The AI revolution will reinforce regional disadvantages unless steps are taken to close the data divide

Michael Kende

July 2024

ChatGPT was unveiled to the world less than 2 years ago as the first publicly available generative AI (GenAI) large language model (LLM). It broke records as the first online application to reach 100 million users in just two months. The record speed of adoption was matched by the speed with which policy makers began to grapple with the implications. However, lost in the shuffle is the fact that the benefits of GenAI are not flowing equally to everyone, everywhere, due to a significant data divide, which results from the persistent digital divide.

Data is at the heart of all AI systems

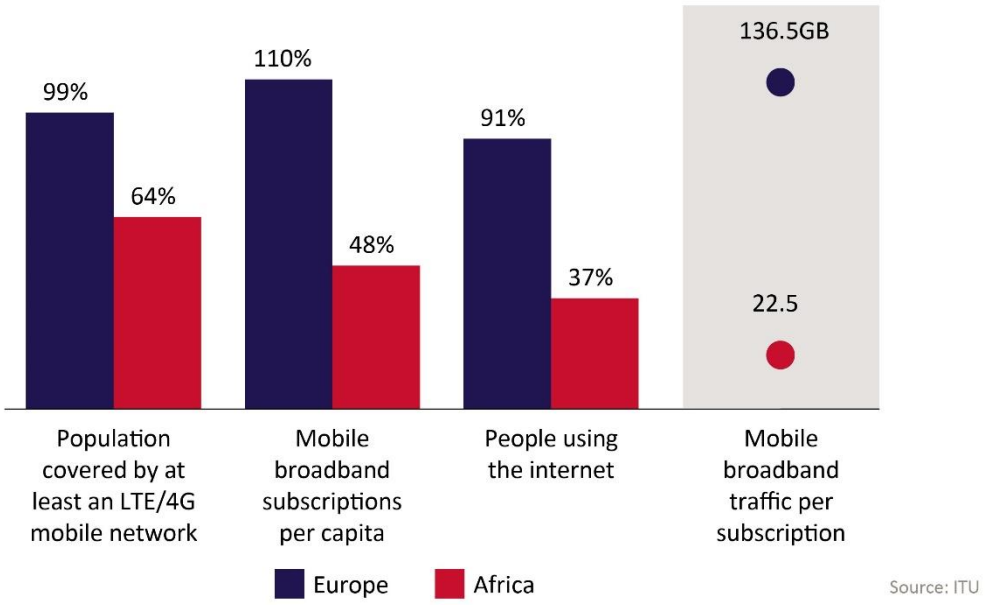
Data is the foundation of the digital age, and the amount of data is growing exponentially. Every online action and interaction we take leaves a trail of digital breadcrumbs underlying our personal and professional lives. In much of the world, policy is focused on containment – limiting how data is used, where it is sent and keeping it out of the hands of hackers. While responsible limits on the use of data are critical, effort should be devoted to generating useful data in areas where it is in short supply.

According to the Organisation for Economic Co-operation and Development (OECD) definition, an AI system "infers, from the inputs it receives, how to generate outputs such as predictions, content, recommendations, or decisions." GenAI models have been trained on data from the internet to generate outputs and to answer our queries. The quantity and quality of the data determine the quality of the output. The problem of the 'data divide' lies in the fact that the internet data used to train the GenAI models is not fully representative of the entire global population, which may lead to no answers, incomplete answers or false answers to queries.

Online access and the ability to access generate content are far from universal

The data divide arises from a waterfall of gaps. In many countries there is still a connectivity gap, as not everyone has access to broadband internet service. Among those with access to the internet, not all of them have adopted its use, and among those online, usage is lower in lower income regions. As shown in the graph below, in Europe access to 4G mobile broadband is practically universal, and almost everyone subscribes to and uses the internet. In Africa, only 64% of the population is covered by 4G, and less than half have an internet subscription. Among those who are online, not everyone has access to a computer or smart device that is capable of producing data, rather than just consuming it. These factors combine to create a level of traffic per subscriber in Europe that is six times higher than in Africa.

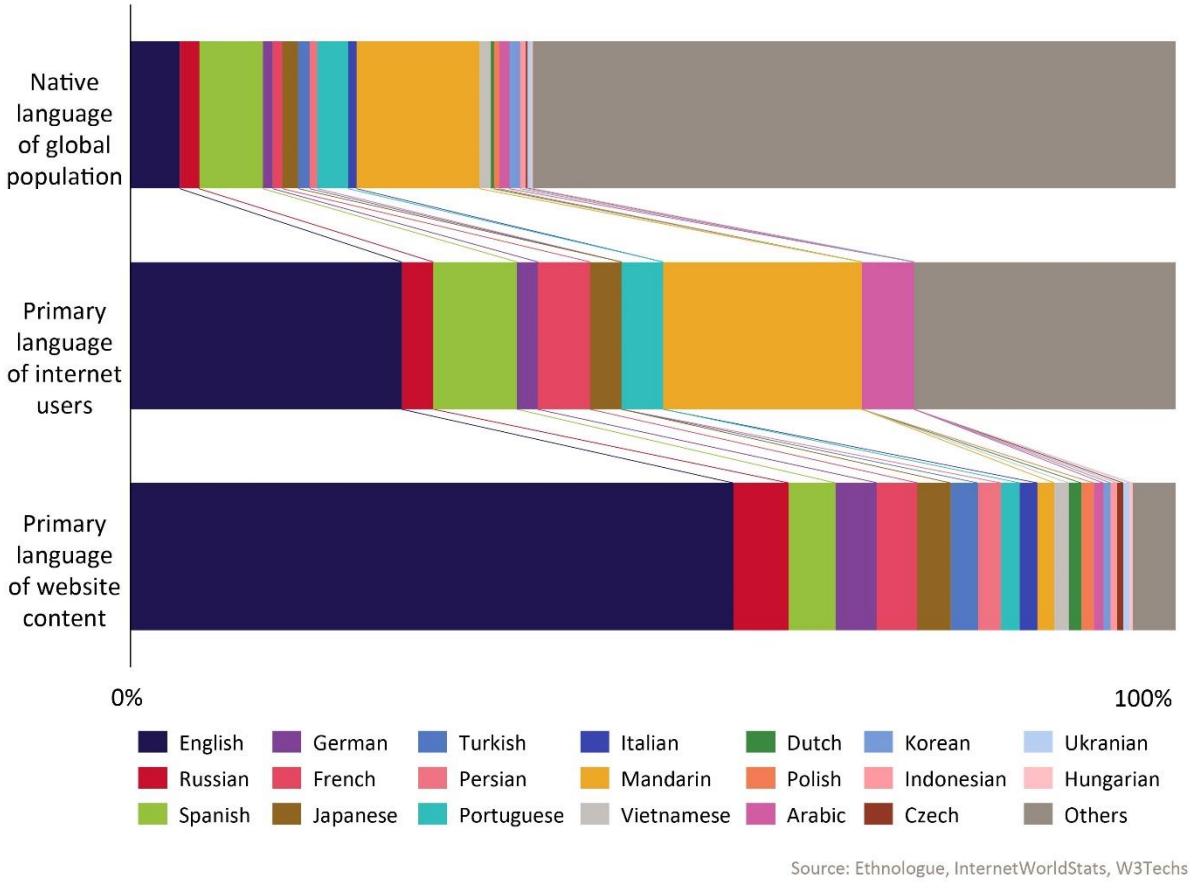
Figure 1: Internet statistics and data traffic comparison between Europe and Africa



As a result, there are fewer internet users in lower income regions, and each user is generating less data; in addition, there are far fewer fixed broadband connections in lower income regions, further reducing the amount of data that is being generated. The data being generated includes local content upon which AI could be trained.

Using language as a proxy for local content, the language of websites is heavily skewed. As shown below, the number of websites in English far exceeds the number of English speakers on the internet, which in turn exceeds those speaking English as a native language, while the other 19 languages listed are far behind in website languages. While the multitude of languages other than the 20 listed here representing the bulk of native speakers, they represent a small minority of websites. The result is far less web content on which to train GenAI models, and it shows. According to one observer, for instance, ChatGPT can understand some African languages, but its answers are ‘very poor’.

Figure 2: Proportion of internet users, websites and native language speakers for the top 20 website content languages



Source: Ethnologue, InternetWorldStats, W3Techs

Thus, the benefits of GenAI, which are currently being enjoyed by millions around the world, are limited for those choosing or needing to use their local languages. The same is true for voice assistants such as Siri and Alexa, which also train on text and speech, and (so far) are available only in a relative handful of languages. Further, as GenAI models begin to underpin medical applications, robots, digital agents and other services, the impact of the gap spreads accordingly.

Expanding the use of such AI systems, including their ability to support and operate in additional languages, will provide significant socio-economic benefits and help realise the UN sustainable development goals (SDGs).

The lack of data constrains the value of AI systems

The digital divide – the gap in access to ICT and the internet in particular – has persisted in lower-income countries since the dawn of the information age. Global efforts are helping to close this divide, but there is a risk that the data divide will persist for years. Failure to close the data divide will result in only partial realisation of the full potential of AI systems including, but not limited to, GenAI models.

Of course, countries can take direct action to facilitate training LLMs by digitising content and making it available to providers, but this will not provide the full riches of content that would be available with increased general internet adoption and usage by individuals and organisations.

To fully close the data divide, policy makers should first focus on closing the digital divide, from which the data access stems – closing it not just for access, but also for usage. This requires several steps:

- **Close the coverage gap** by lowering the cost of deploying networks, for instance by allowing the sharing of infrastructure (including towers), making suitable spectrum available and facilitating the use of new technologies such as low Earth orbit satellite constellations.

- **Increase adoption** of broadband by lowering the cost of access and smart devices, for instance through lower taxes or 'pay as you go' programmes to offer smartphone financing over a period of months or years.

- **Increase usage** of broadband by lowering the cost of internet data and increasing the attractiveness and availability of relevant content, including e-government services.

Closing the digital divide is necessary, but not sufficient, to close the data divide, which also requires policy to create – and fill – data infrastructure. There are again several steps that will help to achieve this:

- **Develop data infrastructure** including data centres and cloud companies, with favourable licensing and reliable and affordable access to energy (preferably renewable), to lower the cost and latency of accessing content.

- **Protect personal data** in domestic data centres with privacy and data protection policies, and enable the free flow of data with trust, to enable data to be suitably shared and aggregated for further benefits.

- **Promote digital skills** with training and capacity building to spur the development of websites and content and limit the liability of intermediaries hosting the content of others.

The urgency to close the digital divide should not just focus on the need for digital infrastructure and benefits from increased internet adoption, but on the need to close the digital divide as a step towards closing the data divide, to ensure that the benefits of AI are realised by and for all.

About us

For almost 40 years, Analysys Mason has helped shape the evolution of regulation in the technology, media and telecoms industries. We have helped explore, measure and understand emerging problems, and have created innovative and effective solutions. To find out more about how Analysys Mason can help you, please contact Michael Kende (Senior Adviser) or James Allen (Partner).