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

Tecnológico de Monterrey

Agenda

- Objetivo
- ¿Por qué Wikipedia?
- Brechas de conocimiento
- Metodología y framework
- Brechas de conocimiento y lenguaje
- Big data vs. traditional data

Note: All the main elements are based on the cited sources.Some images, examples, summaries or artifacts in this presentation could have been generated using AI.

Objetivo

- Explorar cómo los datos de **Wikipedia** pueden utilizarse para **medir las brechas del conocimiento** (knowledge divides) en el consumo de conocimiento
  - Medir las brechas en el **consumo de conocimiento** por nivel de ingresos, países y regiones
  - Analizar las **brechas** del conocimiento **por idioma**
  - Evaluar la calidad, limitaciones y oportunidades de big data
  - Implicaciones para políticas de desarrollo

Fuente:(Wikimedia, 2019c; 2019,d; 2019;e)

Tecnológico de Monterrey

¿Por qué Wikipedia?

- Wikipedia: Enciclopedia **colaborativa** digital
  - Permite producir y consumir conocimiento libremente
- Alcance global **plurilingüe**
  - Disponible en más de 300 idiomas ("Wikipedias")
- Acceso **mundial** con excepciones
  - Bloqueada en China, Turquía…
- Modelo de conocimiento **abierto**
  - Producción y consumo gratuito de información digital

Fuente:(Wikimedia, 2019c; 2019,d; 2019;e)

¿Por qué Wikipedia?

- **Este trabajo no evalúa la calidad de los artículos**
- La comunidad (editores, bots) y el modelo han asegurado la calidad

Sí, en el diseño y entrenamiento de modelos de lenguaje como ChatGPT **sí se utilizó contenido de Wikipedia**, específicamente de la Wikipedia en inglés.

Wikipedia es una fuente clave por varias razones:

- Es de acceso libre y abierta, con una licencia compatible (CC BY-SA).
- Contiene información estructurada, bien redactada y abarca una amplia variedad de temas.
- Ayuda a los modelos a aprender lenguaje natural, hechos generales y relaciones entre conceptos.

Fuente: (Wikimedia, 2019c; 2019,d; 2019;e)

Conocimiento

- Conocimiento: entendimiento sobre datos, información, hechos y habilidades
- Proceso de adquisición: aprendizaje o descubrimiento activo
- Relación bidireccional entre tecnologías digitales (TD) y conocimiento:
  - El conocimiento es esencial para usar tecnologías digitales
  - Facilitan la adquisición y el intercambio de conocimiento

Fuente: (Castells, 2009; Russell, 2009; UNESCO, 2010; UNESCO/IFAP, 2016)

Tecnológico de Monterrey

Brechas de conocimiento

- Disparidades globales en producción y consumo de conocimiento entre países
- Acceso desigual a TIC combinado con diferencias en educación, alfabetización y capital humano
- Países conectados participan en economía del conocimiento; otros obtienen beneficios limitados
- La interrelación TIC-conocimiento intensifica las brechas digitales preexistentes

Fuente: (Mishra, 2015, Heeks, 2017; EC, 2019; UNCTAD, 2019; WB, 2019c)

Metodología y framework

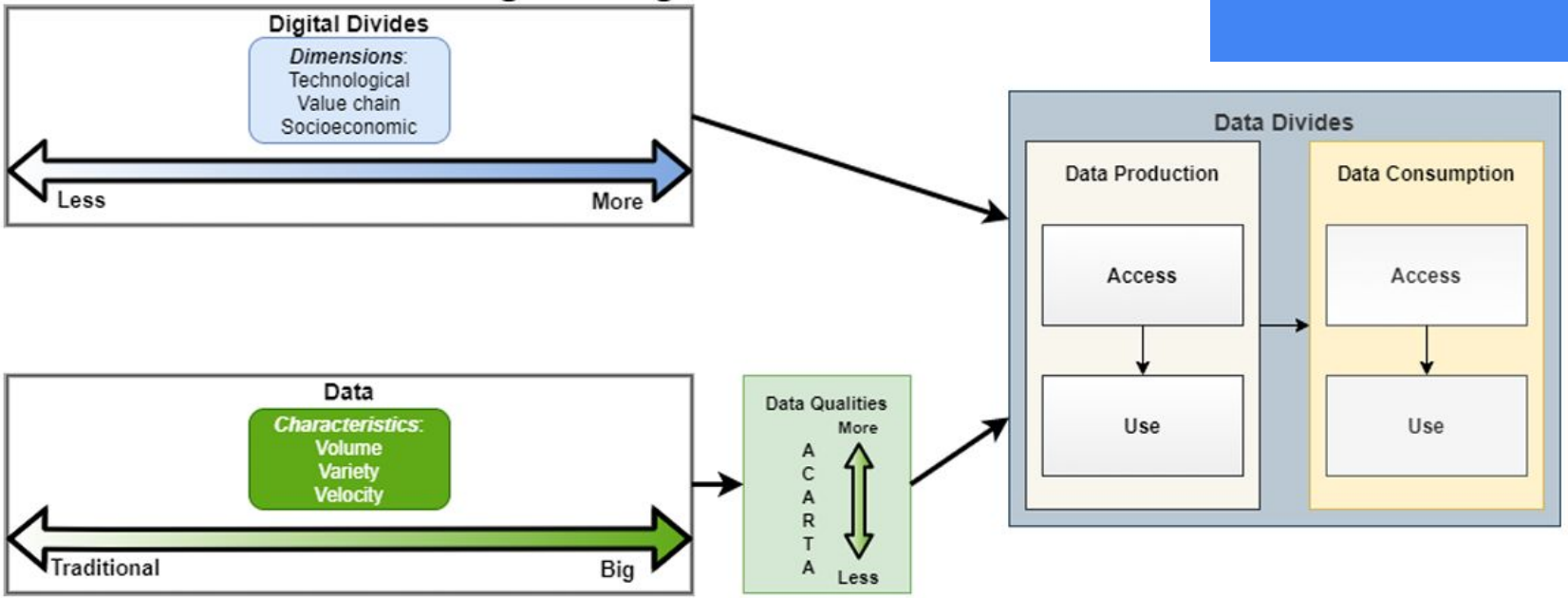
-560 mil millones de registros

-247 Wikipedias

-249 países

-Años: 2016-2018

Figure 1: Digital Divides-Data Framework



Fuente: (Rivera-Illingworth et al., 2025)

Tecnológico de Monterrey

Wikipedias por familia de lenguaje

Family	Wiki Size Category 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

-25 familias idiomáticas

Categorías:

A. 1 millón o más

B. 250K a menos 1M

C. 100K a 249K

D. 50K a 99K

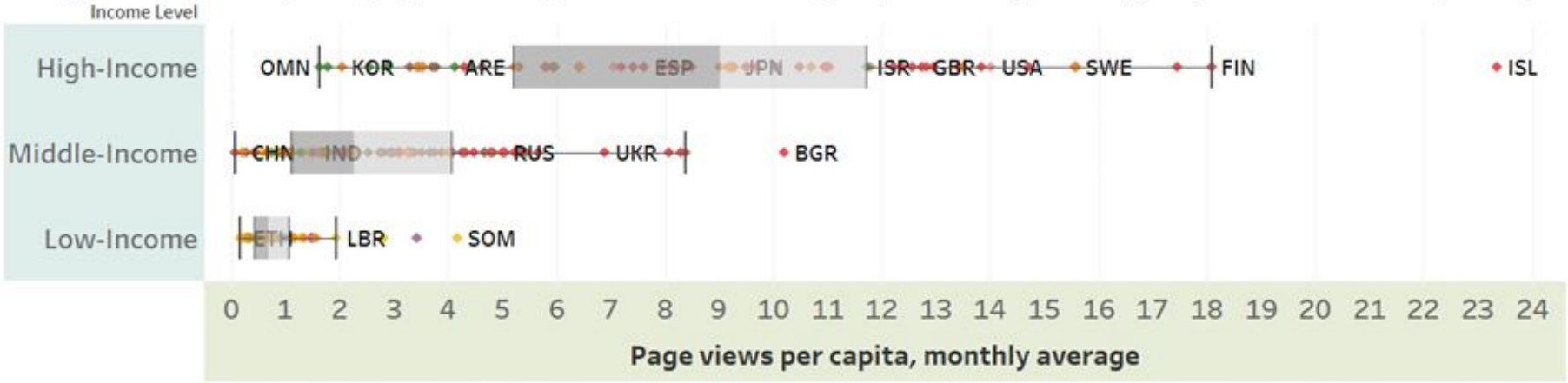
E.10K a 49K

F. 1k a 9K

Fuente: (Wikipedia, 2019; Ethnologue, 2019)

Tecnológico de Monterrey

Páginas vistas per cápita conectada (promedios mensuales por nivel de ingreso), 2018



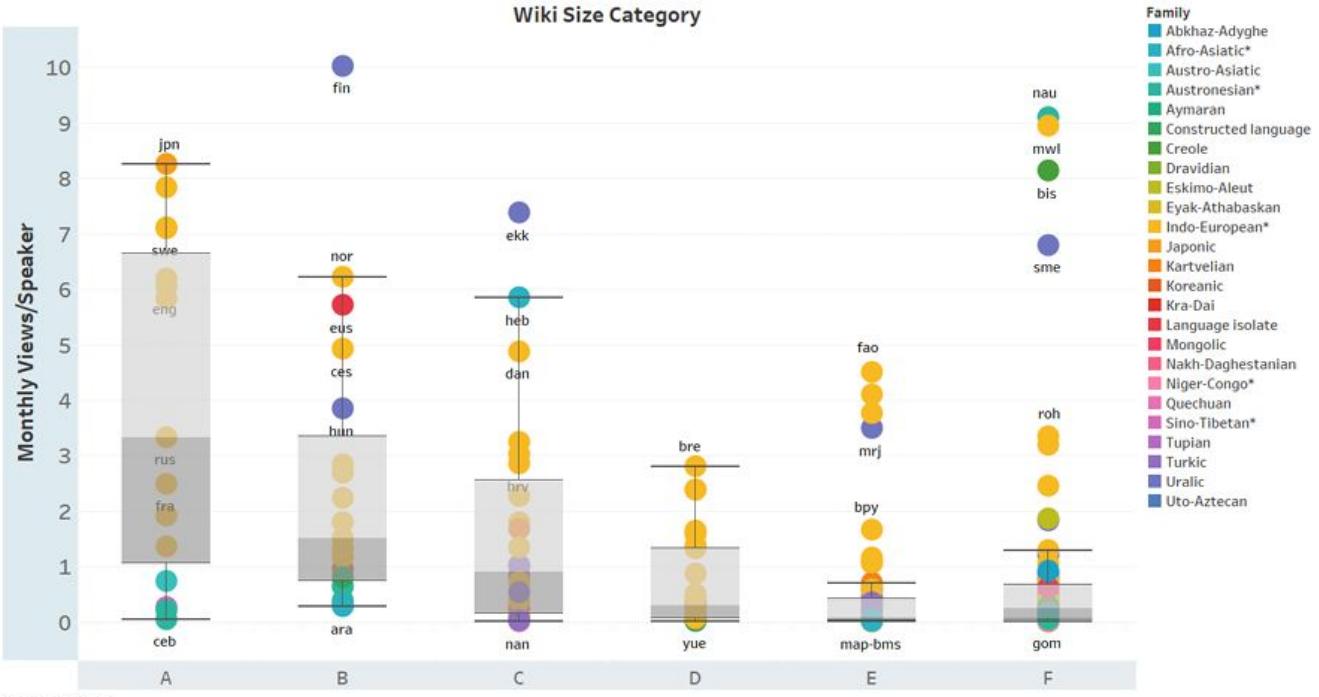
Region WB
- East Asia & Pacific
- Europe & Central Asia
- Latin America & Caribbean
- Middle East & North Africa
- North America
- South Asia
- Sub-Saharan Africa

205 países
Kruskal-Wallis ($H=124.525$, $p<0.001$)

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.

Tecnológico de Monterrey

Páginas vistas por hablante del idioma, por tamaño de Wikipedia, 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

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

CDD

Characteristics	Fortalezas	Limitaciones
Availability	-Todos los países incluidos -~250 idiomas incluyendo todos los principales -Más económico de producir	-No todos los idiomas pequeños representados
Completeness	-Proporciona dimensiones adicionales	-Falta de dimensiones tradicionales: edad, género, urbano/rural
Accuracy	-Mayor precisión en la generación de datos -Metodológicamente homogéneo	-Sesgo potencial: conocimiento disponible y perfiles de usuarios -Competidores, barreras
Relevance	-Relevante para el consumo en esas subdimensiones medidas	-Limitado a cierta noción de conocimiento (Occidental)
Timeliness	-Tiempo real/diario -Archivos analíticos mensuales	
Accessibility	-Los datos en bruto son abiertos y gratuitos -Herramientas disponibles para acceder y analizar datos -Permite replicabilidad -Fuente estable	-Alto poder computacional y capacidades de ciencia de datos -Sin documentación -Algunos datos georreferenciados son limitados

Tecnológico de Monterrey

MANCHESTER 1824
The University of Manchester

CDD

Rivera-Illingworth, Alfonso and Heeks, Richard and Renken, Jaco, Knowledge Divides in the Era of Big Data: Using Wikipedia Data to Identify and Measure Divides across Countries and Languages (February 26, 2025). Available at SSRN: https://ssrn.com/abstract=5157995 or http://dx.doi.org/10.2139/ssrn.5157995

larivera@tec.mx