. Overall, it is possible to identify that high-income countries tend to have higher *pvpcc* when compared to the other two groups.[12] However, some regional divides emerge as Middle East & North African (MENA) countries have very low *pvpcc* within the group of high-income countries.[13] Surprisingly, South Korea has a very low *pvpcc* value within the same income group due to competition from another wiki in the country. [14] Middle- and low-income countries have less spread-out distributions with more countries concentrated in the lower part. The map in Figure 3b helps to illustrate income divides between countries and regions. At this point considering the skewness of the distributions, medians are a more adequate measure of average.

Figure 4 presents the estimated medians of monthly page views per capita when aggregating countries by income levels.

[12] Some outliers within each income level are present like in the case of Iceland (ISL=24 avgpvpc), Bulgaria (BGR=10 avgpvpc), and Somalia (SOM=4 avgpvpc).
[13] Issues in the consumption of Wikipedia contents might be related to those on the production side identified by Graham and Hogan (2014) including the font used in the Arabic Wikipedia.
[14] In South Korea Wikipedia faces high competition from Namu.Wiki (Alexa, 2020).

Figure 4: Monthly page views per connected capita by income level

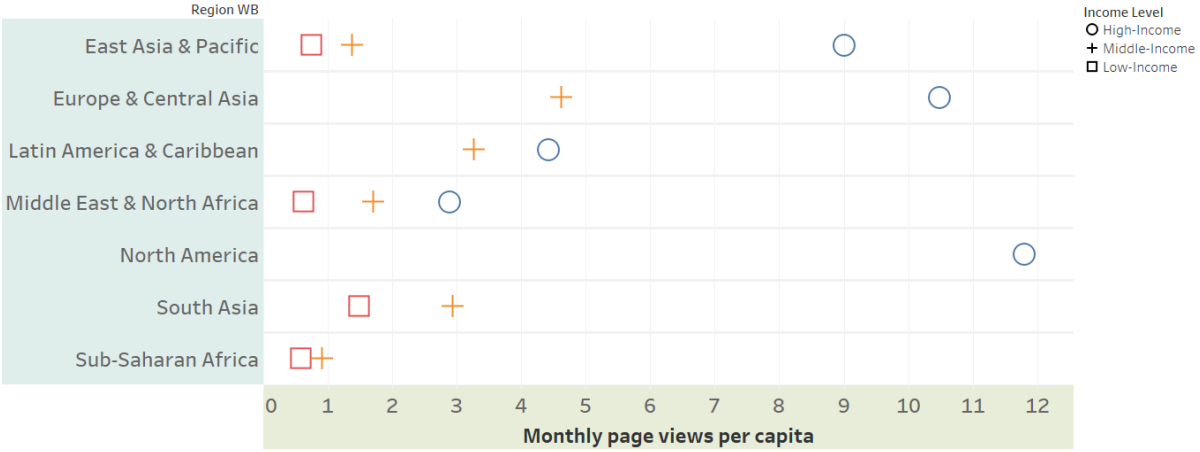


Note: Medians of country averages

Sharp divides in the consumption of knowledge between income levels are summarised in Figure 4. Connected individuals in high-income countries are viewing around nine pages per month, a little more than two pages in middle-income countries, and in low-income countries less than one page. Statistically significant divides in the consumption of knowledge were confirmed using a Kruskal-Wallis test (see Appendix 4).[15] When comparing the consumption of connected individuals, those in high-income countries tend to consume more and this confirms the existence of divides between income levels. Divides between income levels are not surprising; however the use of big data allows the quantification of knowledge consumption divides that are not possible to estimate with this level of completeness using traditional data.

Disparities between and within regions started to reveal in Figure 3; now, these regional divides are explored by income level in Figure 5.

Figure 5: Monthly page views per connected capita by income level and region



Note: Medians of country averages

Some interesting patterns are confirmed with Figure 5. First, high-income countries are the ones consuming the most within all regions; however, those in Latin America & the Caribbean (LAC) and the MENA regions are consuming less than half when compared to Europe & Central Asia and North America.[16] The disaggregated data shows that within each region the pattern is consistent with high-income countries consuming the most, followed by middle-income, and low-income countries; however, disparities are broad between

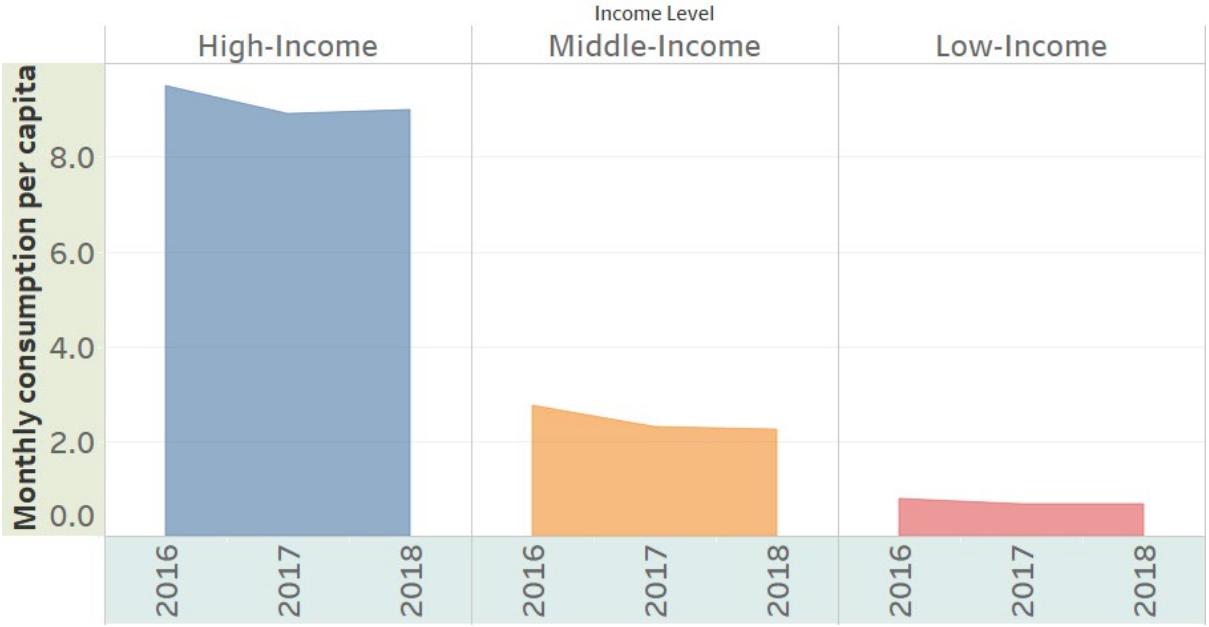
[15] In this case using the Kruskal-Wallis test, H_0 (knowledge consumption is equal across all income levels) is rejected in favour of H_a (knowledge consumption is not equal across all income levels).

[16] In the case of high-income countries in the LAC region, some of these are 'new graduates' from the middle-income category which might explain their lower consumption when compared to other high-income countries.

regions.[17] Overall, big data-related divides are identified with the MENA and Sub-Saharan Africa regions consuming less knowledge when contrasted with the other groups.

The Wikipedia API allows access to some historical data that is useful to explore the consumption per capita across time. In this case the medians of the monthly page views are estimated for the years 2016-2018. Figure 6 is helpful to visually identify a couple of trends in this short period of time.

Figure 6: Monthly page views per capita by income level (2016-2018)



Notes: For all years countries are classified by income level based on the 2018 DAC-OECD catalogue
Medians of country averages

Two interesting trends emerge from Figure 6. First, when looking at the initial and final years the per capita consumption has decreased for the three groups; however, with only three years available it is not possible to confirm a trend. Second, the decrease in the per capita consumption of Wikipedia contents by the group of high-income countries might suggest that divides with the other groups are closing, yet, middle-income and low-income countries have not increased their per capita consumption of knowledge. Divides between income levels remain almost without change. These trends might be the result of existing users consuming less knowledge and/or new users less interested in consuming knowledge from Wikipedia – this is addressed in the discussion in Section G2.

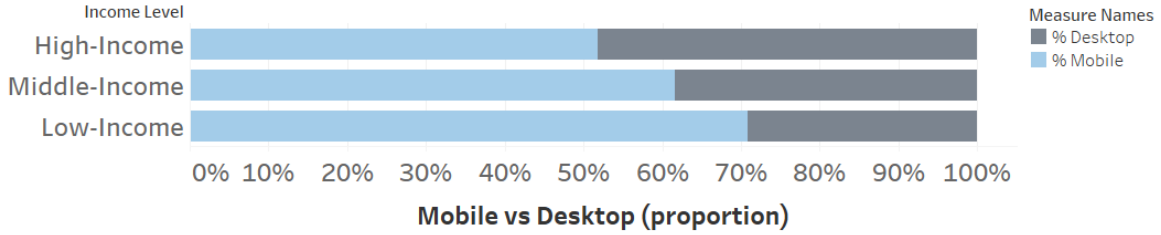
Wikipedia data appears valuable to measure divides in the volumes of particular types of knowledge consumed globally, and the knowledge per connected capita consumed between countries, income levels, and regions. These measures of knowledge consumption cannot be performed with the use of traditional data.

[17] These patterns are very similar, for the majority of the regions, to those related to GDP per capita.

D3. Technological Divides in the Consumption of Knowledge

The role of mobile devices in middle-income and low-income countries has been highlighted due to the opportunities for development brought by these devices. Mobile devices give the possibility of acquiring knowledge to individuals that do not have access or the skills to use a desktop computer. Using Wikipedia data it is possible to estimate the proportion of pages that were viewed using a desktop or a mobile device; first by income level and region for 2018; and then by income level for 2016-2018. Figure 7 shows the aggregated results of page views by device and income level.

Figure 7: Page views by device and income level, country medians mobile vs. desktop 2018 (%)



Note: Medians of country aggregates

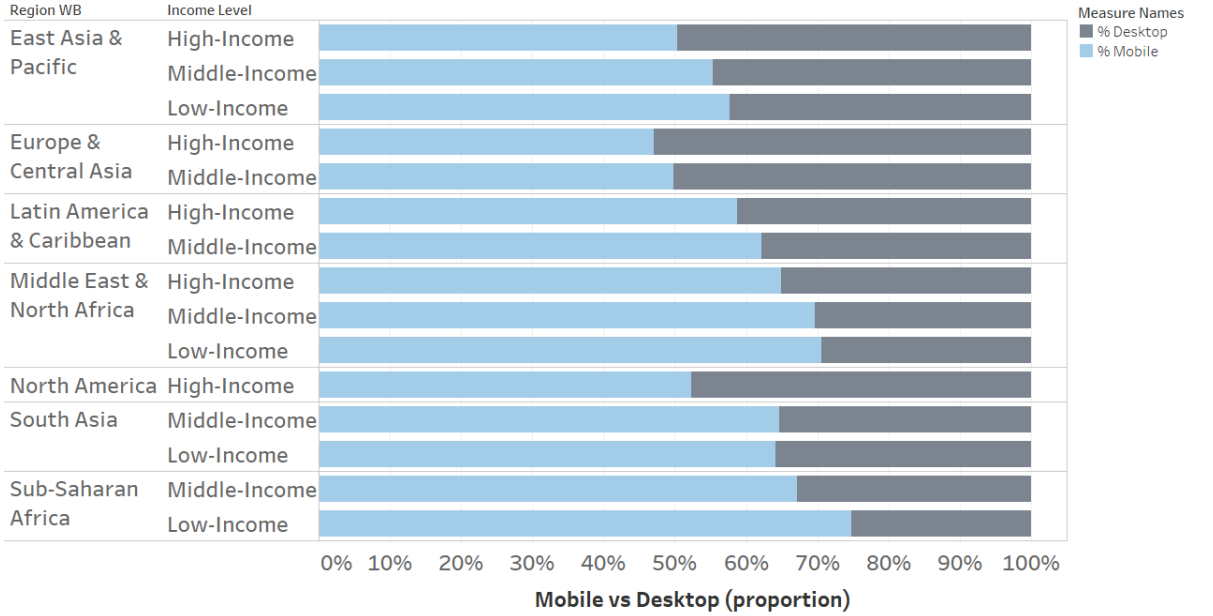
Technological divides in the way individuals access Wikipedia contents are identified in Figure 7. The group of low-income countries consumes the majority of their knowledge using mobile devices (71%), followed by middle-income countries (62%) and high-income countries (52%); this divide appears statistically significant using a Kruskal-Wallis tests (see Appendix 4).[18] Low-income countries and middle-income countries might be benefitting from the opportunities brought by mobile devices; technological divides are clearly identified between income levels. These divides are very similar to the income divides between countries but less pronounced, and to some extent similar to measures of computer ownership.[19] These technological divides interacting with knowledge divides have not been measured using traditional data.

Similarly to the breakdown presented in the previous section, Figure 8 shows technological divides by regions and income level.

[18] In this case using the Kruskal-Wallis test, H_0 (knowledge consumption using a mobile device is equal across all income levels) is rejected in favour of H_a (knowledge consumption using a mobile device is not equal across all income levels).

[19] The correlation between the consumption of knowledge via desktop (Wikipedia) and the traditional measures of computer ownership at home (ITU, 2019) is 0.460.

Figure 8: Page views by device, region and income level, mobile vs. desktop (%)

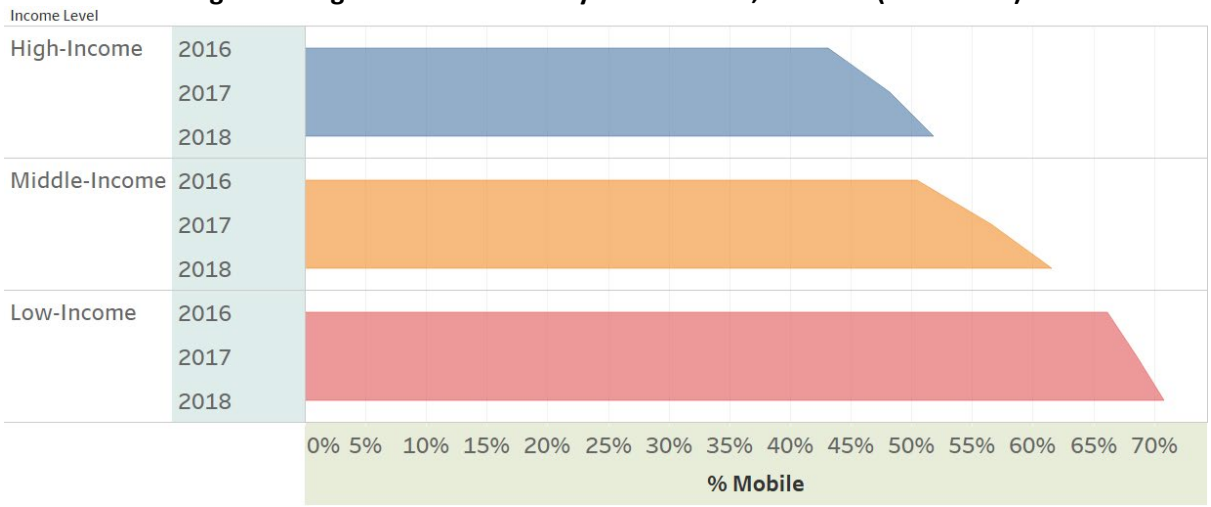


Note: Medians of country aggregates

Once more the divides between regions and income levels are revealed when presenting the page views by device, even though patterns are slightly different this time. In all the cases high-income countries seem to be consuming less content via mobile than their counterparts within each region. The proportions in middle-income and low-income countries tend to be similar in most of the cases within each region. Within all regions with low-income countries, it is those countries that are consuming more contents via mobile devices; yet the proportions fluctuate from 75 percent in the MENA region, to 58 percent in East Asia & Pacific. The use of mobile devices in low-income countries and the group of middle-income countries confirms income divides between regions; again, the trends are similar to those of the income divides between regions, yet less pronounced.

Wikipedia data allows us also to explore trends in the use of devices for 2016-2018. Figure 9 presents these trends by income level over this three-year period.

Figure 9: Page views mobile % by income level, medians (2016-2018)



Note: Medians of country aggregates

The proportion of pages viewed using mobile devices has been clearly increasing for all income levels, as shown in Figure 9. The group of middle-income countries is the one that shows the largest increase (12 percentage points), followed by the group of high-income countries (9 percentage points); low-income countries had the smallest increase (3 percentage points); however, the proportion is still the highest with nearly 70 percent of all contents accessed via mobile devices. The growth in middle-income and low-income countries might be reflecting the fact that new users are connecting only using mobile devices; in the case of high-income countries, it might be showing that users are substituting the use of desktop computers with mobile devices.

In sum, consumption data are valuable to understand how individuals access knowledge and confirm the importance of mobile devices in middle-income and low-income countries. Traditional data does not capture how knowledge is consumed so the use of Wikipedia data sheds some new light through these figures. In terms of policy these results are valuable to track the use of mobile devices and to define policies oriented to the development of capabilities that can help lower-income countries to benefit from the use of mobile devices to leapfrog and participate in the knowledge economy. The next section focuses on language divides in the consumption of knowledge.

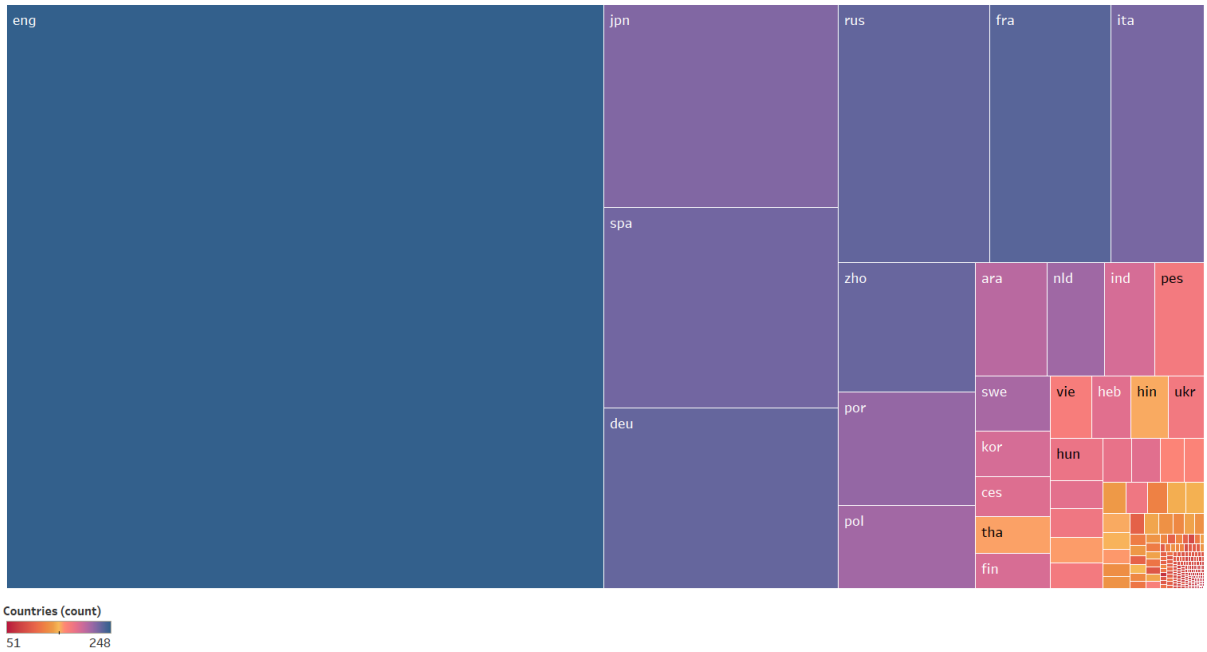
E. Language Divides in the Consumption of Knowledge

After analysing the consumption of knowledge focusing on inter-country divides, this section analyses consumption from the language standpoint. To identify language divides in the consumption of knowledge three main elements are explored: the volume of page views consumed for each Wikipedia (language); the per capita consumption of contents based on the number of speakers of each language; and the consumption considering the size of the Wikipedias (articles available) and country income levels. In this section the results are presented for the year 2018; due to limitations in the access to historical data about languages (traditional data) some trends (2016-2018) are just included in Appendix 5.

E1. Volumes by Language

The analysis of the volume of page views in 2018 is useful to start exploring inter-language divides in the consumption of knowledge and identify disparities based on the number of countries consuming each one of the 247 Wikipedias (languages) in this study. The volume of page views by language (shape size) and the count of countries (colour scale) that consumed each language is presented in Figure 10.

Figure 10: Page views by language and income of primary country of language (2018)



Labels using ISO 639-3: eng=English; jpn=Japanese; spa=Spanish; deu=German; rus=Russian; zho=Chinese; por=Portuguese; pol=Polish; fra=French; ita=Italian; ara=Arabic; nld=Dutch; ind=Indonesian; pes=Persian; swe=Swedish; vie=Vietnamese; heb=Hebrew; hin=Hindi; ukr=Ukrainian; kor=Koreanic; hun= Hungarian; ces=Czech; tha=Thai; fin=Finnish.

Figure 10 is helpful to measure two types of divides that are not possible to estimate using traditional data: the disparities in the consumption volumes between languages and global divides based on the number of countries consuming each language. Language divides are clear when looking at the size of the shapes with major Indo-European languages and a few other languages from East Asia dominating the consumption. This is the case of the English Wikipedia with a monthly average of 7.7 billion views, almost half of all the page views, followed behind by the Japanese and Spanish Wikipedias with 1 billion views, the German (927 million), and Russian (858 million).[20] A second type of divides emerges when looking at the number of countries consuming each language (colour scale), with a very similar pattern: major languages are consumed by more countries (e.g. English was consumed by all countries); languages with fewer page views were also consumed by users from fewer countries. Overall, major Indo-European languages and the English Wikipedia particularly seem to have an important share in the consumption of contents in terms of volume and global views, with broad divides when compared to smaller languages.

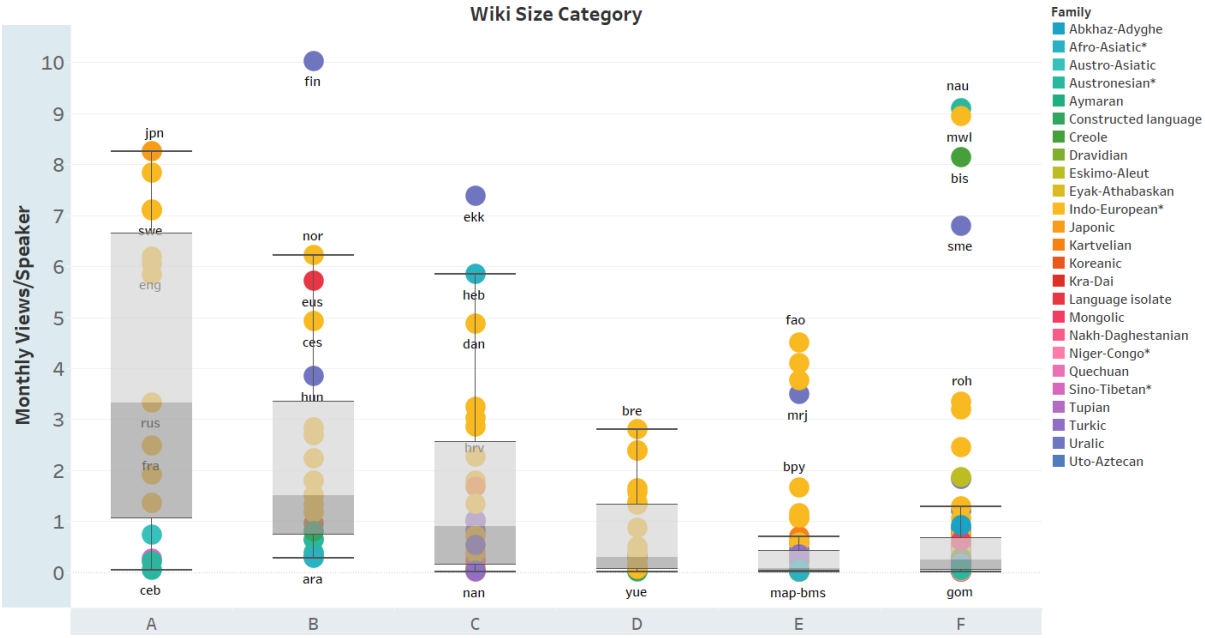
E2. Page Views per Language Speaker

To further explore these inter-language divides, the per capita consumption of Wikipedia contents is estimated considering the number of speakers of each language. As described in Section C2 these estimates include the total number of speakers (first and second language [L1, L2]) and the language families. For this case, it was not possible to access the data

[20] It is important to keep in mind that at the end of 2018 more articles were available in English, almost 6 million articles, followed by the Cebuano (5 million), Swedish (4 million), German, French, and Dutch projects (2 million). An outlier in this top-5 is the case of the Cebuano Wikipedia with around 5 million articles translated from other projects (Tkacz, 2011).

disaggregated by country, and there are no known estimates available about the number of speakers of each language that are connected to the Internet. The estimates in this sub-section are related to 219 languages out of the 247 Wikipedias due to the available estimates and the exclusion of extreme outliers.[21] Figure 11 presents boxplots for each Wikipedia size category (articles available) (horizontal axis: A-F) and the estimated average monthly views per speaker are presented for each language on the vertical axis; colours represent language families.

Figure 11: Box plot page views per language speaker by wiki size



Categories:
A = 1 million or more articles
B = 250,000 – less than 1 million
C = 100,000 – less than 250,000
D = 50,000 – less than 100,000
E = 10,000 – less than 50,000
F = 1,000 – less than 10,000

Labels: jpn=Japanese; swe=Swedish; eng=English; rus=Russian; fra=French; ceb=Cebuano; fin=Finnish; nor=Norwegian (Bokmål); eus=Basque; ces=Czech; hun=Hungarian; ara=Arabic; ekk=Estonian; heb=Hebrew; dan=Danish; hrv=Croatian; nan=Min Nan; bre=Breton; yue=Cantonese; fao=Faroese; mrj=Hill Mari; bpy=Bishnupriya Manipuri; map-bms=Banyumasan; nau=Nauruan; mwl=Mirandese; bis=Bislama; sme=Northern Sami; roh=Romansh; gom=Goan Konkani

Once more, box plots are helpful to identify divides not estimated with traditional data. First, when comparing the six box plots aggregating language categories (A-F) in Figure 11, it is clear when looking at the medians that those Wikipedias (languages) with more articles available have on average more views per speaker.[22] Second, within each category size there are disparities showing some languages being consumed more per speaker; usually, languages whose primary country is categorised as high-income (more on this in the next section). Low views/speaker rates might be reflecting other divides for languages consumed in countries with low Internet connectivity rates (e.g. Cebuano in Philippines) or Wikipedia's

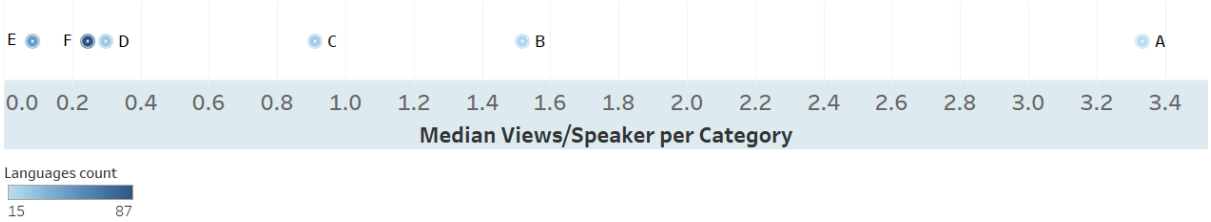
[21] 15 languages with more than 10.1 or less than 0.00 views/speakers per month were excluded from the analysis; no data were available for 13 languages.
[22] Overall the correlation is positive.

popularity issues (e.g. Arabic in Arab countries).[23] Additional language divides are revealed where Indo-European languages are almost half of all the languages in the dataset, with higher consumption rates.

Outliers outside the top whiskers present additional insights. Some Uralic languages appear among the ones with higher views per speaker like in the case of Finnish and Estonian, both from high-income countries. When looking at outliers in categories E and F the estimates of the number of speakers for some languages seem to be underestimating the real number of speakers of these languages.[24] This might be one of the reasons that multiple languages appeared as outliers with higher consumption rates; however, this might also be showing high interest in these minor languages. Overall, languages with the highest average monthly views per speaker are those whose primary country is categorised as high-income. Similar knowledge-language divides have not been identified before with this level of detail using traditional data.

The summaries of knowledge consumption for each one of the six Wikipedia size categories are presented in Figure 12 using medians per category. The colour scale indicates the number of languages within each category.

Figure 12: Median page views per language speaker by wiki size



The median page views per language category reflect the point that Wikipedias with more articles are being consumed more per estimated speaker of that language; again this might be a reflection of the fact that Wikipedias with more articles are being consumed in countries where more individuals are connected to the Internet; this will be analysed in more detail in the following sections.

Overall, the combination of big data from Wikipedia and traditional data estimates about speakers, allow the exploration of inter-language divides that are not possible to analyse using only traditional data. Cultural and digital policies that are reflected in the consumption of knowledge can be reinforced and tracked using these measures. For instance, policies associated to the analogue foundations of knowledge such as educational attainment and literacy can be strengthened; also policies related to improving online connectivity – such as improving coverage and lowering data costs – and digital literacy, could lead to higher knowledge consumption.

[23] For more about issues in the production of contents in the Arabic Wikipedia see Graham et al. (2014) and Graham and Hogan (2014).

[24] For instance the Cornish Wikipedia – excluded from the graph – has an average of 227 pages consumed per speaker per month, based on only 600 speakers being estimated by Ethnologue. Wikipedia data might be helpful to identify language speakers around the world in order to adjust some of these estimates.

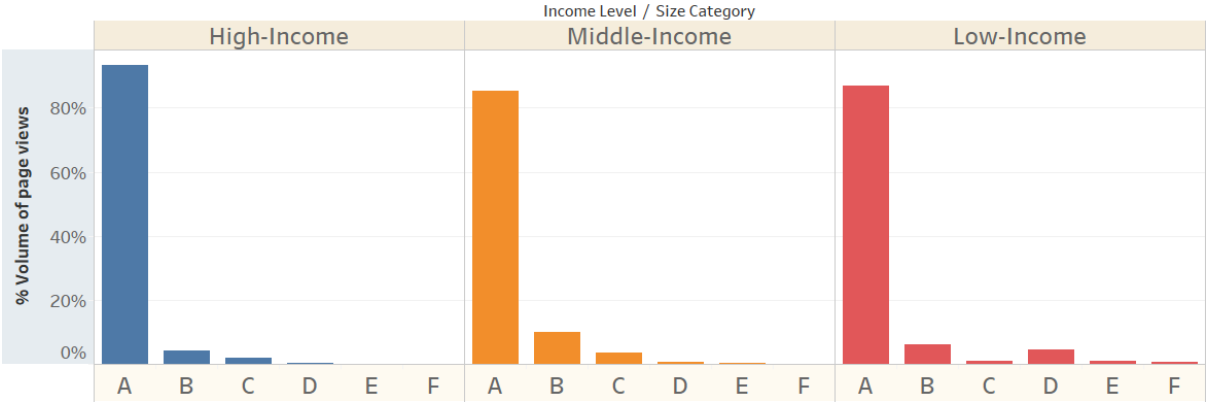
E3. Income – Language Knowledge Divides

Finally, the language divides for different income levels from the country standpoint are analysed. Two approaches are followed: first analysing for each income group the proportion of contents consumed from each Wikipedia category (size). Then, the opposite is investigated: for each of the six Wikipedia size categories the income levels are used to identify the share of contents that each group of countries are consuming.

- **Views by income level and Wikipedia size category**

To better understand the volumes consumed in 2018, the contents are aggregated by income level and broken down by Wikipedia size. In Section D1 the results showed that, overall, the majority of contents (71.0%) were consumed by high-income countries, 28.1 percent by middle-income, and less than one percent by low-income countries. Figure 13 presents on the top of the x-axis the income level (three panels); the bottom of the x-axis indicates Wikipedia size (A-F); the vertical axis shows the percentage of page views for each category within each income level. The analyses of trends for 2016-2018 are available in Appendix 5.

Figure 13: Page views by income level and wiki size (2018)



Note: percentages add up to 100 percent within each income level.

Figure 13 is useful to identify patterns within the three income levels and compare between them. When aggregating by income, countries consume almost all of their contents from Category A languages. Low-income countries have a little more diverse consumption patterns from minor languages, and categories D-F appear as more relevant for this group – more about this is discussed in the next paragraph. Despite the big disparities in the volumes consumed by income level, Category A languages clearly dominate the consumption within the three income groups.[25] In the aggregate, countries from all income levels seem to be consuming more contents from Category A languages; those with more articles. Divides are evident within all income groups.

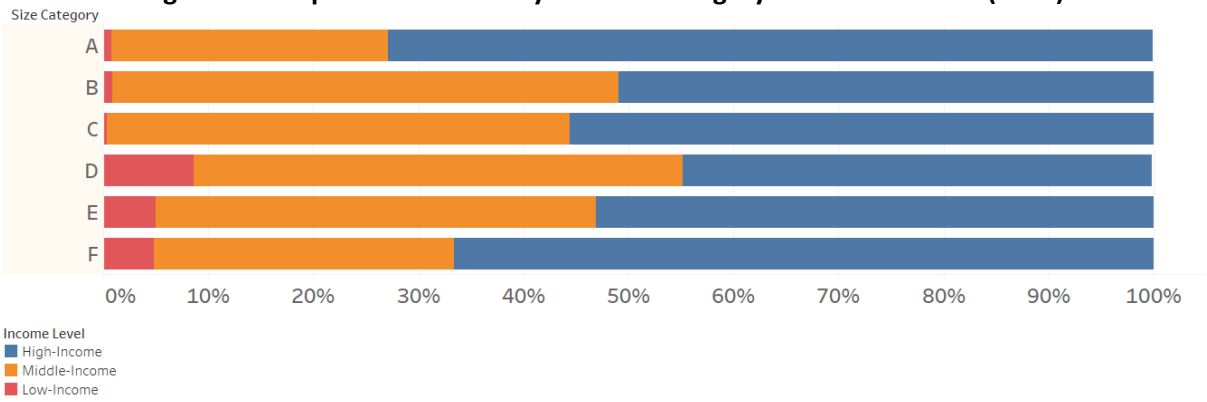
- **Views by Wikipedia size category and income level**

To further examine the previous findings, the analysis of consumption broken down by Wikipedia size is conducted, to find what type of countries are consuming each Wikipedia

[25] These analyses were also conducted excluding the English Wikipedia with very similar results.

category. Despite the fact that major languages (Category A) dominate the consumption volumes (90% of the total), it is important to analyse and compare all categories. This analysis is conducted for 2018; the analysis of trends between 2016-2018 is available in Appendix 5. Figure 14 shows the proportion of consumption by income level as a colour on the x-axis for each Wikipedia size category in the y-axis.

Figure 14: Proportion of views by wiki size category and income level (2018)

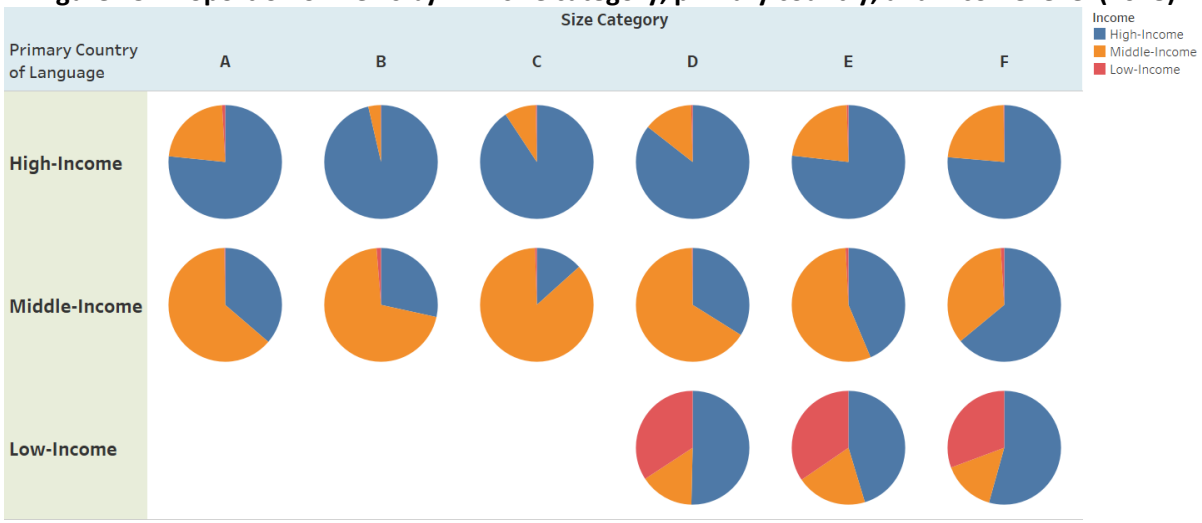


Categories: A = 1 million or more articles; B = 250,000 – less than 1 million; C = 100,000 – less than 250,000; D = 50,000 – less than 100,000; E = 10,000 – less than 50,000; F = 1,000 – less than 10,000.
Note: percentages add up to 100 percent within each Size category.

The majority of the contents in all but one of the size categories in Figure 14 are consumed by individuals in high-income countries, with more than two-thirds of contents in languages A and F being consumed in high-income countries. Then, middle-income countries consume a big share of B-E languages. Finally, low-income countries show an interesting pattern with almost no presence in languages A-C, but with a more important presence in smaller languages D-E. In the case of category D, the Bengali language dominates this category with an important share of consumption from Bangladesh, a low-income country. Despite the low volumes of consumption, and the lower number of articles available in Wikipedias D-F, low-income countries emerge in these categories.

The combination of traditional data and big data allows further exploration of these divides considering the income level of the primary country of the language. Overall, a little more than 85 percent of all knowledge consumed is from languages whose primary country is a high-income country; on the other hand, languages whose primary country is a low-income one represent less than one percent of all the contents consumed. Figure 15 presents the proportion of views by Wikipedia size (x-axis), income classification of the primary country of each language (y-axis), and the income level of countries consuming these contents (slice colour).

Figure 15: Proportion of views by wiki size category, primary country, and income level (2018)



Note: All pie charts add up to 100% within 'Size Category' and 'Primary Country of Language'.

Multiple income divides in the consumption of knowledge are revealed. The first thing that stands out is that languages whose primary country is categorised as high-income are being consumed mostly by high-income countries (first row, blue slices). Second, languages from middle-income countries are consumed by mostly middle-income countries in all categories but one (F) (second row, orange slices).[26] Third, languages from low-income countries (e.g. Afro-Asiatic, Creole and Niger-Congo languages) are only represented in the smallest size categories (D-F) and the majority of their contents are consumed together by high- and middle-income countries; yet low-income countries have a relevant share here (third row, red slices). Figure 15 confirms divides where languages from low-income countries have less knowledge available in Wikipedia, and these countries are not the ones consuming most of these contents. Overall, smaller Wikipedias representing languages of high-income countries are being consumed mostly by these high-income countries; the smallest Wikipedias from languages of non-high-income countries are consumed in a very high proportion by what are perhaps immigrant populations in high-income countries.

The combination of traditional data and big data allows a detailed exploration of language-income divides in consumption of the forms of knowledge held on Wikipedias. In general, Wikipedias representing languages from low-income countries have less knowledge available, very low consumption volumes, and are consumed by a smaller number of countries. However, the consumption rates per speaker show the interest in the consumption of minor languages; smaller Wikipedias seem to be the source of knowledge for migrant populations in non-low-income countries. Language divides in the consumption of knowledge are confirmed as speakers of major languages are benefitted with the fact that there is more knowledge available for them (mostly Indo-European); this is reflected in the consumption volumes and per speaker of these languages. Overall, these results can help to justify the support that digital, cultural, and development policies require to improve the digital representation of knowledge from small and endangered languages.

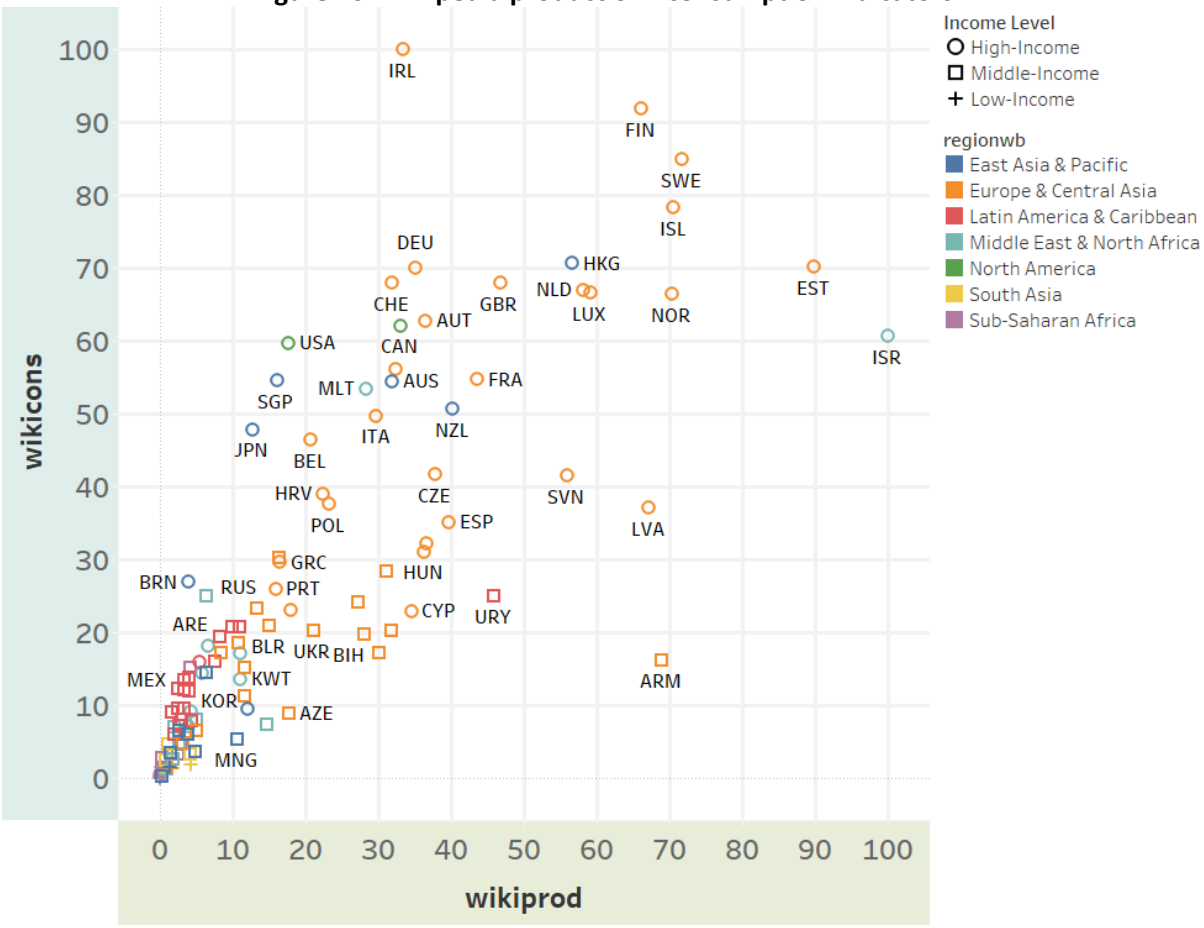
[26] Around three-quarters of all the contents of languages such as Buginese and Yoruba are consumed in high-income countries. In terms of volume Egyptian Arabic is highly consumed in the US and Saudi Arabia.

F. Wikipedia Production – Consumption Index

The previous two sections have focused on analysing knowledge divides on the consumption side between countries and between languages. As mentioned in Section B2 consumption divides have not been widely explored using Wikipedia and the majority of the research and measures focused on the production side. Now that sections D and E have provided a detailed picture of the consumption divides it is important to get a sense of how knowledge production and consumption look together at the country level in order to understand how these correlate – or not. Unfortunately the Wikipedia API does not allow access to the edits/production data disaggregated by country so this case relies on the use of data and methodology utilised by the GII (Cornell University et al., 2018). As described in Section C3, GII estimates from Wikipedia a production indicator (*wikiprod*) using yearly page edits (per million population, 15–69 years old) in a country, and rescales these values to 0 – 100. Following this approach a consumption indicator (*wikiconsum*) is estimated using Wikipedia yearly page views in 2017 (per million population, 15-69 years old). The estimation of both indicators allows a contrast of production and consumption in a selection of 125 countries in 2017. It is important to highlight that the GII is estimated for all individuals in a country and not only for connected individuals as done in Section D2; this will somehow 'penalise' countries where not many individuals are connected to the Internet. Values for both indicators for all countries can be found in Appendix 6.

Figure 16 presents a scatterplot with the estimations of the two indicators; *wikiprod* based on page edits is presented on the x-axis; *wikiconsum* showing page views is plotted on the y-axis.

Figure 16: Wikipedia production - consumption indicators



Labels: IRL=Ireland; FIN=Finland; SWE=Sweden; ISL=Iceland; DEU=Germany; HKG=Hong Kong; EST=Estonia; CHE=Switzerland; GBR=United Kingdom; NLD=Netherlands; Lux=Luxemburg; NOR=Norway; AUT=Austria; ISR=Israel; USA=United States; CAN=Canada; SGP=Singapore; MLT=Malta; AUS=Australia; FRA=France; JPN=Japan; BEL=Belgium; ITA=Italy; NZL=New Zealand; HRV=Croatia; CZE=Czech Republic; SVN=Slovenia; POL=Poland; ESP=Spain; LVA=Latvia; GRC=Greece; HUN=Hungary; BRN= Brunei Darussalam; RUS=Russia; PRT=Portugal; CYP=Cyprus; URY=Uruguay; ARE=United Arab Emirates; BLR=Belarus; UKR=Ukraine; BIH= Bosnia and Herzegovina; ARM=Armenia; MEX=Mexico; KWT=Kuwait; KOR=South Korea; AZE= Azerbaijan; MNG=Mongolia.

Even though there is not a perfect linear tendency between the indicators, Figure 16 shows a positive and high correlation between the two.[27] Considering a 45^{o} line from the origin of the plot the majority of the countries lie above and to the left, reflecting positively skewed distributions for both indices – more positively-skewed in the case of *wikiprod.* Low-income countries appear very close to the origin with very low scores for both indicators. The majority of middle-income countries are located in the lower-left side of the plot with the exception of Armenia with a very high *wikiprod* score – despite a somehow low *wikiconsum* score.[28] Most of the high-income countries are scattered towards the upper-right side of the plot; however, as noticed before a good number of high-income countries in the MENA region are concentrated in the lower-left side with low *wikiprod* and *wikiconsum* values. This

[27] The correlation coefficient is 0.81.

[28] The production of contents in the Armenian Wikipedia has been promoted by the Armenian government (Wikimedia, 2020c; Wikimedia, 2020d).

is also the case of South Korea with very low scores on both indicators due to competition from a local wiki.

When looking at the upper, upper-right, and far-right sides of the graph a group of high-income countries – mostly European – appear with the highest scores for both indicators. Ireland stands out as the country with the highest *wikiconsum* score but with a low *wikiprod* score – a good example of why both indicators should be considered. A group of Nordic countries appear with high scores for both indicators; in the case of Finland, the Finnish Wikipedia appeared in Section E2 as the one with the highest average consumption per speaker (size=B; 10 pages/month). Finally, Israel appears as the one with the highest *wikiprod* score very likely related to the Hebrew and Yiddish Wikipedias; Israel consumed the majority of its contents from the Hebrew Wikipedia which is also one with the highest average consumption per speaker (size=C; 6 pages/month). The divides between countries are evident when looking at this graph, as is the divide between production and consumption. These results show the value of considering both production and consumption indicators as part of composite indices measuring knowledge in order to properly reflect these disparities.

A suggested index (*wikindex*) aggregates production and consumption to rank countries and provides a new measure of knowledge divides. As mentioned before, the production of Wikipedia contents seems to be highly-skewed with some countries producing high amounts of knowledge, and some others not producing much knowledge. The aggregation of both indicators into one measure can help to get a better picture of how countries fare in the knowledge economy. *wikindex* is estimated by combining the *wikiprod* and *wikiconsum* indicators – giving equal weight to both – getting values between 0 and 100 for each country. Figure 17a presents the results of *wikindex* by income level using a box plot; Figure 17b shows the result on a map.

Figure 17a: Boxplot wikindex 2017

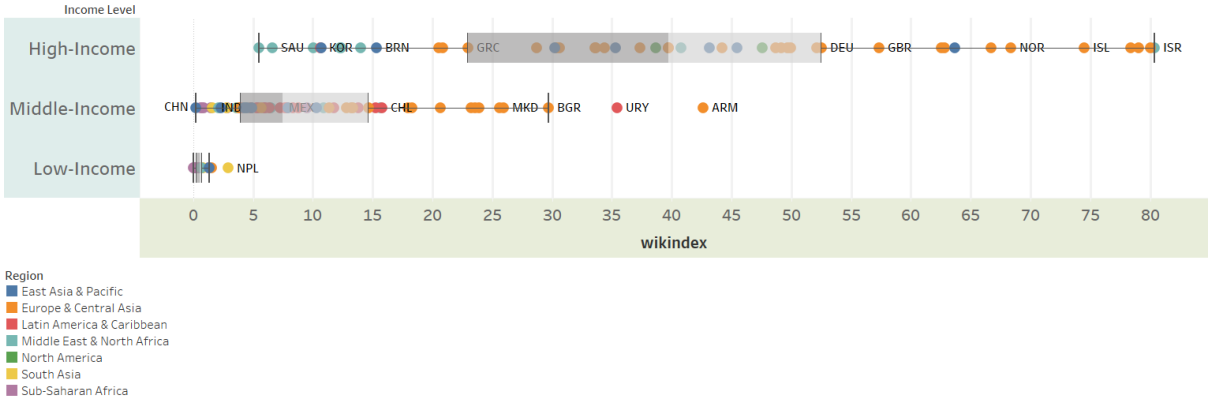
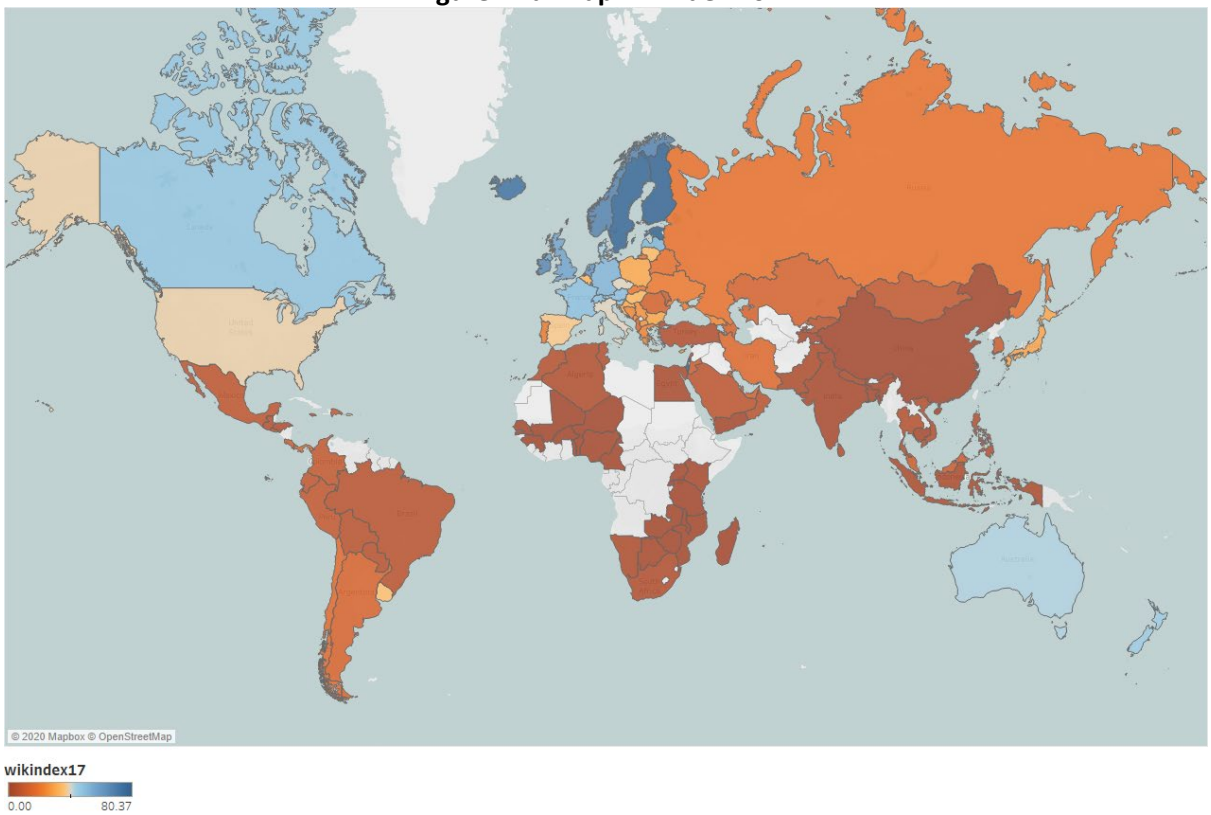


Figure 17b: Map wikindex 2017



The values of *wikindex* confirm divides but also evidence some limitations of using Wikipedia data to measure knowledge divides – including the lack of data about many low-income countries. Overall, income divides are again present between income levels with a median *wikindex* score of 39.7 for high-income countries, 7.5 for middle-income, and 0.5 for low-income countries. When looking at *wikindex* within the high-income group, disparities are very pronounced. Again a group of countries from Europe & Central Asia get very high scores whilst some countries from East Asia & Pacific and several from the MENA region get very low scores. This is showing that these countries are not consuming or producing knowledge using Wikipedia as might be expected according to their income levels. As mentioned before this is perhaps related to competition (e.g. South Korea and the Korean Wikipedia) or to a preference (e.g. MENA countries not consuming the Arabic Wikipedia). This is a limitation when considering using this index as a sole measure (proxy) to reflect the state of knowledge divides around the world; some countries might not be using Wikipedia despite having the means to produce and consume knowledge. The group of middle-income countries reflects less pronounced disparities; however, a group of countries from Europe & Central Asia, and others from Latin American take the top scores. In this group India appears with a very low score due to low Wikipedia consumption and production, perhaps due to the combination of low Internet connectivity in the country and the lack of contents in local languages; the score for China is close to zero due to censorship. In the case of low-income countries the combination of very low production and consumption scores, most likely due to very low connectivity rates, lack of contents in local languages, low literacy and education attainment, among others, place this group well behind the rest of the world.

Overall *wikindex* seems to be a useful tool to map the production and consumption of knowledge and to confirm divides between income levels. However, there are some

limitations in the use of Wikipedia explored in the previous sections and reflected in this index, such as language divides in the availability of contents, divides in the consumption rates due to competition or preference, and the impact of the connectivity divides. The estimation of this index considering all the population of 15-69 years old instead of only individuals connected to the Internet – such as the estimates in Section D2 – shows broader divides between countries and income groups. Additional discussion around these results is presented in the next section.

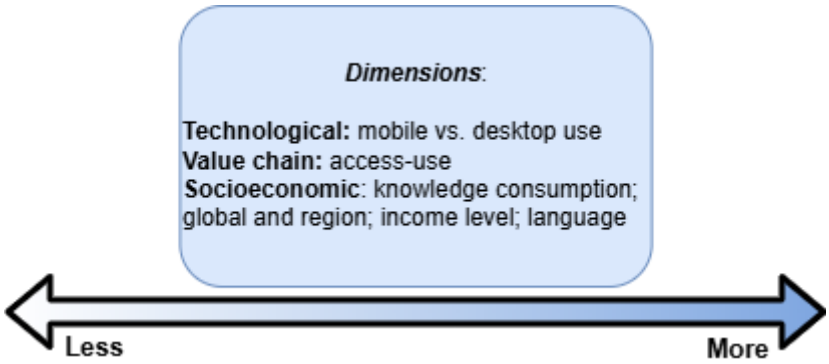
G. Discussion of Results

The results presented in the prior sections provided valuable insights around the use of Wikipedia data to measure knowledge divides. This section presents the discussion following the elements included in the conceptual framework. First the insights into the knowledge divides identified are presented; second the discussion related to the type of big data analysed and the benefits and limitations of using Wikipedia data are assessed using the ACARTA model; finally, the data divides identified in this work are summarised. These three topics are helpful to address the research question using the framework previously defined.

G1. Knowledge Divides

The use of big data allowed the identification of knowledge divides that cannot be measured using traditional data. These divides are identified between countries (global), between groups of countries (regions), by income level of these countries (high-, middle-, and low-income), and by language. In addition, a technological dimension is identified in the use of mobile phones or desktop computers, and value chain divides at the access-use level. Figure 18 presents the main knowledge divides that were identified using the conceptual framework.

Figure 18: Knowledge divides identified



As presented in the earlier sections, these divides are best identified on a more-less continuum rather than in terms of some binary divide. In addition, the multidimensional character of these knowledge divides can be attested with the multiple combinations of divides that are present at the same time. The use of Wikipedia exhaust data allows the identification of methodological and analytical insights that cannot be identified with the analysis of traditional data.

When looking at knowledge consumed, two divides stand out: a divide in the volume of contents consumed by income level and a divide in page views per connected capita. Firstly, the majority of Wikipedia contents are consumed by high-income countries, with the USA at the head. In contrast the volume consumed by middle- and low-income countries was very low, evidencing broad divides. Second, the *pvpcc* indicator confirms income divides with high-income countries consuming almost four times more knowledge per capita than middle-income countries, and almost 13 times more than low-income countries. These divides are also present between and within regions; unsurprisingly, the poorest regions face the lowest levels of consumption. However, some high-income countries – like those in the MENA region – have very low levels of per connected capita consumption. Overall, despite the fact that some individuals in middle-income and low-income countries have access to the Internet they are consuming less of this type of knowledge when compared to connected individuals in high-income countries. The estimation of these divides is valuable to understand the state of knowledge consumption as part of measures of the knowledge economy, to track the progress of countries and regions, and to define policies oriented towards closing these divides. These policies could be related to strengthening the analogue foundations of knowledge such as education and literacy, and to improve connectivity by extending coverage and lowering data costs.

The longitudinal data for 2016-2018 related to the per connected capita consumption was helpful to notice two patterns. First, the consumption of Wikipedia contents per capita seems to be decreasing for all income levels. This might be related to changes in the behaviour and preferences of current users and the arrival of a different type of Internet users. For instance, current users might be becoming less interested in consuming knowledge generally, or in the type of knowledge represented in Wikipedia or perhaps a Wikipedia competitor appeared in their countries. New users might not be interested in consuming this type of knowledge, or there is no knowledge available for them (for instance in smaller languages, as discussed below), or they have less skills/knowledge on how to consume knowledge. Second, data shows that the divides between income levels are not closing throughout this period; albeit that three years did not provide enough evidence to confirm this trend. At the beginning of 2018 Wikipedia was the 5th most popular around the world; by the end of 2024 it was the 7th (Alexa, 2021; Similarweb, 2024). The longitudinal data was helpful to identify divides; however it shows that Wikipedia log data might become less relevant and accurate in the future.

The importance of mobile devices in middle-income and low-income countries can also be confirmed with these data. Technological divides in the way individuals consume knowledge using desktop or mobile devices become evident with the analyses by income level. In the aggregate, countries consume the majority of Wikipedia contents using mobile devices, with low-income ones consuming almost three-quarters of all their contents in this manner. In addition, the longitudinal data shows that over time knowledge is being consumed more and more via mobile devices – within all income levels. This is consistent with the measurements coming from other big data sources that show a decline in the use of desktop devices in favour of mobile technology (Statcounter, 2020). Overall, it is possible to confirm the importance of mobile devices in low-income countries and middle-income countries in the consumption of knowledge, and the increasing preference for the use of mobile technology at the global level. The use of mobile devices might help in reducing or

avoiding the amplification of knowledge divides. These measures might be a helpful element to track broadband policies and specific programmes that promote the consumption of online knowledge, and to strengthen policies oriented towards the generation and improvement of capacities that can help to grab the benefits of mobile devices.

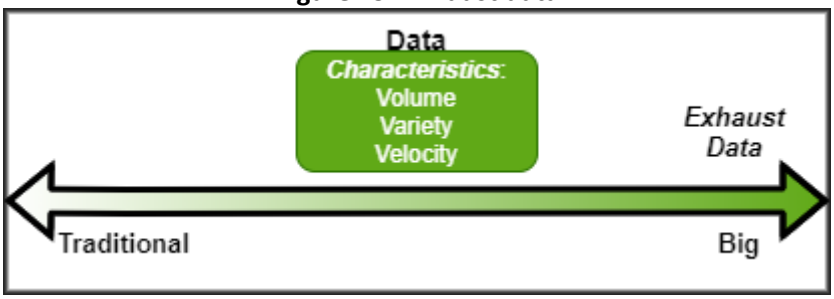
The combination of big data and traditional data about languages exposed additional knowledge divides. First, when looking at language divides in the consumption of knowledge it was clear that major Indo-European languages and a few East Asian languages led consumption; the role of high-income countries as major consumers was undeniable – mostly Western countries. Estimates per language speaker show that Wikipedias with more articles are consumed more; yet there are some languages with a high number of articles available but low consumption rates (such as the case of Vietnamese or Arabic). Surprisingly, these estimates show that there is interest in the consumption of minor languages from all over the world (such as Faroese or Nauruan). Second, the combination of income variables with these data confirmed the disadvantages of low-income countries, as languages from these have less knowledge available and very low consumption volumes; yet these smaller Wikipedias seem to be the source of knowledge for migrant populations outside low-income countries. Even if some major Western and East Asian languages dominate the consumption of knowledge, benefitting those who can read these languages, there seems to be a rising interest to consume minor languages. However, the disparities in volumes consumed and content available between major and other languages pose a challenge for the democratisation of knowledge and the reduction of knowledge divides. These results are a useful tool to define and support digital, cultural, and development policies following non-Western views, to promote knowledge and the representation of minor and endangered languages, or languages with very few Wikipedia contents.

The value of exploring knowledge production and consumption became evident when estimating a consumption indicator and a joint index to aggregate knowledge production and consumption. The estimation of *wikindex* based on *wikiprod* and *wikiconsum* indicators confirmed knowledge divides between income levels and revealed some limitations of these data. In general, high-income countries dominate both production and consumption and get higher *wikindex* scores. Again, middle-income countries appeared behind with lower scores, and low-income countries got the lowest scores. Yet, outliers on the positive and negative sides were helpful to confirm the limitations of Wikipedia data. For instance, South Korea has a very low Wikindex – as a result of local competition – which is clearly not a reflection of the state of knowledge in the country; the blockage in China makes this index an irrelevant measure in this country. Conversely, Armenia shows a high Wikindex score possibly as a result of government support for the production of contents; yet the interest in the consumption of Wikipedia content in the country, and contents from the Armenian Wikipedia was low (Wikimedia, 2020c; Wikimedia, 2020d). The production and consumption of knowledge in Wikipedia is biased as a result of the profile of the creators of contents, competition, barriers, or incentives to production and consumption; thus, the state of knowledge in some countries might not be accurately represented. Despite these caveats, a Wikipedia production-consumption index appears as a valuable element to be included as part of measures of the knowledge economy and to explore and track divides.

G2. Exhaust Big Data, Benefits and Limitations

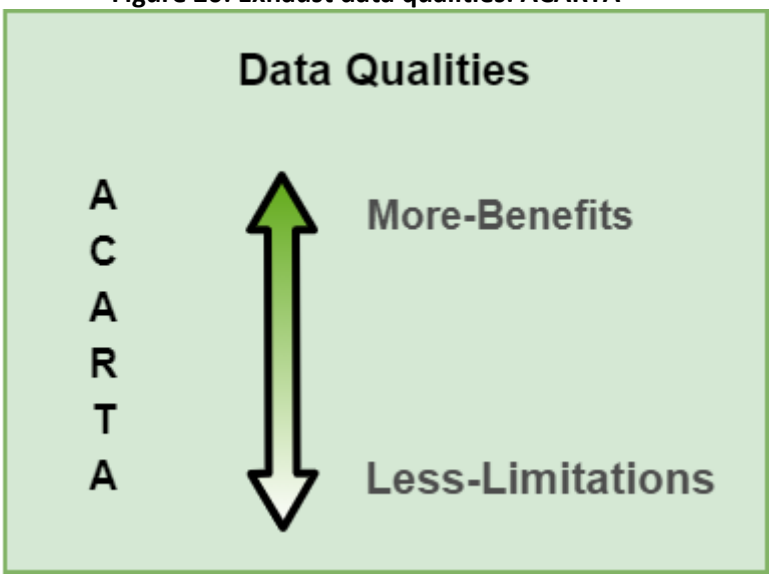
This paper identified the use of exhaust data generated as a byproduct of the consumption of Wikipedia contents as the basis of the analyses of knowledge divides. The conceptual framework considers two opposites along a continuum: traditional data and big data. In this case the use of exhaust data from Wikipedia – 560 billion page views – represents a case where big data is used to measure digital divides. Figure 19 presents the segment of the conceptual framework that places exhaust data on this continuum.

Figure 19: Exhaust data



Overall, Wikipedia big data used to measure knowledge divides can be considered as high-volume, is produced at high speed, and incorporates a variety of data types. Wikipedia data presented some benefits and limitations when used to measure knowledge divides. The analysis of these benefits and limitations is presented below following the ACARTA model.

Figure 20: Exhaust data qualities: ACARTA



Availability- At the country level, Wikipedia big data are more widely available than traditional data about the particular type of knowledge studied here. These data are generated for all countries, territories and small islands which makes Wikipedia a valuable resource, considering that some of these are not usually included in traditional datasets. From the language standpoint all major languages are included in the analyses; however, the almost-250 languages analysed are still far behind the near-8,000 languages recorded. Big data-related divides are identified where minor languages – most likely from lower-

income countries – are not represented in Wikipedia, excluding them from getting knowledge and being measured.

Completeness- Wikipedia data are less complete than traditional data. Wikipedia data focuses on a particular type of knowledge and provides measures related to consumption volumes, language, and preferred devices aggregated at the country level. Important knowledge divides that can be captured by traditional data – for instance using surveys – such as gender and age divides are not identified with these big data. Wikipedia consumption data was useful to measure only a few divides; yet, when combined with other traditional data these big data proved to be valuable.

Accuracy- Using Wikipedia, the accuracy of big data is expected to be higher than when measuring using traditional data. Considering that Wikipedias are supported by the structure provided by the Wikimedia Foundation (a non-profit) there are no commercial interests involved in the generation of the data used in these analyses. Also, the dataset used is the result of the aggregation of billions of records and is not affected by typical data collection efforts. Instead of relying on data provided by NSOs or other agencies, in this case, the data are methodologically homogeneous as it comes from a single source. However, accuracy might be affected as a result of biases in the consumption of Wikipedia contents. Disparities in the number of articles created in each language and differences in Internet access rates across countries can impact the consumption of knowledge. In addition there might be some biases in the consumption of knowledge derived from the profile of Wikipedia users based on characteristics such as gender, age, geography, among others.[29] Finally, Wikipedia faces competition from other sources of knowledge which might vary across countries.

Relevance- As discussed in Section B there is no unique definition of knowledge and knowledge divides have been measured using a variety of indicators, designed by international organisations and mostly related to Western countries. These indicators are focusing typically on the production side. The use of big data allows measurement of real individual knowledge consumption adding a range of subdimensions – technological, language, geography – that make these more relevant than traditional data, when measuring these subdimensions. However, despite the fact that Wikipedia is a popular digital encyclopaedia the derived measures might only be relevant to a notion of knowledge somehow shared by Wikipedia users, that might not be appealing for users in some countries/regions or speakers of certain languages (Wikimedia, 2020a). This type of data might not be relevant to those who, for example, wish to measure divides relating to indigenous knowledge.

Timeliness- Wikipedia data are released in some cases in real-time (online data) with some log files published within days. Despite the fact that some analytical files are available on a monthly basis this is still much faster than the compilation of traditional data.

[29] A global survey about consumption patterns conducted in 14 languages found gender disparities in the consumption of Wikipedia: women are underrepresented and consume less knowledge (Johnson et al., 2020). Results from the Community Engagement Insights 2018 show that on the production side 90 percent of the Wikipedia editors are male, 81 percent come from high-income countries, and 34 percent have a university degree (Wikimedia, 2019a).

Accessibility- When compared to traditional data the level of accessibility of Wikipedia data is higher as it is possible to get open and free access to raw data. In addition, the community working with these data and the Wikimedia Foundation provide tools that can help with the extraction and initial analysis of these data. The openness and costlessness of these data ease the replicability process, fundamental in the production of research. In addition, Wikipedia log files can be considered as a stable source of data throughout time as the log files have been available in some cases for more than 10 years, and it is expected that these will be available in the future. Yet some accessibility issues were encountered. The analysis of raw data and the use of some tools require high-computational power and data science skills, and the lack of documentation also adds to the challenges. Another limitation is that due to privacy issues not all data can be identified at the country level, and it is not possible to identify what type of contents/knowledge is being consumed by countries. Finally, the geolocated consumption data have not been widely used and the API is still in "experimental" phase.

Table 5 summarises the results of the ACARTA model for this case.

Table 5: Strengths and limitations of Wikipedia big data

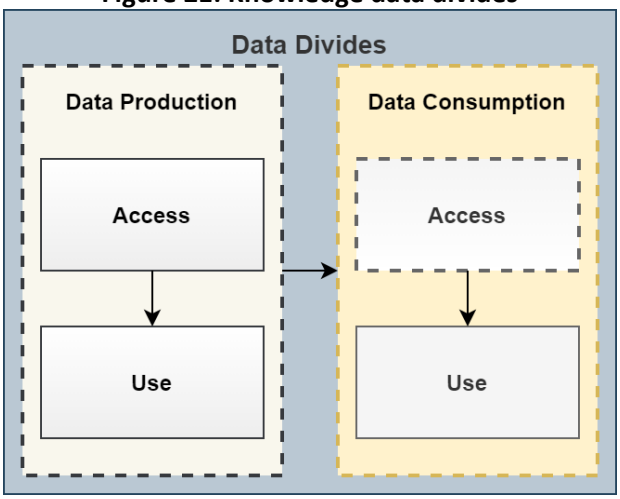
Characteristics	Strengths	Limitations
Availability	-All countries included -~250 languages including all major ones -Cheaper to produce	-Not all small languages represented
Completeness	-Provides additional dimensions	-Lack of traditional dimensions: age, gender, urban/rural
Accuracy	-Increased accuracy in the generation of data -Methodologically homogeneous	- Potential bias: knowledge available and users' profiles -Competitors, barriers
Relevance	-Relevant to consumption in those subdimensions measured	-Limited to a certain notion of knowledge (Western)
Timeliness	-Real-time/daily -Monthly analytical files	
Accessibility	-Raw data are open and free -Tools available to access and analyse data -Allows replicability -Stable source	-Some analyses require high-computational power and data science capabilities -No documentation -Some georeferenced data are limited -API is still in experimental phase

Source: Own based on Heeks (2006).

G3. Knowledge Data Divides

Figure 21 presents the elements of data divides on the production and consumption sides that are included in the conceptual framework, highlighting with a dashed line those areas that are evidenced by the analyses above.

Figure 21: Knowledge data divides



First, regional and income disparities emerged with the use of Wikipedia data that were helpful to identify big data-related divides on the consumption side. Overall, regional disparities presented the same income level patterns, with high-income countries consuming more knowledge than middle-income countries, and the latter consuming more than low-income countries. Yet, disparities between regions were very high even when comparing between groups with the same income level. These disparities were profound in the case of the MENA region where high- and middle-income countries had very low page views per connected capita values when compared to similar countries – this seemed mostly the result of low consumption of the Arabic Wikipedia. Wikipedia data was helpful to identify disparities in the consumption of knowledge and regional disparities. Yet, regional big data-related divides in the consumption of knowledge from Wikipedia were clearly identified.

At the access level on the consumption side, it is important to note that when individuals consume contents from the Wikipedia they are at the same time producing the exhaust data that is being used to measure these knowledge divides. In this case, the bidirectionality of this consumption-production relationship is not considered by the conceptual framework. Overall, knowledge divides in the consumption of knowledge are reflected as well in the production of the data that are used to measure these knowledge divides. As mentioned before there are big disparities in the consumption of contents following the multiple divides that were mentioned in this paper.

When contrasting production to consumption through the estimation of the consumption indicator and the joint Wikipedia index it was possible to identify specific big data-related divides. Once more, income and regional divides appeared together with the knowledge consumption and production divides. Overall, high-income countries from the Europe & Central Asia region got the highest scores on the production and consumption indices; low-income countries from Sub-Saharan Africa got the lowest scores. The joint production-consumption index confirmed a big data-related divide where high-income countries from Europe & Central Asia lead the production and consumption of knowledge. Also, it was interesting to see that only a few countries appeared to be producing more knowledge than they are consuming, in relative terms; the opposite is true for the majority of the countries.

Divides are broad with some countries leading the production and consumption of knowledge which might amplify existing divides.

H. Concluding Remarks

Knowledge of all kinds has a fundamental role in development and wellbeing including the particular type of knowledge studied in this case. That type of formal knowledge is intertwined with ICTs: ICTs have become a crucial element to produce and consume this type of knowledge, and that knowledge is also necessary for users to benefit from the use of ICTs. Disparities in the production and consumption of multiple dimensions of this type of knowledge have been measured between countries. These disparities have been referred as knowledge divides. Traditional data used to measure divides such as literacy rates, R&D indicators, scientific articles published, among others, have usually been defined by international organisations and present multiple limitations. A small number of big data sources have been used to measure knowledge divides, typically from the production side. This case focused on using big data from Wikipedia log files and explored the way that these big data can be used to measure knowledge divides.

Following the conceptual framework, multiple divides can be identified. These big data allowed the identification of consumption, per capita estimates, income, technological, and language divides related to knowledge that cannot be explored with traditional sources. The use of the Wikipedia API to extract billions of pageviews and the repurposing of these data to measure knowledge consumption and the aforementioned subdimensions becomes a valuable methodological contribution. The combination of Wikipedia big data with traditional data allowed additional insights into knowledge divides, and revealed the magnitude of these divides between high-income countries and the rest of the world. The use of consumption data also allowed the estimation of a consumption indicator that combined with an existing production indicator, offered a bidimensional approach to knowledge divides.

Using the conceptual framework the data qualities were analysed. Wikipedia big data was shown to be a useful source to measure knowledge divides, despite presenting multiple limitations. Overall, it was found that when compared to traditional datasets these data increase geographical availability and are expected to be cheaper to produce. These data are also more accurate in some ways, timely, and accessible. Yet, these data are less complete as these are not useful to measure sociodemographic dimensions. Also, it is important to keep in mind that these measures are reflecting the popularity and interest in consuming contents using the “wiki” model of representing knowledge in a Western manner. Moreover, biases in the production and consumption of contents, related to the profile of the creators of contents – mostly men from Western countries – and users, might impact these data and should be acknowledged.

Some implications for development and policy were identified. Considering the limitations of these data, and the lack of a sole definition of knowledge, this source should be used as a complement of other measures. These data become valuable when used alongside other sources that measure diverse dimensions of knowledge to integrate multidimensional indicators. Measures of the knowledge economy can benefit from the inclusion of the

consumption measures provided by these data and can be used by policy makers to track the state and progress of countries. Measures of language divides are a helpful element to understand how languages are being digitally represented and consumed. These measures can help policy makers to identify those countries and languages that are not being represented and whose speakers are at risk of not being able to access this type of knowledge. Endangered languages can benefit from the support of national and multilateral programmes that promote their digital representation in Wikipedia, and similar projects. In addition, data about consumption can be useful to track the impacts of policies related to the analogue foundations of digital divides such as education and literacy.

Acknowledgements

This paper is an excerpt of the thesis: Rivera Illingworth, L. (2022). *Digital divides in the era of Big Data: new dimensions and measurements*. University of Manchester.

References

Alexa (2019). *The Top 500 Sites on the Web*: Alexa, an amazon.com company. Available at: https://www.alexa.com/topsites (Accessed: 26-June-2019).

Alexa (2020). *Top Sites in South Korea*: Alexa, an amazon.com company. Available at: https://www.alexa.com/topsites/countries/KR (Accessed: 25-February-2020).

Alexa (2021). *The Top 500 Sites on the Web*: Alexa, an amazon.com company. Available at: https://www.alexa.com/topsites (Accessed: 2-October-2021).

BSI (2020). *BS ISO 639-3:2007*, London: The British Standards Institution.

Brown, D. M., Soto-Corominas, A., Suárez, J. L. & Rosa, J. d. l. (2017). 'The Social Media Data Processing Pipeline', in Quan-Haase, A. & Sloan, L. (eds.) The SAGE Handbook of Social Media Research Methods. 1st ed. London: SAGE Publishing, pp. 125-145.

Buchel, O. & Pennington, D. R. (2017). 'Geospatial Analysis', in Quan-Haase, A. & Sloan, L. (eds.) The SAGE Handbook of Social Media Research Methods. 1st ed. London: SAGE Publishing, pp. 285-308.

Cady, F. (2017). 'The Data Science Road Map', in Cady, F. (ed.) The Data Science Handbook. 1st ed. NJ, USA: John Wiley & Sons, Inc., pp. 7-17.

Castells, M. (2009). *The Rise of the Network Society* (New ed. Vol. 1). Oxford, UK: Wiley-Blackwell.

Chen, D. H. C. & Dahlmann, C. J. (2006). The Knowledge Economy, the KAM Methodology and World Bank Operations. p. 42. Available at: http://documents.worldbank.org/curated/en/695211468153873436/pdf/358670WBI0The11dge1Economy01PUBLIC1.pdf (Accessed: 10-June-2019).

Cornell University, INSEAD & WIPO (2015). *The Global Innovation Index 2015: Effective Innovation Policies for Development*, Fontainebleau, Ithaca, and Geneva: Cornell University, INSEAD, and WIPO. Available at: https://www.wipo.int/edocs/pubdocs/en/wipo_gii_2015.pdf.

Cornell University, INSEAD & WIPO (2018). *The Global Innovation Index 2018: Energizing the World with Innovation*, Ithaca, Fontainebleau, and Geneva: Cornell_University, INSEAD and WIPO. Available at: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_gii_2018.pdf.

Cornell University, INSEAD & WIPO (2019). *Global Innovation Index 2019. Creating Healthy Lives—The Future of Medical Innovation*, Ithaca, Fontainebleau, and Geneva: Cornell University, INSEAD, WIPO. Available at: https://www.globalinnovationindex.org/gii-2019-report.

Dutta, S. (2012). *The Global Innovation Index 2012: Stronger Innovation Linkages for Global Growth*, Fontainebleau, France: INSEAD-WIPO. Available at: https://www.wipo.int/edocs/pubdocs/en/economics/gii/gii_2012.pdf.

Dutta, S., Lanvin, B. & Wunsch-Vincent, S. (2019). *Interactive Database of the GII 2019 Indicators: 7.3.3 Wikipedia Yearly Edits*. Global Innovation Index 1. Available at: https://www.globalinnovationindex.org/analysis-indicator (Accessed: 19-January-2020).

Eberhard, D. M., Simons, G. F. & Fennig, C. D. (2020). Ethnologue Global Dataset. *Ethnologue: Languages of the World* [Online]. Available at: https://www.ethnologue.com/sites/default/files/Ethnologue-23-Global%20Dataset%20Doc.pdf (Accessed: 04-April-2020).

EBRD (2019). *Introducing the EBRD Knowledge Economy Index*, London: European Bank for Reconstruction and Development. Available at:

https://www.ebrd.com/documents/policy/download-the-ebrds-knowledge-economy-index.pdf.

EC (2019). *The Changing Nature of Work and Skills in the Digital Age*, Brussels, Belgium: European Commission. Available at: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC117505/executive_new_way_of_work_online.pdf.

Graham, M. (2011). 'Wiki Space: Palimpsests and the Politics of Exclusion', in Geert Lovink, N. T. (ed.) *Critical point of view: A Wikipedia reader.* 1st ed. Amsterdam, Netherlands: Institute of Network Cultures, pp. 269-282.

Graham, M. & Hogan, B. (2014). *Uneven Openness: Barriers to MENA Representation on Wikipedia*, Oxford: Oxford Internet Institute. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2430912.

Graham, M., Hogan, B., Straumann, R. K. & Medhat, A. (2014). 'Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty', *Annals of the Association of American Geographers,* 104(4), pp. 746-764.

Graham, M., Straumann, R. K. & Hogan, B. (2015). 'Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia', *Annals of the Association of American Geographers,* 105(6), pp. 1158-1178.

Heeks, R. (2006). *Implementing and Managing Egovernment an International Text* (1st ed.). London: SAGE Publishing.

Heeks, R. (2017). Decent Work and the Digital Gig Economy: A Developing Country Perspective on Employment Impacts and Standards in Online Outsourcing, Crowdwork, etc. *Development Informatics Working Paper* [Online], (7). Available at: http://hummedia.manchester.ac.uk/institutes/gdi/publications/workingpapers/di/di_wp71.pdf (Accessed: 15-October-2019).

Heeks, R. (2018). *Information and Communication Technology for Development (ICT4D)* (1st ed.). Abingdon, UK: Routledge.

Hegelich, S. (2017). 'R for Social Media Analysis', in Quan-Haase, A. & Sloan, L. (eds.) The SAGE Handbook of Social Media Research Methods. 1st ed. London: SAGE Publishing, pp. 486-498.

Hilbert, M. (2011). 'The End Justifies the Definition: The Manifold Outlooks on the Digital Divide and their Practical Usefulness for Policy-Making', *Telecommunications Policy,* 35(8), pp. 715-736.

IPA-WIPO (2016). *The Global Publishing Industry in 2016*, Geneva: International Publishers Association; World Intellectual Property Organisation. Available at: https://www.wipo.int/edocs/pubdocs/en/wipo_ipa_pilotsurvey_2016.pdf.

ISO (2013). *Codes for the Representation of Names of Languages, ISO 639.2*, Vienna, Austria: Library of Congress. Available at: https://www.loc.gov/standards/iso639-2/ascii_8bits.html.

ITU (2018a). *Measuring the Information Society Report 2018*, Geneva: International Telecommunication Union. Available at: https://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2018/MISR-2018-Vol-1-E.pdf.

ITU (2018b). ITU World Telecommunication/ICT Indicators Database 2018. Available at: https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx (Accessed: 20-June-2020).

ITU (2019). ITU World Telecommunication/ICT Indicators Database 2019. Available at: https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx (Accessed: 15-June-2020).

Johnson, I., Lemmerich, F., Sáez-Trumper, D., West, R., Strohmaier, M. & Zia, L. (2020). Global Gender Differences in Wikipedia Readership. *arXiv* [Online]. Available at: https://arxiv.org/abs/2007.10403 (Accessed: 12-August-2019).

Latour, B. (2007). 'A Textbook Case Revisited. Knowledge as a Mode of Existence', in Hackett, E. J., Lynch Michael, Wajcman, Judy, Amsterdamska, Olga (ed.) *The Handbook of Science and Technology Studies* 1st ed. Cambridge, USA: MIT Press, pp. 83-112.

Lundvall, B.-Å. (2016). *The Learning Economy and the Economics of Hope* (1st ed.). London: Anthem Press.

Maaroof, A. (2016). *Big Data and the 2030 Agenda for Sustainable Development*, Bangkok, Thailand: UNESCAP. Available at: http://www.unescap.org/sites/default/files/1_Big%20Data%202030%20Agenda_stock-taking%20report_25.01.16.pdf.

Machlup, F. (1980). *Knowledge : its creation, distribution, and economic significance. Volume 1, Knowledge and knowledge production*. Princeton, New Jersey: Princeton University Press.

Mayr, P. & Weller, K. (2017). 'Think before you Collect: Setting up a Data Collection Approach for Social Media Studies', in Quan-Haase, A. & Sloan, L. (eds.) The SAGE Handbook of Social Media Research Methods. 1st ed. London: SAGE Publishing, pp. 107-124.

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å. & Lanamäki, A. (2015). '"The Sum of All Human Knowledge": A Systematic Review of Scholarly Research on the Content of Wikipedia', *Journal of the Association for Information Science and Technology,* 66(2), pp. 219-245.

Mishra, D. (2015). Will the Spread of Digital Technologies Spell the End of the Knowledge Divide? *Prepared for the 2015 Brookings Blum Roundtable* [Online]**,** p. 6. Available at: https://www.brookings.edu/wp-content/uploads/2016/07/MishraEndoftheKnowledgeDivide.pdf (Accessed: 15-Mar-2019).

NumFOCUS (2018). pandas Python Data Analysis Library. v0.24.2 [Computer Program]. Available at: https://pandas.pydata.org/ (Accessed: 15-June-2019).

OECD (1996). *The Knowledge-Based Economy*, Paris: OECDOCDE/GD(96)102). Available at: https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=OCDE/GD%2896%29102&docLanguage=En.

OECD (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide* (1st ed.). Paris, France: OECD Publishing; Joint Research Centre-European Commission.

OECD (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development* (1st ed.). Paris, France: OECD Publishing.

OECD (2017). *DAC List of ODA Recipients*. Paris: OECD. Available at: http://www.oecd.org/dac/financing-sustainable-development/development-finance-standards/daclist.htm (Accessed: 01-Apr-2020).

OECD (2018). *DAC List of ODA Recipients*. Paris: OECD. Available at: http://www.oecd.org/dac/financing-sustainable-development/development-finance-standards/daclist.htm (Accessed: 01-Apr-2020).

OECD/Eurostat (2018). *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation*, Paris-Luxembourg: OECD Publishing, Eurostat. Available at: https://doi.org/10.1787/9789264304604-en.

Ojanperä, S., Graham, M. & Zook, M. (2019). 'The Digital Knowledge Economy Index: Mapping Content Production', *The Journal of Development Studies,* 55(12), pp. 2626-2643.

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å. & Lanamäki, A. (2012). The People's Encyclopedia under the Gaze of the Sages: A Systematic Review of Scholarly Sesearch on Wikipedia. *SSRN* [Online]**,** p. 138. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2021326 (Accessed: 15-June-2018).

Python Software Foundation, P. (2019). JSON Encoder and Decoder. v3.6 [Computer Program]. Available at: https://docs.python.org/3/library/json.html (Accessed: 13-June-2019).

Reitz, K. (2019). Requests: HTTP for Humans™. v2.22.0. [Computer Program]. Available at: https://2.python-requests.org/en/master/ (Accessed: 13-June-2019).

Russell, B. (2009). *Human Knowledge its Scope and Limits* (New ed.). Abingdon, UK: Routledge.

Scheff, S. W. (2016). 'Chapter 8 - Nonparametric Statistics', in Scheff, S. W. (ed.) *Fundamental Statistical Principles for the Neurobiologist.* 1st ed. London: Academic Press, pp. 157-182.

Segev, E. (2010). *Google and the Digital Divide: The Bias of Online Knowledge* (1st ed.). Oxford, UK: Elsevier.

SimilarWeb. (n.d.). Top Websites Ranking. Retrieved from: https://www.similarweb.com/es/top-websites/

Statcounter (2020). *Desktop vs Mobile Market Share Worldwide*. Available at: https://gs.statcounter.com/platform-market-share/desktop-mobile/worldwide/ (Accessed: 25-February-2020).

Sullivan, J. & Escaravage, S. (2015). The Field Guide to Data Science [Online]. Virginia, USA: Booz Allen Hamilton. Available at: https://www.boozallen.com/content/dam/boozallen_site/sig/pdf/publications/2015-field-guide-to-data-science.pdf.

Tkacz, G. L. a. N. (2011). *Critical Point of View: A Wikipedia Reader* (1st ed.). Amsterdam, Netherlands: Institute of Network Cultures.

UNCTAD (2019). *Digital Economy Report 2019. Value Creation and Capture: Implication for Developing Countries*, New York, NY: United Nations Conference on Trade and Development. Available at: https://unctad.org/en/PublicationsLibrary/der2019_en.pdf.

UNECE (2014). A Suggested Framework for the Quality of Big Data. Big Data Quality Framework v4.01 [Online]. Available at: https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2 (Accessed: 11-September-2018).

UNESCO-UIS (2003). *Measuring and Monitoring The Information and Knowledge Societies: A Statistical Challenge*, Montreal, Canada: UNESCO Institute for Statistics, Montreal. Available at: http://www.unesco.org/new/en/communication-and-information/resources/publications-and-communication-materials/publications/full-

list/measuring-and-monitoring-the-information-and-knowledge-societies-a-statistical-challenge/.

UNESCO (2010). Presentation of the World Digital Library Global Initiative. *The World Digital Library and Universal Access to Knowledge* [Online], (20-May-2019), p. 7. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000187073 (Accessed: 5-May-2019).

UNESCO/IFAP (2016). *Knowledge Societies Policy Handbook* [Online]. Paris: United Nations Educational, Scientific and Cultural Organization-Information for All Programme. Available at: https://en.unesco.org/sites/default/files/knowledge_socities_policy_handbook.pdf.

UNESCWA (2005). *Information Society Indicators*, New York, NY: United Nations Economic and Social Commission for Western Asia. Available at: https://www.unescwa.org/sites/www.unescwa.org/files/publications/files/ictd-05-1.pdf.

Unger, R. (2019). *The Knowledge Economy* (1st ed.). London: Verso Books.

UNGP (2016). A Guide to Data Innovation for Development - From Idea to Proof-of-Concept. Available at: https://www.undp.org/content/undp/en/home/librarypage/development-impact/a-guide-to-data-innovation-for-development---from-idea-to-proof-.html (Accessed: 12-April-2017).

WB (2007). *Building Knowledge Economies: Advanced Strategies for Development* (1st ed.). Washington, DC, USA: World Bank Publications.

WB (2012). *Knowledge Economy Index (World Bank), 2012*. Knowledge, World Rankings: World Bank. Available at: https://knoema.com/WBKEI2013/knowledge-economy-index-world-bank-2012 (Accessed: 05-May-2017).

WB (2016). *World Development Report 2016: Digital Dividends*, Washington, D.C.: World Bank. Available at: http://www.worldbank.org/en/publication/wdr2016.

WB (2019a). *Education Indicators*: World Bank. Available at: https://data.worldbank.org/topic/education (Accessed: 01-March-2019).

WB (2019b). *World Development Indicators: Literacy Rate, Adult Total (% of People Ages 15 and Above)*. Washington D.C.: World Bank. Available at: https://databank.worldbank.org/ (Accessed: 15-April-2020).

WB (2019c). *World Development Report 2019. The Changing Nature of Work*, Washington, D.C.: World Bank. Available at: http://documents.worldbank.org/curated/en/816281518818814423/pdf/2019-WDR-Report.pdf.

WB (2020a). *Individuals using the Internet (% of Population)*: World Bank. Available at: https://data.worldbank.org/indicator/IT.NET.USER.ZS (Accessed: 01-February-2020).

WB (2020b). *World Bank Country and Lending Groups*. Washington, DC: World Bank. Available at: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (Accessed: 05-May-2020).

Wikimedia (2019a). *Community Engagement Insights/2018 Report*. Available at: https://meta.wikimedia.org/wiki/Community_Engagement_Insights/2018_Report (Accessed: 01-May-2019).

Wikimedia (2019b). *Wikipedia: Size of Wikipedia*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia (Accessed: 10-December-2019).

Wikimedia (2019c). *Wikipedia: Censorship_of_Wikipedia*. Available at: https://en.wikipedia.org/wiki/Censorship_of_Wikipedia (Accessed: 10-December-2019).

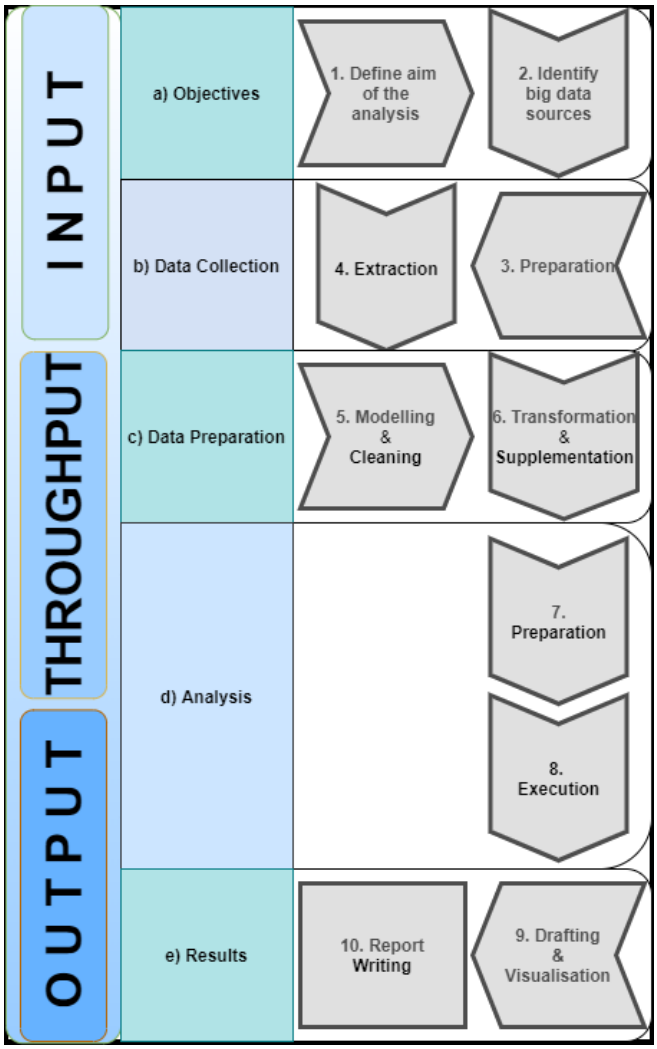
Wikimedia (2019d). *List of Wikipedias, Meta-Wiki*. Available at: https://meta.wikimedia.org/wiki/List_of_Wikipedias (Accessed: 01-May-2019).
Wikimedia (2019e). *Wikipedia: Wikipedia is an Encyclopedia*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_an_encyclopedia (Accessed: 7-May-2019).
Wikimedia (2019f). *Wikimedia REST API 1.0.0 OAS3*: Wikimedia Foundation. Available at: https://wikimedia.org/api/rest_v1/ (Accessed: 01-June-2019).
Wikimedia (2020a). *Systemic Bias*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Systemic_bias (Accessed: 25-February-2020).
Wikimedia (2020b). *List of ISO 639-2 Codes*. Available at: https://en.wikipedia.org/wiki/List_of_ISO_639-2_codes.
Wikimedia (2020c). *Armenian Wikipedia*. Available at: https://en.wikipedia.org/wiki/Armenian_Wikipedia (Accessed: 24-February-2020).
Wikimedia (2020d). *Arminé Aghayan*. Available at: https://wikimediafoundation.org/profile/armine-aghayan/ (Accessed: 24-February-2020).
Zeller, F. (2017). 'Analyzing Social Media Data and Other Data Sources: A Methodological Overview', in Quan-Haase, A. & Sloan, L. (eds.) The SAGE Handbook of Social Media Research Methods. 1st ed. London: SAGE Publishing, pp. 386-404.

Appendices

1. Data Collection and Preparation

Based on the social data science methodology presented below this subsection describes how the data collection and preparation processes were conducted. With the aim of guiding researchers throughout their data tasks, a social data science methodology (SDSM) was defined drawing from multiple sources.

Figure A.1 Social data science methodology



Source: adapted from (UNECE, 2014; Sullivan and Escaravage, 2015; UNGP, 2016; Brown et al., 2017; Buchel and Pennington, 2017; Cady, 2017; Hegelich, 2017; Mayr and Weller, 2017; Zeller, 2017)

i) A GET call within the Wikimedia REST API (representational state transfer application program interface) was identified to extract page views per country for each Wikipedia language (see more below).

ii) A Python client was programmed using the Requests (NumFOCUS, 2018), JSON (JavaScript Object Notation) (Python Software Foundation, 2019), and pandas (NumFOCUS, 2018) libraries in order to handle and extract all the URLs (details below).

The Wikimedia REST API is a tool to process and extract data from the Wikimedia servers (Wikimedia, 2019f). The REST API allowed us to conduct simple web-based requests using GET commands through HTTP based URLs. In this way the data processing was performed on the Wikimedia web servers and the responses obtained were the results.

These API requests were considered to be in "experimental" mode which means that there might be changes in the way the request operates without informing the users, until a stable version is released. Also, the server had rate limits defined to a maximum of 100 requests per second and a maximum timespan of one month, which translates in the creation of a URL per article, per Wikipedia language, per month-year.

As mentioned in ii) a Python client was programmed to handle the requests and extract the URLs. This Python client used the Requests, JSON, and pandas library (NumFOCUS, 2018; Python Software Foundation, 2019; Reitz, 2019). The first client was used to connect to the API and get the data; the second one to read the JSON files obtained from the API; and the third one to manage the data within a data frame. The code is included in Appendix 2. Using the Python client all the URLs for each request were loaded in a csv file and the data extraction executed.

Consumption: page views per country and language

a) Data Collection
The page views count aggregated at the country level were analysed for all the Wikipedia projects/languages with more than 1,000 articles – a total of 247 languages (May 2019) (Wikimedia, 2019d). In this case data were collected between January 2016 and December 2018. The following tasks were conducted to prepare for the data extraction step:

i) The languages and their codes (ISO 639) were identified for the 247 projects with more than 1,000 articles available in January 2019.

ii) The links to the 247 Wikipedia projects were extracted.

iii) The GET call within the Wikimedia REST API was identified to extract the page view counts for each language aggregated by country.

iv) 9,108 URLs were created using the logic established by the API (see below).

v) A Python client was used to handle and extract all the URLs (details below).

One GET call was used on the consumption side:

- To identify page views the "Pageviews data" call was used to "Get pageviews by country and access method". The URL was created in the following manner (parameters in bold are fixed):

  https://wikimedia.org/api/rest_v1/metrics/pageviews/top-by-country/*project*/*access*/*year*/*month*
  Project: Wikipedia project language
  Access: all-access, desktop, mobile-web[30]
  Year: year to query
  Month: month to query

The API connected to the analytical data files derived from the raw log files that included only human traffic. A Python client was used to handle the requests and extract the URLs. The code is included in Appendix 2. Once more, this client managed all the URLs previously loaded in a csv file to execute the extraction.

b) Data Preparation
The data preparation stage was conducted also using Python. In this case once the response from the API was received as a JSON file it was integrated into a data frame using the JSON and pandas libraries. The three files (one for each access type: all-access, desktop, and mobile-web) contain the following variables:

country: the ISO 3166 alpha-2 code
rank: for each project/language the rank of countries (1= most views)
views: estimated views interval
views_ceil: estimated views rounded to the nearest 1,000 for privacy reasons[31]
project: Wikipedia project language
year: year queried
month: month queried

[30] The mobile-web access was discarded due to the low numbers.
[31] Those countries with less than 100 views within a period are not exported by the API for privacy reasons.

2. Python Code, Page Views

```
# -*- coding: utf-8 -*-
"""
Created on Tue May 7 19:38:21 2019

Spyder Editor, Extraction of Pageviews Wikimedia REST API

"""

import requests
import json
import pandas as pd

# Read URLs from Csv
df_urls = pd.read_csv('file with urls.csv')

print(df_urls.head(5))

#creates a new dataframe that's empty
final_df =  pd.DataFrame()

headers = {
   'User-Agent': 'Luis RI',
   'From': ‘xxxx@xxx.com'  # declare email
}

#iterates on dataframe
for row in df_urls.itertuples():

  print(row.url_name)

  url = row.url_name

  r = requests.get(url)

  #print (r.status_code)
  #print (r.headers['content-type'])

  data = r.json()

  #print(json.dumps(data, indent=4))

  # data is a dictionary
  country_info = data['items'][0]['countries']
  df = pd.DataFrame.from_dict(country_info)

  #shape of the results
```

```
  print (df.shape)
  print (df.head())


  df['project'] = data['items'][0]['project']
  df['access'] = data['items'][0]['access']
  df['year'] = data['items'][0]['year']
  df['month'] = data['items'][0]['month']

  final_df = final_df.append(df, ignore_index = True)

print (final_df.shape)

#print (final_df.sample(5))

final_df.to_csv('out_filename.csv', index=False)

print(final_df.describe())
```

3. Summary Statistics

Page views by year (views); page view consumption per capita for connected individuals by year (wikipc); proportion of views using mobile devices by year (%mob).

Table A.3.1 Summary statistics, all countries and income level

All countries	views16	views17	views18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
N	242	245	248	202	205	205	238	243	246
mean	781,000,000	755,000,000	744,000,000	4.7	4.4	4.4	49.0	53.7	57.6
median	42,100,000	41,900,000	37,400,000	3.4	3.3	3.0	48.7	53.7	58.4
min	4,000	3,000	4,000	0.1	0.1	0.1	4.6	6.1	10.1
max	41,900,000,000	41,600,000,000	41,400,000,000	22.3	20.3	23.4	90.9	92.4	89.4

High-Income	views16	views17	views18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
N	73	73	76	64	66	69	73	73	76
mean	1,820,000,000	1,780,000,000	1,720,000,000	9.4	8.9	8.7	43.6	48.2	52.7
median	176,000,000	160,000,000	164,000,000	9.8	9.3	9.0	43.0	47.7	51.8
min	1,523,000	1,267,000	681,000	1.7	1.6	1.6	25.4	28.0	25.8
max	41,900,000,000	41,600,000,000	41,400,000,000	22.3	20.3	23.4	66.4	73.0	77.6

Middle-Income	views16	views17	views18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
N	97	97	96	89	90	89	96	97	96
mean	556,000,000	545,000,000	540,000,000	3.3	3.0	2.8	47.9	53.1	58.2
median	106,000,000	109,000,000	100,000,000	3.1	2.7	2.3	49.7	56.1	61.6
min	4,000	25,000	53,000	0.1	0.1	0.1	5.8	9.9	18.9
max	8,550,000,000	7,660,000,000	8,520,000,000	8.6	8.4	10.2	82.6	82.9	82.8

low-income countries	views16	views17	views18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
N	51	51	49	49	49	47	51	51	49
mean	37,500,000	39,100,000	31,900,000	1.1	1.0	0.9	62.2	66.1	67.5
median	11,900,000	12,800,000	12,900,000	0.8	0.7	0.7	67.6	70.7	70.8
min	20,000	22,000	32,000	0.2	0.2	0.2	4.6	6.1	10.1
max	311,000,000	333,000,000	346,000,000	5.9	5.7	4.2	90.9	92.4	89.4

Table A.3.2 Wikipedia summaries by country

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
AW	Aruba	ABW	Latin America & Caribbean	High-Income	High-Income	6,978,000	6,738,000	6,563,000	5.9	5.5	5.3	0.6	0.6	0.6
AF	Afghanistan	AFG	South Asia	Low-Income	Low-Income	241,510,000	279,768,000	175,126,000	5.5	5.7	3.4	0	0.1	0.1
AO	Angola	AGO	Sub-Saharan Africa	Low-Income	Low-Income	265,650,000	254,354,000	148,865,000	5.9	5	2.8	0.8	0.8	0.8
AI	Anguilla	AIA	N/A	N/A	N/A	604,000	710,000	576,000				0.5	0.5	0.6
AX	Åland	ALA	N/A	N/A	N/A	3,789,000	3,543,000	3,458,000				0.5	0.5	0.5
AL	Albania	ALB	Europe & Central Asia	Middle-Income	Middle-Income	109,709,000	120,403,000	129,129,000	4.8	4.9	5.2	0.5	0.6	0.6
AD	Andorra	AND	Europe & Central Asia	High-Income	High-Income	6,048,000	6,310,000	6,565,000	6.7	6.9	7.2	0.5	0.5	0.6
AE	United Arab Emirates	ARE	Middle East & North Africa	High-Income	High-Income	440,139,000	441,845,000	468,311,000	4.4	4.1	4.1	0.5	0.5	0.6
AR	Argentina	ARG	Latin America & Caribbean	Middle-Income	Middle-Income	1,824,553,000	1,733,498,000	1,727,884,000	4.9	4.3	4.3	0.5	0.6	0.6

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
AM	Armenia	ARM	Europe & Central Asia	Middle-Income	Middle-Income	100,635,000	105,914,000	106,862,000	4.5	4.3	4.3	0.4	0.4	0.5
AS	American Samoa	ASM	East Asia & Pacific	Middle-Income	Middle-Income	1,043,000	974,000	729,000				0.4	0.5	0.5
AQ	All land and ice shelves south of the 60th parallel south	ATA	N/A	N/A	N/A			20,000						0.2
TF	The French Southern and Antarctic Lands	ATF	N/A	N/A	N/A			5,000						0.3
AG	Antigua and Barbuda	ATG	Latin America & Caribbean	Middle-Income	Middle-Income	4,925,000	5,520,000	4,776,000	5.6	5.9	5.4	0.5	0.7	0.6
AU	Australia	AUS	East Asia & Pacific	High-Income	High-Income	2,900,981,000	2,835,553,000	2,777,166,000	11.5	11.1	10.7	0.5	0.5	0.5
AT	Austria	AUT	Europe & Central Asia	High-Income	High-Income	1,236,004,000	1,202,591,000	1,194,595,000	14	13	12.8	0.4	0.5	0.5
AZ	Azerbaijan	AZE	Europe & Central Asia	Middle-Income	Middle-Income	186,589,000	194,599,000	202,930,000	2	2.1	2.1	0.5	0.6	0.6
BI	Burundi	BDI	Sub-Saharan Africa	Low-Income	Low-Income	5,918,000	6,131,000	4,965,000	0.9	0.8	0.7	0.8	0.8	0.7
BE	Belgium	BEL	Europe & Central Asia	High-Income	High-Income	1,171,756,000	1,126,307,000	1,124,070,000	10	9.4	9.3	0.3	0.4	0.4
BJ	Benin	BEN	Sub-Saharan Africa	Low-Income	Low-Income	15,504,000	15,269,000	17,384,000	1	0.8	0.9	0.6	0.6	0.7
BQ	Bonaire, Sint Eustatius and Saba	BES	N/A	N/A	N/A	1,457,000	1,383,000	1,410,000				0.4	0.5	0.6
BF	Burkina Faso	BFA	Sub-Saharan Africa	Low-Income	Low-Income	12,292,000	15,352,000	14,996,000	0.4	0.4	0.4	0.7	0.8	0.8
BD	Bangladesh	BGD	South Asia	Low-Income	Low-Income	310,765,000	333,295,000	346,375,000	0.9	0.9	1	0.6	0.6	0.7
BG	Bulgaria	BGR	Europe & Central Asia	Middle-Income	Middle-Income	425,788,000	453,544,000	557,127,000	8.3	8.4	10.2	0.3	0.4	0.4
BH	Bahrain	BHR	Middle East & North Africa	High-Income	High-Income	60,860,000	56,319,000	54,035,000	3.7	3.3	2.9	0.5	0.6	0.7

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
BS	Bahamas, The	BHS	Latin America & Caribbean	High-Income	High-Income	16,345,000	19,120,000	14,833,000	4.4	4.7	3.8	0.5	0.7	0.7
BA	Bosnia and Herzegovina	BIH	Europe & Central Asia	Middle-Income	Middle-Income	150,827,000	156,312,000	158,161,000	5.9	5.3	5.7	0.4	0.5	0.6
BL	The Collectivity of Saint-Barthélemy	BLM	N/A	N/A	N/A	872,000	666,000	657,000				0.5	0.6	0.6
BY	Belarus	BLR	Europe & Central Asia	Middle-Income	Middle-Income	464,570,000	449,831,000	430,340,000	5.7	5.3	4.8	0.4	0.4	0.5
BZ	Belize	BLZ	Latin America & Caribbean	Middle-Income	Middle-Income	6,813,000	6,986,000	7,009,000	3.5	3.3	3.2	0.5	0.5	0.6
BM	Bermuda	BMU	North America	High-Income	High-Income	5,683,000	5,216,000	5,327,000	7.4	6.8	7.1	0.4	0.5	0.5
BO	Bolivia	BOL	Latin America & Caribbean	Middle-Income	Middle-Income	186,185,000	204,142,000	202,076,000	3.6	3.5	3.4	0.6	0.7	0.7
BR	Brazil	BRA	Latin America & Caribbean	Middle-Income	Middle-Income	3,793,499,000	3,636,499,000	3,822,213,000	2.5	2.1	2.3	0.4	0.5	0.6
BB	Barbados	BRB	Latin America & Caribbean	High-Income	High-Income	16,272,000	14,550,000	13,021,000	6	5.2	4.6	0.5	0.5	0.6
BN	Brunei Darussalam	BRN	East Asia & Pacific	High-Income	High-Income	25,489,000	25,178,000	25,443,000	5.6	5.2	5.2	0.6	0.6	0.6
BT	Bhutan	BTN	South Asia	Low-Income	Low-Income	5,200,000	5,800,000	6,434,000	1.3	1.2	1.5	0.7	0.8	0.8
BV	Bouvet Island	BVT	N/A	N/A	N/A			64,000						0.5
BW	Botswana	BWA	Sub-Saharan Africa	Middle-Income	Middle-Income	13,256,000	11,937,000	12,437,000	1.2	1	1.1	0.5	0.5	0.5
CF	Central African Republic	CAF	Sub-Saharan Africa	Low-Income	Low-Income	513,000	847,000	719,000	0.2	0.3	0.3	0.4	0.7	0.7
CA	Canada	CAN	North America	High-Income	High-Income	5,240,560,000	5,016,764,000	4,866,189,000	13.2	12.3	11.8	0.4	0.4	0.5
CC	The Territory of Cocos (Keeling) Islands	CCK	N/A	N/A	N/A	12,000	12,000	28,000				0.3	0.5	0.4

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
CH	Switzerland	CHE	Europe & Central Asia	High-Income	High-Income	1,272,813,000	1,251,952,000	1,240,699,000	14.2	13.2	13	0.4	0.4	0.5
CL	Chile	CHL	Latin America & Caribbean	Middle-Income	High-Income	833,792,000	828,604,000	817,361,000	4.7	4.6	4.4	0.5	0.5	0.6
CN	China	CHN	East Asia & Pacific	Middle-Income	Middle-Income	1,120,720,000	1,243,946,000	1,766,826,000	0.1	0.1	0.2	0.2	0.2	0.3
CI	Côte d'Ivoire	CIV	Sub-Saharan Africa	Middle-Income	Middle-Income	53,665,000	64,389,000	57,811,000	0.5	0.5	0.4	0.5	0.6	0.7
CM	Cameroon	CMR	Sub-Saharan Africa	Middle-Income	Middle-Income	48,632,000	49,890,000	49,577,000	0.7	0.7	0.7	0.7	0.7	0.7
CD	Congo, Dem. Rep.	COD	Sub-Saharan Africa	Low-Income	Low-Income	39,299,000	46,978,000	45,907,000	0.7	0.6	0.5	0.8	0.9	0.9
CG	Congo, Rep.	COG	Sub-Saharan Africa	Middle-Income	Middle-Income	6,006,000	5,269,000	5,123,000	1.2	1	0.9	0.7	0.7	0.7
CK	The Cook Islands	COK	N/A	Middle-Income	Middle-Income	402,000	458,000	514,000				0.4	0.6	0.7
CO	Colombia	COL	Latin America & Caribbean	Middle-Income	Middle-Income	1,515,193,000	1,397,116,000	1,306,235,000	4.5	3.8	3.5	0.5	0.5	0.6
KM	Comoros	COM	Sub-Saharan Africa	Low-Income	Low-Income	613,000	826,000	1,087,000	0.8	1	1.3	0.5	0.6	0.7
CV	Cabo Verde	CPV	Sub-Saharan Africa	Middle-Income	Middle-Income	5,474,000	5,918,000	6,342,000	1.7	1.6	1.7	0.6	0.7	0.7
CR	Costa Rica	CRI	Latin America & Caribbean	Middle-Income	Middle-Income	177,818,000	173,532,000	166,590,000	4.6	4.1	3.7	0.6	0.6	0.7
CU	Cuba	CUB	Latin America & Caribbean	Middle-Income	Middle-Income	19,455,000	18,877,000	18,109,000	0.3	0.3	0.3	0.1	0.1	0.2
CW	Curaçao	CUW	Latin America & Caribbean	High-Income	High-Income	8,570,000	8,098,000	7,754,000	7.2	6.8	5.9	0.5	0.5	0.5
CX	The Territory of Christmas Island	CXR	N/A	N/A	N/A	42,000	72,000	52,000				0.4	0.5	0.6
KY	Cayman Islands	CYM	Latin America & Caribbean	High-Income	High-Income	5,850,000	5,517,000	5,239,000	10.2	9.2	8.4	0.5	0.6	0.6

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
CY	Cyprus	CYP	Europe & Central Asia	High-Income	High-Income	59,485,000	60,862,000	69,758,000	5.6	5.3	5.8	0.4	0.5	0.5
CZ	Czech Republic	CZE	Europe & Central Asia	High-Income	High-Income	982,497,000	991,518,000	977,898,000	10.1	9.9	9.5	0.3	0.3	0.4
DE	Germany	DEU	Europe & Central Asia	High-Income	High-Income	12,916,885,000	12,380,247,000	11,977,361,000	15.5	14.8	13.4	0.4	0.4	0.4
DJ	Djibouti	DJI	Middle East & North Africa	Low-Income	Low-Income	4,353,000	5,166,000	5,805,000	2.9	0.8	0.9	0.4	0.5	0.6
DM	Dominica	DMA	Latin America & Caribbean	Middle-Income	Middle-Income	3,131,000	2,372,000	1,269,000	5.3	3.8	2.1	0.4	0.5	0.6
DK	Denmark	DNK	Europe & Central Asia	High-Income	High-Income	712,528,000	691,154,000	684,277,000	10.7	10.3	10.1	0.4	0.4	0.5
DO	Dominican Republic	DOM	Latin America & Caribbean	Middle-Income	Middle-Income	289,457,000	291,167,000	282,457,000	3.5	3.5	3	0.6	0.6	0.7
DZ	Algeria	DZA	Middle East & North Africa	Middle-Income	Middle-Income	409,584,000	404,574,000	382,924,000	2	1.7	1.3	0.5	0.5	0.6
EC	Ecuador	ECU	Latin America & Caribbean	Middle-Income	Middle-Income	414,442,000	401,088,000	383,863,000	3.9	3.5	3.3	0.4	0.4	0.5
EG	Egypt, Arab Rep.	EGY	Middle East & North Africa	Middle-Income	Middle-Income	455,189,000	495,362,000	501,670,000	1	0.9	0.9	0.5	0.6	0.6
ER	Eritrea	ERI	Sub-Saharan Africa	Low-Income	Low-Income	281,000	293,000	253,000				0.2	0.3	0.3
EH	The Sahrawi Arab Democratic Republic	ESH	N/A	N/A	N/A		19,000	45,000					0.6	0.7
ES	Spain	ESP	Europe & Central Asia	High-Income	High-Income	3,601,431,000	3,587,428,000	3,672,893,000	8	7.6	7.6	0.5	0.5	0.6
EE	Estonia	EST	Europe & Central Asia	High-Income	High-Income	177,749,000	198,665,000	208,487,000	12.9	14.3	14.7	0.3	0.3	0.3
ET	Ethiopia	ETH	Sub-Saharan Africa	Low-Income	Low-Income	57,963,000	58,271,000	50,101,000	0.3	0.2	0.2	0.7	0.7	0.6
FI	Finland	FIN	Europe & Central Asia	High-Income	High-Income	1,112,296,000	1,083,872,000	1,064,328,000	19.2	18.7	18.1	0.5	0.5	0.5

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
FJ	Fiji	FJI	East Asia & Pacific	Middle-Income	Middle-Income	12,559,000	12,514,000	12,371,000	2.5	2.3	2.3	0.5	0.6	0.6
FK	The Falkland Islands	FLK	N/A	N/A	N/A	222,000	45,000	86,000				0.4	0.4	0.6
FR	France	FRA	Europe & Central Asia	High-Income	High-Income	8,355,289,000	7,402,638,000	7,211,132,000	13.1	11.4	10.9	0.3	0.4	0.5
FO	Faroe Islands	FRO	Europe & Central Asia	High-Income	High-Income	5,625,000	5,658,000	5,503,000	10	9.8	9.7	0.4	0.5	0.5
FM	Micronesia, Fed. Sts.	FSM	East Asia & Pacific	Middle-Income	Middle-Income	602,000	627,000	540,000	1.4	1.4	1.1	0.3	0.4	0.4
GA	Gabon	GAB	Sub-Saharan Africa	Middle-Income	Middle-Income	12,301,000	12,539,000	11,126,000	1.1	1	0.9	0.7	0.7	0.7
GB	United Kingdom	GBR	Europe & Central Asia	High-Income	High-Income	9,752,545,000	9,592,976,000	9,643,775,000	13.1	12.8	12.7	0.5	0.5	0.6
GE	Georgia	GEO	Europe & Central Asia	Middle-Income	Middle-Income	142,964,000	151,096,000	153,295,000	5.4	5.6	5.4	0.3	0.4	0.5
GG	The Bailiwick of Guernsey	GGY	N/A	N/A	N/A	7,221,000	7,788,000	7,200,000				0.5	0.5	0.6
GH	Ghana	GHA	Sub-Saharan Africa	Middle-Income	Middle-Income	100,603,000	97,334,000	88,147,000	0.9	0.7	0.7	0.8	0.8	0.7
GI	Gibraltar	GIB	Europe & Central Asia	High-Income	High-Income	4,735,000	4,099,000	4,219,000	12.1	10.5	11	0.4	0.5	0.5
GN	Guinea	GIN	Sub-Saharan Africa	Low-Income	Low-Income	12,454,000	16,102,000	14,750,000	0.9	0.9	0.9	0.8	0.9	0.9
GP	Guadeloupe	GLP	N/A	N/A	N/A	19,549,000	16,493,000	18,460,000				0.4	0.5	0.6
GM	Gambia, The	GMB	Sub-Saharan Africa	Low-Income	Low-Income	2,962,000	3,056,000	3,214,000	0.7	0.6	0.6	0.7	0.7	0.7
GW	Guinea-Bissau	GNB	Sub-Saharan Africa	Low-Income	Low-Income	963,000	1,158,000	1,022,000	1.2	1.3	1.2	0.6	0.7	0.7
GQ	Equatorial Guinea	GNQ	Sub-Saharan Africa	Middle-Income	Middle-Income	1,583,000	1,988,000	1,631,000	0.5	0.5	0.4	0.6	0.5	0.5
GR	Greece	GRC	Europe & Central Asia	High-Income	High-Income	662,806,000	678,151,000	696,905,000	7.4	7.5	7.4	0.4	0.4	0.5

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
GD	Grenada	GRD	Latin America & Caribbean	Middle-Income	Middle-Income	3,410,000	3,483,000	3,368,000	4.7	4.6	4.3	0.4	0.5	0.5
GL	Greenland	GRL	Europe & Central Asia	High-Income	High-Income	2,083,000	1,999,000	2,009,000	4.5	4.3	4.3	0.4	0.5	0.5
GT	Guatemala	GTM	Latin America & Caribbean	Middle-Income	Middle-Income	222,793,000	218,869,000	212,036,000	3.2	2.6	2.5	0.5	0.5	0.6
GF	Guyane	GUF	N/A	N/A	N/A	7,913,000	7,387,000	7,154,000				0.4	0.5	0.5
GU	Guam	GUM	East Asia & Pacific	High-Income	High-Income	10,318,000	10,448,000	10,260,000	6.9	6.6	6.4	0.6	0.6	0.6
GY	Guyana	GUY	Latin America & Caribbean	Middle-Income	Middle-Income	10,273,000	11,385,000	11,416,000	3.1	3.3	3.3	0.5	0.6	0.6
HK	Hong Kong SAR, China	HKG	East Asia & Pacific	High-Income	High-Income	1,159,915,000	1,222,269,000	1,246,004,000	15.1	15.4	15.6	0.4	0.4	0.5
HM	The Territory of Heard Island and McDonald Islands	HMD	N/A	N/A	N/A			8,000						
HN	Honduras	HND	Latin America & Caribbean	Middle-Income	Middle-Income	106,172,000	109,430,000	107,808,000	3.2	3.1	2.9	0.6	0.6	0.6
HR	Croatia	HRV	Europe & Central Asia	High-Income	High-Income	359,844,000	362,322,000	373,793,000	9.9	10.9	10.5	0.4	0.5	0.5
HT	Haiti	HTI	Latin America & Caribbean	Low-Income	Low-Income	29,401,000	34,604,000	34,085,000	1.8	2.1	0.8	0.7	0.8	0.8
HU	Hungary	HUN	Europe & Central Asia	High-Income	High-Income	678,405,000	685,979,000	723,159,000	7.3	7.6	8.1	0.4	0.4	0.5
ID	Indonesia	IDN	East Asia & Pacific	Middle-Income	Middle-Income	1,872,949,000	1,925,341,000	2,066,912,000	2.3	1.9	1.6	0.6	0.7	0.8
IM	Isle of Man	IMN	Europe & Central Asia	High-Income	High-Income	10,985,000	10,996,000	11,313,000				0.5	0.5	0.6
IN	India	IND	South Asia	Middle-Income	Middle-Income	6,616,134,000	7,222,573,000	8,517,538,000	1.4	1.3	1.5	0.6	0.7	0.8
IO	The British Indian Ocean Territory	IOT	N/A	N/A	N/A	108,000	105,000	98,000				0.4	0.4	0.7

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
IE	Ireland	IRL	Europe & Central Asia	High-Income	High-Income	1,074,698,000	992,442,000	859,178,000	22.3	20.3	17.5	0.6	0.6	0.6
IR	Iran, Islamic Rep.	IRN	Middle East & North Africa	Middle-Income	Middle-Income	4,399,401,000	4,354,904,000	3,462,414,000	8.6	7.4	5	0.1	0.2	0.3
IQ	Iraq	IRQ	Middle East & North Africa	Middle-Income	Middle-Income	158,566,000	172,404,000	170,103,000	1.7	0.8	0.7	0.6	0.7	0.8
IS	Iceland	ISL	Europe & Central Asia	High-Income	High-Income	53,329,000	55,753,000	98,123,000	13.5	13.8	23.4	0.3	0.4	0.3
IL	Israel	ISR	Middle East & North Africa	High-Income	High-Income	943,734,000	960,271,000	1,019,940,000	11.6	11.3	11.7	0.4	0.5	0.6
IT	Italy	ITA	Europe & Central Asia	High-Income	High-Income	6,707,436,000	6,479,383,000	6,667,815,000	15	14.6	12.4	0.6	0.6	0.7
JM	Jamaica	JAM	Latin America & Caribbean	Middle-Income	Middle-Income	51,715,000	56,691,000	48,227,000	3.4	3.4	2.8	0.6	0.6	0.6
JE	The Bailiwick of Jersey	JEY	N/A	N/A	N/A	11,737,000	11,548,000	11,358,000				0.5	0.5	0.6
JO	Jordan	JOR	Middle East & North Africa	Middle-Income	Middle-Income	123,511,000	129,686,000	134,867,000	1.7	1.7	1.7	0.6	0.7	0.7
JP	Japan	JPN	East Asia & Pacific	High-Income	High-Income	13,493,770,000	12,850,143,000	12,807,032,000	9.5	9.3	9.3	0.5	0.6	0.6
KZ	Kazakhstan	KAZ	Europe & Central Asia	Middle-Income	Middle-Income	549,402,000	562,897,000	572,751,000	3.4	3.4	3.3	0.5	0.6	0.7
KE	Kenya	KEN	Sub-Saharan Africa	Low-Income	Middle-Income	180,449,000	175,311,000	159,163,000	1.9	1.6	1.4	0.7	0.7	0.7
KG	Kyrgyz Republic	KGZ	Europe & Central Asia	Middle-Income	Middle-Income	70,986,000	79,685,000	88,723,000	2.8	2.8	3.1	0.6	0.7	0.8
KH	Cambodia	KHM	East Asia & Pacific	Low-Income	Low-Income	39,542,000	50,980,000	57,536,000	0.6	0.8	0.7	0.5	0.5	0.6
KI	Kiribati	KIR	East Asia & Pacific	Low-Income	Low-Income	160,000	234,000	317,000	0.9	1.1	1.6	0.4	0.5	0.6
KN	St. Kitts and Nevis	KNA	Latin America & Caribbean	High-Income	High-Income	2,461,000	2,381,000	1,883,000	4.9	4.4	3.7	0.4	0.5	0.6

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
KR	Korea, Rep.	KOR	East Asia & Pacific	High-Income	High-Income	1,225,832,000	1,156,863,000	1,208,146,000	2.1	2	2	0.5	0.5	0.5
KW	Kuwait	KWT	Middle East & North Africa	High-Income	High-Income	125,701,000	124,436,000	126,361,000	3.3	2.6	2.6	0.7	0.7	0.8
LA	Lao PDR	LAO	East Asia & Pacific	Low-Income	Low-Income	12,734,000	14,455,000	16,303,000	0.7	0.7	0.8	0.7	0.7	0.7
LB	Lebanon	LBN	Middle East & North Africa	Middle-Income	Middle-Income	108,908,000	111,384,000	112,365,000	2	2	1.7	0.6	0.6	0.7
LR	Liberia	LBR	Sub-Saharan Africa	Low-Income	Low-Income	6,701,000	7,914,000	8,912,000	1.7	1.7	1.9	0.8	0.8	0.7
LY	Libya	LBY	Middle East & North Africa	Middle-Income	Middle-Income	36,061,000	36,643,000	36,152,000	2.4	2.2	2.1	0.6	0.7	0.7
LC	St. Lucia	LCA	Latin America & Caribbean	Middle-Income	Middle-Income	5,447,000	5,731,000	5,177,000	5.5	5.3	4.7	0.5	0.5	0.6
LI	Liechtenstein	LIE	Europe & Central Asia	High-Income	High-Income	5,398,000	5,389,000	4,895,000	12.2	12.1	11	0.3	0.4	0.4
LK	Sri Lanka	LKA	South Asia	Middle-Income	Middle-Income	143,363,000	152,885,000	157,256,000	1.8	1.7	1.8	0.5	0.6	0.6
LS	Lesotho	LSO	Sub-Saharan Africa	Low-Income	Low-Income	4,690,000	4,732,000	3,798,000	0.6	0.6	0.5	0.8	0.8	0.8
LT	Lithuania	LTU	Europe & Central Asia	High-Income	High-Income	202,377,000	207,371,000	212,675,000	7.9	7.9	8	0.3	0.4	0.4
LU	Luxembourg	LUX	Europe & Central Asia	High-Income	High-Income	91,094,000	85,105,000	88,992,000	13.3	12.2	12.6	0.4	0.4	0.5
LV	Latvia	LVA	Europe & Central Asia	High-Income	High-Income	176,380,000	160,147,000	158,519,000	9.4	8.5	8.2	0.3	0.3	0.4
MO	Macao SAR, China	MAC	East Asia & Pacific	High-Income	High-Income	53,490,000	55,292,000	57,197,000	8.9	8.9	9	0.4	0.4	0.5
MF	St. Martin (French part)	MAF	Latin America & Caribbean	High-Income	High-Income	1,523,000	1,267,000	753,000				0.5	0.5	0.6
MA	Morocco	MAR	Middle East & North Africa	Middle-Income	Middle-Income	408,814,000	450,283,000	487,104,000	1.7	1.7	1.7	0.5	0.6	0.7

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
MC	Monaco	MCO	Europe & Central Asia	High-Income	High-Income	5,267,000	5,031,000	4,955,000	12	11.2	11	0.3	0.4	0.4
MD	Moldova	MDA	Europe & Central Asia	Middle-Income	Middle-Income	110,520,000	109,479,000	100,790,000	3.7	3.4	3.1	0.3	0.4	0.5
MG	Madagascar	MDG	Sub-Saharan Africa	Low-Income	Low-Income	30,029,000	28,910,000	28,464,000	2.1	1	0.9	0.4	0.5	0.5
MV	Maldives	MDV	South Asia	Middle-Income	Middle-Income	13,776,000	14,499,000	15,986,000	4.5	4.4	4.1	0.5	0.6	0.7
MX	Mexico	MEX	Latin America & Caribbean	Middle-Income	Middle-Income	3,376,393,000	3,146,779,000	2,975,842,000	3.7	3.2	3	0.5	0.6	0.6
MH	Marshall Islands	MHL	East Asia & Pacific	Middle-Income	Middle-Income	515,000	403,000	443,000	2.7	1.6	1.6	0.3	0.4	0.4
MK	Macedonia, FYR	MKD	Europe & Central Asia	Middle-Income	Middle-Income	93,868,000	96,925,000	99,927,000	5.2	5.1	5	0.3	0.4	0.5
ML	Mali	MLI	Sub-Saharan Africa	Low-Income	Low-Income	11,859,000	12,820,000	12,862,000	0.5	0.5	0.4	0.7	0.8	0.8
MT	Malta	MLT	Middle East & North Africa	High-Income	High-Income	48,875,000	52,344,000	47,413,000	11.6	11.6	10	0.4	0.4	0.4
MM	Myanmar	MMR	East Asia & Pacific	Low-Income	Low-Income	48,288,000	54,233,000	58,808,000	0.3	0.3	0.3	0.7	0.7	0.7
ME	Montenegro	MNE	Europe & Central Asia	Middle-Income	Middle-Income	38,638,000	41,933,000	44,305,000	7.4	7.9	8.3	0.5	0.6	0.7
MN	Mongolia	MNG	East Asia & Pacific	Middle-Income	Middle-Income	30,585,000	34,339,000	36,148,000	3.8	3.9	4	0.3	0.4	0.5
MP	Northern Mariana Islands	MNP	East Asia & Pacific	High-Income	High-Income	2,263,000	2,022,000	1,394,000				0.3	0.4	0.5
MZ	Mozambique	MOZ	Sub-Saharan Africa	Low-Income	Low-Income	49,665,000	51,091,000	36,036,000	0.8	0.7	0.5	0.8	0.8	0.8
MR	Mauritania	MRT	Sub-Saharan Africa	Low-Income	Low-Income	7,339,000	7,463,000	7,471,000	0.8	0.7	0.7	0.6	0.7	0.8
MS	Montserrat	MSR	N/A	Middle-Income	Middle-Income	129,000	236,000	149,000				0.4	0.3	0.4

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
MQ	Martinique	MTQ	N/A	N/A	N/A	19,009,000	17,727,000	17,610,000				0.5	0.6	0.6
MU	Mauritius	MUS	Sub-Saharan Africa	Middle-Income	Middle-Income	44,347,000	43,700,000	42,483,000	5.6	5.2	4.8	0.5	0.5	0.6
MW	Malawi	MWI	Sub-Saharan Africa	Low-Income	Low-Income	9,164,000	8,688,000	8,262,000	0.4	0.3	0.3	0.8	0.7	0.7
MY	Malaysia	MYS	East Asia & Pacific	Middle-Income	Middle-Income	948,146,000	960,067,000	999,890,000	3.2	3.2	3.3	0.5	0.6	0.6
YT	The Department of Mayotte	MYT	N/A	N/A	N/A	3,045,000	2,970,000	2,960,000				0.3	0.4	0.4
NA	Namibia	NAM	Sub-Saharan Africa	Middle-Income	Middle-Income	15,240,000	14,893,000	14,609,000	1.7	1.3	1.3	0.6	0.6	0.6
NC	New Caledonia	NCL	East Asia & Pacific	High-Income	High-Income	11,308,000	10,840,000	9,848,000		3.9	3.5	0.3	0.4	0.4
NE	Niger	NER	Sub-Saharan Africa	Low-Income	Low-Income	7,029,000	5,683,000	5,950,000	0.7	0.2	0.4	0.8	0.8	0.8
NF	The Territory of Norfolk Island	NFK	N/A	N/A	N/A	13,000	12,000	49,000				0.2	0.5	0.5
NG	Nigeria	NGA	Sub-Saharan Africa	Middle-Income	Middle-Income	392,137,000	423,522,000	376,974,000	0.7	0.7	0.6	0.8	0.8	0.8
NI	Nicaragua	NIC	Latin America & Caribbean	Middle-Income	Middle-Income	65,926,000	66,656,000	59,285,000	3.6	3.2	2.7	0.6	0.6	0.7
NU	Niue	NIU	N/A	Middle-Income	Middle-Income	26,000	25,000	53,000				0.3	0.3	0.4
NL	Netherlands	NLD	Europe & Central Asia	High-Income	High-Income	2,373,123,000	2,476,994,000	2,707,940,000	12.8	12.9	13.8	0.4	0.4	0.4
NO	Norway	NOR	Europe & Central Asia	High-Income	High-Income	763,094,000	752,761,000	751,109,000	12.5	12.3	12.2	0.4	0.5	0.5
NP	Nepal	NPL	South Asia	Low-Income	Low-Income	113,933,000	101,905,000	107,943,000	1.7	1.4	1.5	0.7	0.6	0.6
NR	Nauru	NRU	East Asia & Pacific	Middle-Income	Middle-Income	101,000	71,000	58,000		0.8	0.7	0.4	0.3	0.3
NZ	New Zealand	NZL	East Asia & Pacific	High-Income	High-Income	526,266,000	503,666,000	488,240,000	10.6	9.6	9.2	0.4	0.4	0.5

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
OM	Oman	OMN	Middle East & North Africa	High-Income	High-Income	70,987,000	73,099,000	75,221,000	1.7	1.6	1.6	0.6	0.6	0.7
PK	Pakistan	PAK	South Asia	Middle-Income	Middle-Income	966,058,000	1,765,838,000	1,846,721,000	2.7	4.8	4.7	0.3	0.2	0.2
PA	Panama	PAN	Latin America & Caribbean	Middle-Income	Middle-Income	158,009,000	169,480,000	153,865,000	6	6	5.3	0.5	0.5	0.6
PN	Pitcairn Islands	PCN	N/A	N/A	N/A	4,000	4,000							
PE	Peru	PER	Latin America & Caribbean	Middle-Income	Middle-Income	784,167,000	749,612,000	738,537,000	4.5	4	3.7	0.4	0.4	0.5
PH	Philippines	PHL	East Asia & Pacific	Middle-Income	Middle-Income	1,301,166,000	1,348,709,000	1,363,696,000	1.9	1.8	1.8	0.4	0.5	0.5
PW	Palau	PLW	East Asia & Pacific	Middle-Income	Middle-Income	531,000	408,000	581,000				0.3	0.4	0.6
PG	Papua New Guinea	PNG	East Asia & Pacific	Middle-Income	Middle-Income	6,691,000	7,964,000	8,031,000	0.7	0.7	0.7	0.6	0.7	0.7
PL	Poland	POL	Europe & Central Asia	High-Income	High-Income	3,277,780,000	3,281,829,000	3,362,620,000	9.8	9.5	9.5	0.3	0.4	0.5
PR	Puerto Rico	PRI	Latin America & Caribbean	High-Income	High-Income	128,298,000	95,495,000	98,432,000	4.6	3.3	3.5	0.6	0.6	0.7
KP	Korea, Dem. People's Rep.	PRK	East Asia & Pacific	Low-Income	Low-Income	154,000	262,000	140,000				0.2	0.1	0.2
PT	Portugal	PRT	Europe & Central Asia	High-Income	High-Income	584,278,000	586,615,000	593,073,000	6.7	6.4	6.4	0.3	0.4	0.5
PY	Paraguay	PRY	Latin America & Caribbean	Middle-Income	Middle-Income	117,789,000	109,833,000	110,719,000	2.7	2.2	2	0.6	0.7	0.7
PS	West Bank and Gaza	PSE	Middle East & North Africa	Middle-Income	Middle-Income	46,084,000	54,210,000	60,201,000	1.4	1.5	1.7	0.5	0.6	0.6
PF	French Polynesia	PYF	East Asia & Pacific	High-Income	High-Income	8,526,000	8,401,000	8,299,000	3.7	3.4	3.4	0.4	0.4	0.5
QA	Qatar	QAT	Middle East & North Africa	High-Income	High-Income	101,875,000	98,183,000	95,593,000	3.5	3.2	2.9	0.5	0.6	0.6
RE	Réunion	REU	N/A	N/A	N/A	44,308,000	40,139,000	39,330,000				0.4	0.5	0.5

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
RO	Romania	ROU	Europe & Central Asia	Middle-Income	Middle-Income	773,164,000	770,273,000	738,361,000	5.5	5.1	4.5	0.4	0.4	0.5
RU	Russian Federation	RUS	Europe & Central Asia	Middle-Income	Middle-Income	8,547,862,000	7,657,877,000	7,011,798,000	6.8	5.8	5	0.4	0.4	0.4
RW	Rwanda	RWA	Sub-Saharan Africa	Low-Income	Low-Income	14,693,000	14,257,000	14,075,000	0.5	0.4	0.4	0.6	0.6	0.6
SA	Saudi Arabia	SAU	Middle East & North Africa	High-Income	High-Income	657,886,000	654,357,000	669,932,000	2.3	2	1.8	0.6	0.7	0.8
SD	Sudan	SDN	Sub-Saharan Africa	Low-Income	Low-Income	61,891,000	62,215,000	58,010,000	0.5	0.4	0.4	0.8	0.9	0.9
SN	Senegal	SEN	Sub-Saharan Africa	Low-Income	Low-Income	39,863,000	38,562,000	38,604,000	0.8	0.7	0.7	0.6	0.6	0.7
SG	Singapore	SGP	East Asia & Pacific	High-Income	High-Income	741,275,000	766,093,000	806,054,000	13	13.5	13.5	0.5	0.5	0.5
GS	South Georgia and the South Sandwich Islands	SGS	N/A	N/A	N/A		3,000	4,000					0.4	0.5
SH	Saint Helena, Ascension and Tristan da Cunha	SHN	N/A	Middle-Income	Middle-Income	4,000	31,000	88,000					0.2	0.2
SJ	Svalbard and Jan Mayen	SJM	N/A	N/A	N/A	6,000	21,000	22,000					0.2	0.5
SB	Solomon Islands	SLB	East Asia & Pacific	Low-Income	Low-Income	1,193,000	1,242,000	1,243,000	1.5	1.4	1.3	0.6	0.6	0.6
SL	Sierra Leone	SLE	Sub-Saharan Africa	Low-Income	Low-Income	4,997,000	4,883,000	5,272,000	0.5	0.4	0.4	0.8	0.8	0.8
SV	El Salvador	SLV	Latin America & Caribbean	Middle-Income	Middle-Income	128,634,000	123,252,000	116,212,000	5.8	5.2	4.8	0.5	0.6	0.6
SM	San Marino	SMR	Europe & Central Asia	High-Income	High-Income	2,340,000	1,971,000	2,073,000		8.2	8.5	0.4	0.5	0.5
SO	Somalia	SOM	Sub-Saharan Africa	Low-Income	Low-Income	10,345,000	12,387,000	15,040,000	3.2	3.5	4.2	0.6	0.7	0.8
PM	The Overseas Collectivity of Saint-Pierre and Miquelon	SPM	N/A	N/A	N/A	567,000	527,000	621,000				0.3	0.3	0.3

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
RS	Serbia	SRB	Europe & Central Asia	Middle-Income	Middle-Income	459,855,000	478,492,000	495,803,000	8.1	8.1	8.1	0.4	0.4	0.5
SS	South Sudan	SSD	Sub-Saharan Africa	Low-Income	Low-Income	2,744,000	2,584,000	2,472,000	0.3	0.2	0.2	0.8	0.8	0.8
ST	São Tomé and Principe	STP	Sub-Saharan Africa	Low-Income	Low-Income	732,000	750,000	858,000	1.1	1	1.1	0.5	0.6	0.7
SR	Suriname	SUR	Latin America & Caribbean	Middle-Income	Middle-Income	12,812,000	12,711,000	13,146,000	4.2	3.8	3.9	0.6	0.6	0.6
SK	Slovak Republic	SVK	Europe & Central Asia	High-Income	High-Income	285,932,000	291,150,000	306,733,000	5.5	5.5	5.8	0.3	0.4	0.4
SI	Slovenia	SVN	Europe & Central Asia	High-Income	High-Income	189,276,000	190,492,000	182,484,000	10.1	9.7	9.2	0.3	0.4	0.4
SE	Sweden	SWE	Europe & Central Asia	High-Income	High-Income	1,807,282,000	1,755,640,000	1,753,845,000	16.9	15.1	15.6	0.5	0.5	0.5
SZ	Swaziland	SWZ	Sub-Saharan Africa	Middle-Income	Middle-Income	3,128,000	3,538,000	3,559,000	0.7	0.7	0.9	0.7	0.7	0.7
SX	Sint Maarten (Dutch part)	SXM	Latin America & Caribbean	High-Income	High-Income	2,650,000	2,308,000	2,128,000				0.5	0.5	0.5
SC	Seychelles	SYC	Sub-Saharan Africa	Middle-Income	High-Income	4,491,000	4,305,000	6,476,000	7	6.4	9.5	0.4	0.5	0.4
SY	Syrian Arab Republic	SYR	Middle East & North Africa	Middle-Income	Middle-Income	46,174,000	57,344,000	70,926,000	0.7	0.8	1	0.7	0.8	0.8
TC	Turks and Caicos Islands	TCA	Latin America & Caribbean	High-Income	High-Income	1,836,000	1,718,000	1,622,000				0.5	0.6	0.6
TD	Chad	TCD	Sub-Saharan Africa	Low-Income	Low-Income	3,797,000	4,437,000	2,028,000	0.4	0.4	0.2	0.9	0.9	0.8
TG	Togo	TGO	Sub-Saharan Africa	Low-Income	Low-Income	7,526,000	8,437,000	9,487,000	0.7	0.7	0.8	0.6	0.6	0.7
TH	Thailand	THA	East Asia & Pacific	Middle-Income	Middle-Income	935,258,000	978,569,000	1,050,125,000	2.4	2.2	2.2	0.5	0.6	0.6
TJ	Tajikistan	TJK	Europe & Central Asia	Low-Income	Middle-Income	22,602,000	23,932,000	22,763,000	1.1	1	0.9	0.7	0.8	0.8

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
TK	Tokelau	TKL	N/A	N/A	N/A	5,000	8,000	12,000					0.7	0.6
TM	Turkmenistan	TKM	Europe & Central Asia	Middle-Income	Middle-Income	13,524,000	22,050,000	29,940,000	1.1	1.5	2	0.6	0.5	0.5
TL	Timor-Leste	TLS	East Asia & Pacific	Low-Income	Low-Income	3,277,000	3,427,000	3,030,000	0.9	0.8	0.7	0.5	0.6	0.6
TO	Tonga	TON	East Asia & Pacific	Middle-Income	Middle-Income	541,000	543,000	484,000	1.1	1	0.9	0.3	0.3	0.4
TT	Trinidad and Tobago	TTO	Latin America & Caribbean	High-Income	High-Income	50,904,000	49,170,000	44,775,000	4.2	3.9	3.5	0.5	0.5	0.6
TN	Tunisia	TUN	Middle East & North Africa	Middle-Income	Middle-Income	167,887,000	171,830,000	159,830,000	2.5	2.2	1.8	0.4	0.5	0.6
TR	Turkey	TUR	Europe & Central Asia	Middle-Income	Middle-Income	2,020,146,000	813,677,000	55,329,000	3.6	1.3	0.1	0.6	0.6	0.4
TV	Tuvalu	TUV	East Asia & Pacific	Low-Income	Low-Income	20,000	22,000	32,000	0.3	0.3	0.5	0.3	0.3	0.4
TW	Taiwan, China	TWN	East Asia & Pacific	High-Income	High-Income	2,340,961,000	2,516,165,000	2,732,209,000				0.4	0.4	0.5
TZ	Tanzania	TZA	Sub-Saharan Africa	Low-Income	Low-Income	71,651,000	70,458,000	61,240,000	0.8	0.6	0.6	0.8	0.8	0.8
UG	Uganda	UGA	Sub-Saharan Africa	Low-Income	Low-Income	45,976,000	48,853,000	38,934,000	0.4	0.4	0.3	0.7	0.7	0.7
UA	Ukraine	UKR	Europe & Central Asia	Middle-Income	Middle-Income	2,206,314,000	2,068,007,000	2,105,759,000	7.7	6.7	6.9	0.3	0.4	0.4
UM	Baker Island, Howland Island, Jarvis Island, Johnston Atoll, Kingman Reef, Midway Atoll, Navassa Island, Palmyra Atoll, and Wake Island	UMI	N/A	N/A	N/A		25,000	31,000						
UY	Uruguay	URY	Latin America & Caribbean	Middle-Income	High-Income	183,902,000	179,696,000	168,568,000	6.7	6.3	6	0.5	0.6	0.6
US	United States	USA	North America	High-Income	High-Income	41,942,039,000	41,631,207,000	41,415,111,000	14.4	14.2	14	0.5	0.5	0.6

country	countryname	ISO3	regionwb	oecd14-17	oecd	views_ceil16	views_ceil17	views_ceil18	wikipc16	wikipc17	wikipc18	%mob16	%mob17	%mob18
UZ	Uzbekistan	UZB	Europe & Central Asia	Middle-Income	Middle-Income	73,289,000	81,199,000	92,639,000	0.4	0.4	0.4	0.6	0.6	0.7
VA	The Holy See	VAT	N/A	N/A	N/A	915,000	865,000	764,000				0.1	0.2	0.2
VC	St. Vincent and the Grenadines	VCT	Latin America & Caribbean	Middle-Income	Middle-Income	2,010,000	2,726,000	2,483,000	2.7	3.2	8.4	0.5	0.6	0.6
VE	Venezuela, RB	VEN	Latin America & Caribbean	Middle-Income	Middle-Income	751,967,000	645,733,000	509,546,000	3.3	2.6	2.3	0.4	0.4	0.4
VG	British Virgin Islands	VGB	Latin America & Caribbean	High-Income	High-Income	1,781,000	1,735,000	681,000				0.4	0.3	0.5
VI	Virgin Islands (U.S.)	VIR	Latin America & Caribbean	High-Income	High-Income	4,358,000	3,270,000	2,725,000	5.7	3.9	3.3	0.5	0.5	0.6
VN	Vietnam	VNM	East Asia & Pacific	Middle-Income	Middle-Income	726,830,000	783,137,000	874,769,000	1.4	1.4	1.1	0.5	0.5	0.6
VU	Vanuatu	VUT	East Asia & Pacific	Low-Income	Low-Income	1,023,000	1,197,000	1,037,000	1.3	1.4	1.1	0.3	0.4	0.4
WF	The Territory of the Wallis and Futuna Islands	WLF	N/A	Middle-Income	Middle-Income	152,000	151,000	155,000				0.3	0.4	0.5
WS	Samoa	WSM	East Asia & Pacific	Middle-Income	Middle-Income	632,000	579,000	759,000	0.9	0.7	1	0.4	0.5	0.6
YE	Yemen, Rep.	YEM	Middle East & North Africa	Low-Income	Low-Income	29,221,000	31,631,000	31,271,000	0.4	0.3	0.3	0.7	0.8	0.8
ZA	South Africa	ZAF	Sub-Saharan Africa	Middle-Income	Middle-Income	570,812,000	549,854,000	567,636,000	1.6	1.4	1.5	0.6	0.6	0.7
ZM	Zambia	ZMB	Sub-Saharan Africa	Low-Income	Low-Income	31,551,000	30,175,000	30,238,000	0.6	0.5	1	0.8	0.8	0.8
ZW	Zimbabwe	ZWE	Sub-Saharan Africa	Low-Income	Low-Income	31,716,000	31,363,000	28,110,000	0.7	0.6	0.6	0.6	0.6	0.6

4. Kruskal-Wallis Tests

Table A.4.1 Kruskal-Wallis tests, page views per connected capita 2018

```
. kwallis wikipc18, by (oecd)

Kruskal-Wallis equality-of-populations rank test
```

oecd	Obs	Rank Sum
High	69	11133.00
Low	47	1841.00
Middle	89	8141.00

```
chi-squared =   124.525 with 2 d.f.
probability =     0.0001

chi-squared with ties =   124.525 with 2 d.f.
probability =     0.0001
```

Table A.4.2 Kruskal-Wallis tests, technological subdimension, mobile devices 2018

```
. kwallis mob18, by (oecd)

Kruskal-Wallis equality-of-populations rank test
```

oecd	Obs	Rank Sum
High	76	6050.00
Low	49	7685.00
Middle	96	10796.00

```
chi-squared =    43.552 with 2 d.f.
probability =     0.0001

chi-squared with ties =    43.552 with 2 d.f.
probability =     0.0001
```

5. Wikipedia Size Categories and Income Level Trends

Figure A.2 shows the proportion of views by Wikipedia size category disaggregated by income level for 2016-2018. The top of the horizontal axis shows categories (A-F); then it presents income level for each category; at the bottom it presents the years. The vertical axis presents the share of the volume of pageviews consumed by each income level within categories for each year.

Figure A.2: Proportion of views by wiki size category and income level (2016-2018)

Figure A.2 is helpful to identify patterns within two groups of Wikipedias: A-C; and D-F. Within the first group the trends seem to be stable. Category A languages remain without noticeable changes within this period; high-income and middle-income countries show opposite trends in categories B and C; and within the three categories high-income countries consumed the majority of contents in 2018. The second group, categories D-F shows mixed trends. Category D languages were the only category that was consumed more by middle-income countries in 2018, yet the trend is not clear. Categories E and F show some mixed patterns over time; it is interesting to note that these smaller Wikipedias are being consumed mostly by high-income countries. Within categories D-F the group of low-income countries seem to have a greater share in the consumption that is increasing over time. Overall, high-income countries appear as the ones consuming the majority of the contents in almost all the categories – including the smallest ones – however middle-income countries appear very close in some categories.

Figure A.3 presents page views by income level for each Wikipedia size category for 2016-2018. The top row on the horizontal axis shows income level; then is broken down by category size; and the bottom shows the years 2016-2018. The vertical axis shows the proportion of page views consumed within each income level, for each one of the category sizes during the three years.

Figure A.3 Page views by income level and wiki size (2016-2018)

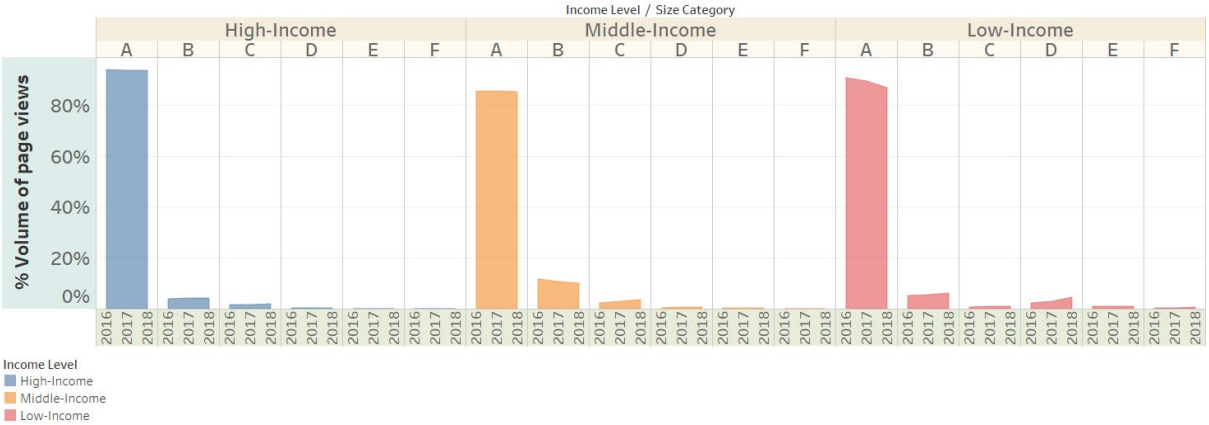


Figure A.3 is useful to identify the trends within the three income levels and compare between them. High-income countries consume the majority of their contents from Category A languages and a small fraction from categories B and C, with almost no changes during this period; just a tiny fraction comes from categories D-F. A similar pattern is identified for middle-income countries, with a higher consumption of categories B and C, and an increase over time for the latter category. Low-income countries have a more varied consumption pattern even though the majority of their consumption comes from Category A languages: contents from categories B and D appear as more relevant than they were for other income groups, and for this group the remaining categories account for at least around one percent each. Over time the consumption of Category A languages seems to be decreasing in favour of the other language categories. However, for the three groups Category A languages clearly dominate the consumption.[32]

[32] These analyses were also conducted excluding the English Wikipedia to contrast the analyses; the results were very similar so these are not included here.

6. Wiki Indicators and Wikindex values

Table A.6.1 wikiprod, wikiconsum, and wikindex scores by country

Rank	Country	ISO3	Region World Bank	Income (2017)	wikiprod (GII)	wikiconsum	wikindex (2017)
1	Israel	ISR	Middle East & North Africa	High-Income	100	60.7	80.4
2	Estonia	EST	Europe & Central Asia	High-Income	89.9	70.2	80.0
3	Finland	FIN	Europe & Central Asia	High-Income	66.1	92.0	79.0
4	Sweden	SWE	Europe & Central Asia	High-Income	71.8	84.9	78.4
5	Iceland	ISL	Europe & Central Asia	High-Income	70.6	78.4	74.5
6	Norway	NOR	Europe & Central Asia	High-Income	70.3	66.5	68.4
7	Ireland	IRL	Europe & Central Asia	High-Income	33.4	100.0	66.7
8	Hong Kong (China)	HKG	East Asia & Pacific	High-Income	56.7	70.7	63.7
9	Luxembourg	LUX	Europe & Central Asia	High-Income	59.1	66.6	62.8
10	Netherlands	NLD	Europe & Central Asia	High-Income	58.1	67.0	62.6
11	United Kingdom	GBR	Europe & Central Asia	High-Income	46.7	68.0	57.4
12	Germany	DEU	Europe & Central Asia	High-Income	35.1	70.0	52.5
13	Latvia	LVA	Europe & Central Asia	High-Income	67.2	37.1	52.1
14	Switzerland	CHE	Europe & Central Asia	High-Income	31.9	68.0	49.9
15	Austria	AUT	Europe & Central Asia	High-Income	36.5	62.8	49.7
16	France	FRA	Europe & Central Asia	High-Income	43.6	54.7	49.2
17	Slovenia	SVN	Europe & Central Asia	High-Income	55.9	41.5	48.7
18	Canada	CAN	North America	High-Income	33	62.1	47.6
19	New Zealand	NZL	East Asia & Pacific	High-Income	40.2	50.7	45.5
20	Denmark	DNK	Europe & Central Asia	High-Income	32.3	56.1	44.2
21	Australia	AUS	East Asia & Pacific	High-Income	31.8	54.5	43.1
22	Armenia	ARM	Europe & Central Asia	Middle-Income	69	16.2	42.6
23	Malta	MLT	Middle East & North Africa	High-Income	28.2	53.4	40.8
24	Czech Republic	CZE	Europe & Central Asia	High-Income	37.8	41.6	39.7
25	Italy	ITA	Europe & Central Asia	High-Income	29.7	49.7	39.7
26	United States of America	USA	North America	High-Income	17.6	59.7	38.7
27	Spain	ESP	Europe & Central Asia	High-Income	39.6	35.1	37.3
28	Uruguay	URY	Latin America & Caribbean	Middle-Income	45.9	25.0	35.5
29	Singapore	SGP	East Asia & Pacific	High-Income	16	54.6	35.3
30	Lithuania	LTU	Europe & Central Asia	High-Income	36.6	32.2	34.4

Rank	Country	ISO3	Region World Bank	Income (2017)	wikiprod (GII)	wikiconsum	wikindex (2017)
31	Hungary	HUN	Europe & Central Asia	High-Income	36.2	31.1	33.6
32	Belgium	BEL	Europe & Central Asia	High-Income	20.7	46.5	33.6
33	Croatia	HRV	Europe & Central Asia	High-Income	22.3	39.0	30.6
34	Poland	POL	Europe & Central Asia	High-Income	23.1	37.6	30.4
35	Japan	JPN	East Asia & Pacific	High-Income	12.6	47.9	30.2
36	Bulgaria	BGR	Europe & Central Asia	Middle-Income	31.1	28.3	29.7
37	Cyprus	CYP	Europe & Central Asia	High-Income	34.5	22.9	28.7
38	TFYR Macedonia	MKD	Europe & Central Asia	Middle-Income	31.7	20.2	25.9
39	Serbia	SRB	Europe & Central Asia	Middle-Income	27.1	24.1	25.6
40	Bosnia and Herzegovina	BIH	Europe & Central Asia	Middle-Income	28.1	19.7	23.9
41	Georgia	GEO	Europe & Central Asia	Middle-Income	30	17.1	23.6
42	Montenegro	MNE	Europe & Central Asia	Middle-Income	16.3	30.2	23.2
43	Greece	GRC	Europe & Central Asia	High-Income	16.4	29.6	23.0
44	Portugal	PRT	Europe & Central Asia	High-Income	15.8	25.9	20.9
45	Ukraine	UKR	Europe & Central Asia	Middle-Income	21	20.3	20.6
46	Slovakia	SVK	Europe & Central Asia	High-Income	18	23.0	20.5
47	Russian Federation	RUS	Europe & Central Asia	Middle-Income	13.3	23.3	18.3
48	Belarus	BLR	Europe & Central Asia	Middle-Income	15	20.9	18.0
49	Chile	CHL	Latin America & Caribbean	Middle-Income	10.9	20.7	15.8
50	Iran, Islamic Republic of	IRN	Middle East & North Africa	Middle-Income	6.3	24.9	15.6
51	Brunei Darussalam	BRN	East Asia & Pacific	High-Income	3.8	26.9	15.3
52	Panama	PAN	Latin America & Caribbean	Middle-Income	9.8	20.7	15.2
53	Albania	ALB	Europe & Central Asia	Middle-Income	10.7	18.6	14.6
54	Bahrain	BHR	Middle East & North Africa	High-Income	10.9	17.1	14.0
55	Argentina	ARG	Latin America & Caribbean	Middle-Income	8.2	19.3	13.8
56	Kazakhstan	KAZ	Europe & Central Asia	Middle-Income	11.6	15.1	13.4
57	Azerbaijan	AZE	Europe & Central Asia	Middle-Income	17.6	8.9	13.2
58	Romania	ROU	Europe & Central Asia	Middle-Income	8.4	17.2	12.8
59	United Arab Emirates	ARE	Middle East & North Africa	High-Income	6.6	18.1	12.4
60	Kuwait	KWT	Middle East & North Africa	High-Income	11	13.5	12.3

Rank	Country	ISO3	Region World Bank	Income (2017)	wikiprod (GII)	wikiconsum	wikindex (2017)
61	Costa Rica	CRI	Latin America & Caribbean	Middle-Income	7.4	16.0	11.7
62	Moldova, Republic of	MDA	Europe & Central Asia	Middle-Income	11.5	11.2	11.4
63	Jordan	JOR	Middle East & North Africa	Middle-Income	14.6	7.3	10.9
64	Korea, Republic of	KOR	East Asia & Pacific	High-Income	11.9	9.4	10.7
65	Trinidad and Tobago	TTO	Latin America & Caribbean	High-Income	5.4	15.9	10.6
66	Malaysia	MYS	East Asia & Pacific	Middle-Income	6.3	14.4	10.3
67	Qatar	QAT	Middle East & North Africa	High-Income	5.7	14.4	10.0
68	Mauritius	MUS	Sub-Saharan Africa	Middle-Income	4	15.0	9.5
69	Dominican Republic	DOM	Latin America & Caribbean	Middle-Income	3.9	13.7	8.8
70	Colombia	COL	Latin America & Caribbean	Middle-Income	3.2	13.5	8.3
71	Mongolia	MNG	East Asia & Pacific	Middle-Income	10.6	5.3	7.9
72	Peru	PER	Latin America & Caribbean	Middle-Income	3.9	11.8	7.9
73	Ecuador	ECU	Latin America & Caribbean	Middle-Income	3.3	12.0	7.6
74	Mexico	MEX	Latin America & Caribbean	Middle-Income	2.3	12.3	7.3
75	Saudi Arabia	SAU	Middle East & North Africa	High-Income	4.1	9.2	6.6
76	Lebanon	LBN	Middle East & North Africa	Middle-Income	5	8.0	6.5
77	El Salvador	SLV	Latin America & Caribbean	Middle-Income	3.2	9.4	6.3
78	Brazil	BRA	Latin America & Caribbean	Middle-Income	4.2	7.9	6.0
79	Bolivia, Plurinational State of	BOL	Latin America & Caribbean	Middle-Income	2.4	9.6	6.0
80	Kyrgyzstan	KGZ	Europe & Central Asia	Middle-Income	4.9	6.5	5.7
81	Oman	OMN	Middle East & North Africa	High-Income	3.9	7.1	5.5
82	Paraguay	PRY	Latin America & Caribbean	Middle-Income	2.9	8.0	5.4
83	Jamaica	JAM	Latin America & Caribbean	Middle-Income	1.6	9.0	5.3
84	Thailand	THA	East Asia & Pacific	Middle-Income	3.8	6.0	4.9
85	Guatemala	GTM	Latin America & Caribbean	Middle-Income	2.6	7.1	4.8
86	Morocco	MAR	Middle East & North Africa	Middle-Income	3.5	6.1	4.8
87	Philippines	PHL	East Asia & Pacific	Middle-Income	2.5	6.5	4.5
88	Tunisia	TUN	Middle East & North Africa	Middle-Income	1.9	6.9	4.4
89	Viet Nam	VNM	East Asia & Pacific	Middle-Income	4.8	3.6	4.2
90	Honduras	HND	Latin America & Caribbean	Middle-Income	1.9	6.0	4.0
91	Turkey	TUR	Europe & Central Asia	Middle-Income	3	4.7	3.9

Rank	Country	ISO3	Region World Bank	Income (2017)	wikiprod (GII)	wikiconsum	wikindex (2017)
92	South Africa	ZAF	Sub-Saharan Africa	Middle-Income	2.8	4.6	3.7
93	Sri Lanka	LKA	South Asia	Middle-Income	4.1	3.2	3.7
94	Algeria	DZA	Middle East & North Africa	Middle-Income	2.5	4.8	3.6
95	Nepal	NPL	South Asia	Low-Income	4.1	1.8	2.9
96	Namibia	NAM	Sub-Saharan Africa	Middle-Income	2.4	3.3	2.8
97	Pakistan	PAK	South Asia	Middle-Income	1	4.5	2.8
98	Indonesia	IDN	East Asia & Pacific	Middle-Income	1.3	3.3	2.3
99	Egypt	EGY	Middle East & North Africa	Middle-Income	1.7	2.6	2.1
100	India	IND	South Asia	Middle-Income	0.7	2.5	1.6
101	Tajikistan	TJK	Europe & Central Asia	Low-Income	1.7	1.3	1.5
102	Botswana	BWA	Sub-Saharan Africa	Middle-Income	0.2	2.7	1.5
103	Cambodia	KHM	East Asia & Pacific	Low-Income	1.2	1.4	1.3
104	Kenya	KEN	Sub-Saharan Africa	Low-Income	0.7	1.9	1.3
105	Côte d'Ivoire	CIV	Sub-Saharan Africa	Middle-Income	0.3	1.4	0.9
106	Bangladesh	BGD	South Asia	Low-Income	0.7	0.8	0.8
107	Senegal	SEN	Sub-Saharan Africa	Low-Income	0.1	1.4	0.7
108	Nigeria	NGA	Sub-Saharan Africa	Middle-Income	0.2	1.2	0.7
109	Zimbabwe	ZWE	Sub-Saharan Africa	Low-Income	0.2	1.1	0.7
110	Yemen	YEM	Middle East & North Africa	Low-Income	0.8	0.5	0.6
111	Benin	BEN	Sub-Saharan Africa	Low-Income	0.6	0.7	0.6
112	Cameroon	CMR	Sub-Saharan Africa	Middle-Income	0.1	1.1	0.6
113	Zambia	ZMB	Sub-Saharan Africa	Low-Income	0.1	1.0	0.5
114	Mozambique	MOZ	Sub-Saharan Africa	Low-Income	0.1	1.0	0.5
115	Uganda	UGA	Sub-Saharan Africa	Low-Income	0.4	0.6	0.5
116	Madagascar	MDG	Sub-Saharan Africa	Low-Income	0.3	0.5	0.4
117	Tanzania, United Republic of	TZA	Sub-Saharan Africa	Low-Income	0.1	0.6	0.4
118	Guinea	GIN	Sub-Saharan Africa	Low-Income	0	0.7	0.3
119	Rwanda	RWA	Sub-Saharan Africa	Low-Income	0.1	0.5	0.3
120	Togo	TGO	Sub-Saharan Africa	Low-Income	0.1	0.5	0.3
121	China	CHN	East Asia & Pacific	Middle-Income	0.2	0.2	0.2
122	Burkina Faso	BFA	Sub-Saharan Africa	Low-Income	0	0.3	0.2
123	Mali	MLI	Sub-Saharan Africa	Low-Income	0	0.3	0.1
124	Malawi	MWI	Sub-Saharan Africa	Low-Income	0	0.1	0.1
125	Niger	NER	Sub-Saharan Africa	Low-Income	0	0.0	0.0