Statistical Commission
Fifty-sixth session
New York, 4 - 7 March 2025

Room document
Available in English only

Fostering AI-Readiness and Responsible Redistribution of Official Statistics

Room document submitted by the World Bank

Fostering AI-Readiness and Responsible Redistribution of Official Statistics

Room Document by the World Bank

Introduction

The Fundamental Principles of Official Statistics (FPOS) highlight the critical role of high-quality statistical information in promoting sustainable development, peace, security, and international cooperation. Principle 1 of the FPOS states: *"Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy, and the public with data about the economic, demographic, social, and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information."*[1]

To fulfill this role in an interconnected world that demands openness and transparency, many national statistical agencies and international organizations publish data and statistics under open or permissive licenses to maximize their use and value. But in a data ecosystem characterized by diverse data sources and rapidly evolving technology, official statistics face three challenges:

1. **Multiplicity of data sources available to users.** Data from the private sector, citizen-generated data, and other sources provide useful inputs for producing official statistics; however, they also introduce competitive dynamics. While non-official data may lack the guarantees of statistical robustness, they often attract users due to their discoverability and ease of access and use. To increase the visibility and utilization of official data, statistical agencies must invest not only in producing high-quality data but also in ensuring that these data are easily discoverable, accessible, interpretable, and user-friendly.
2. **AI applications currently possess limited capabilities, and the AI-readiness of data presents a significant challenge.** Official data is frequently redistributed by third parties through AI-enabled search engines and other generative artificial intelligence (AI) applications, such as chatbots. These applications exhibit limited capability in directly interacting with structured data and statistics published by statistical organizations. When official data is utilized by AI applications, it is often by exploiting secondary sources (official data reported in publications, news, or other documents and websites). As a result, statistics provided by AI platforms may lack accuracy and are not guaranteed to represent the most recent available estimates. There exists potential for enhanced interaction between AI systems and databases, but this will require AI-ready data.
3. **Limited control over redistribution by third parties.** Open data licenses allow third parties to redistribute and transform data and do not require the publication of detailed metadata. This

[1] https://unstats.un.org/fpos/

practice may result in variations in how official data is represented and could impact the integrity of the official statistics being redistributed.

To address these issues, official data producers should:

1. **Modernize data curation practices to ensure that official data is AI-ready**. This involves making both data and metadata machine-readable and machine-understandable, suitable for automated processing and analysis. Solutions involve adopting metadata standards and best practices of data curation, using data exchange standards such as SDMX, systematically employing APIs for data and metadata sharing, and utilizing technology for data and metadata quality control and optimization. This goal necessitates a committed and coordinated program. During its meeting on 10 October 2024, the Committee for the Coordination of Statistical Activities (CCSA) agreed to "*jointly promote the adoption of a common set of metadata standards by official data producers and contribute to their implementation in a coordinated manner to foster AI-readiness of official data.*"
2. **Facilitate and advocate for the responsible redistribution of official data.** This involves clearly defining the requirements and expectations for third-party data redistribution, providing comprehensive guidance, and working collaboratively with third parties.

Making official statistics AI-ready

Many data users do not directly access the websites of data producers. Instead, they utilize search engines or rely on AI systems such as ChatGPT. These tools function as essential intermediaries between data producers and consumers. However, they often fail to adhere to best practices concerning clarity and interpretability.

The evolution of internet search engines

Traditionally, internet search engines have served users by providing relevant links to their queries. Although the exact ranking criteria are not disclosed and may not always highlight the most authoritative sources, users seeking information had a variety of links to choose from. With technological advancements and Artificial Intelligence (AI), the manner in which data users locate and access information is evolving. Search engines now endeavor to provide direct answers, rendering ranked links secondary. Ideally, these systems should identify and exploit the most reliable sources, and verify their agreement when multiple sources are used or highlight inconsistencies; however, this is not occurring. While factors such as data timeliness, reliability, and credibility may influence the selection of resources to be used by the generative AI process and recommended to data users, technical and commercial considerations also play a role in search algorithms. This affects the relevance and accuracy of information provided by internet search engines, as shown below in this document.

The use of generative AI

Generative AI applications such as chatbots face challenges in selecting sources and generating responses to user queries. They do not interact with live databases, which means they may not access the most recent data. When official data is cited, it is often derived from secondary sources, potentially leading to

alterations and a loss of interpretability. This can result in improper attributions and other issues; the AI-generated content may not always be factual and can include false or controversial interpretations.

The main concerns about internet search engines and AI chatbots operating as re-distributors of official data are as follows:

- While there have been improvements, AI-generated answers still frequently contain hallucinations. This issue is due to AI systems' limited ability to interact with structured data and the lack of available metadata accompanying this data.
- AI systems exercise non-transparent discretion when selecting information sources for generating responses to users' queries. This lack of transparency can lead to questions about the reliability and accuracy of the information provided. Commercial interests may affect the informational accuracy of search engine results. Personalized search algorithms may also contribute to biases in data consumption.
- It is often assumed that users possess both the capability and willingness to fact-check AI-generated responses. However, this task can be labor-intensive, requiring statistical literacy.
- Data producers miss out on valuable insights about their data's use and users through web analytics. As redistributors capture demand, statistics on visits, visitors, and downloads from producers' systems may not accurately show true product and service demand.

We demonstrate some of these issues by submitting the following query to Google search, Perplexity.ai, and ChatGPT (OpenAI): "*What is the youth unemployment rate of Indonesia and Thailand?*" We also extract the information from the World Bank database (citing ILO as the source), for comparison purpose.

Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate) - Indonesia, Thailand

Indonesia (2023)	Thailand (2023)	Thailand (2022)
13.1%	*4.5%*	*5.2%*

ILOStat data as published in World Bank Open Data platform, as of 28 February 2025

The answers provided by **Google** demonstrate a high degree of unpredictability or inconsistency in AI-generated answers. The same query, submitted 3 times in a short interval of time (< 15 minutes), provided three different answers (Figures 1 to 3). The data are provided without information on the year of the estimates (only mentioning "recent data"). The data for Indonesia in Figure 1 attributes the estimate of 16.3% to the World Bank, but this value does not correspond to a value published in the World Bank online database.

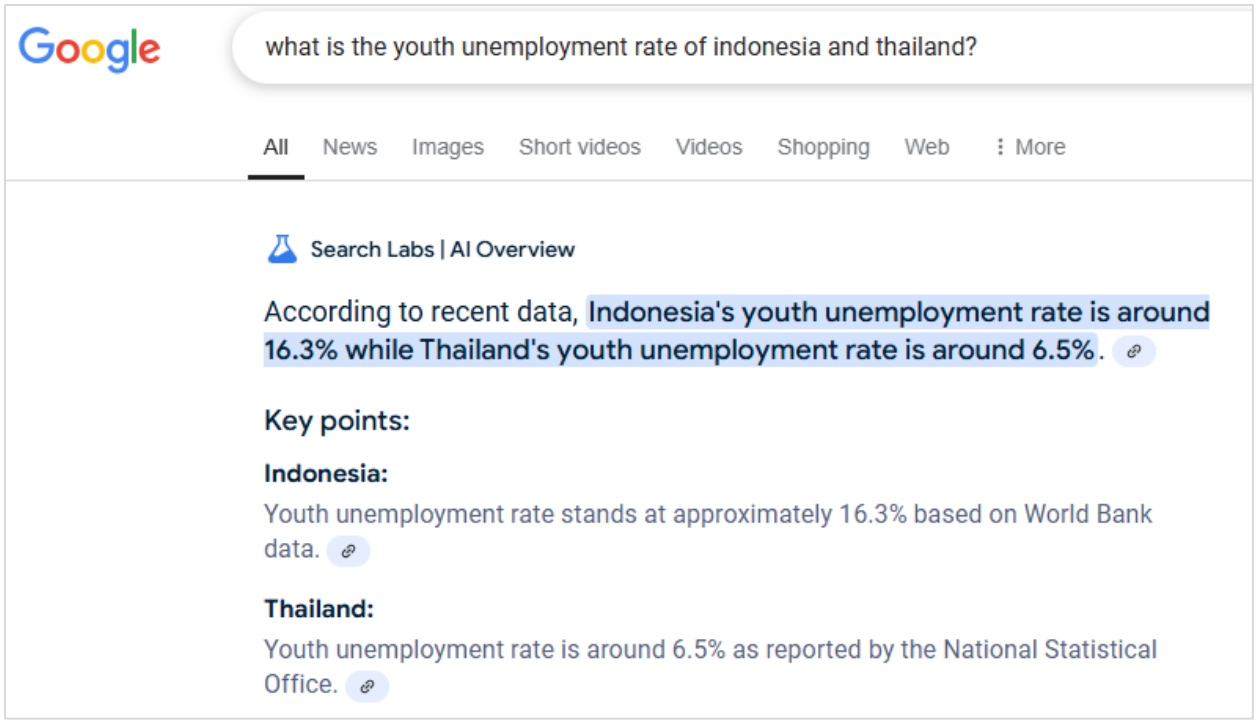

Figure 1 - Query "What is the youth unemployment rate of Indonesia and Thailand?" (as of 28 February 2025)

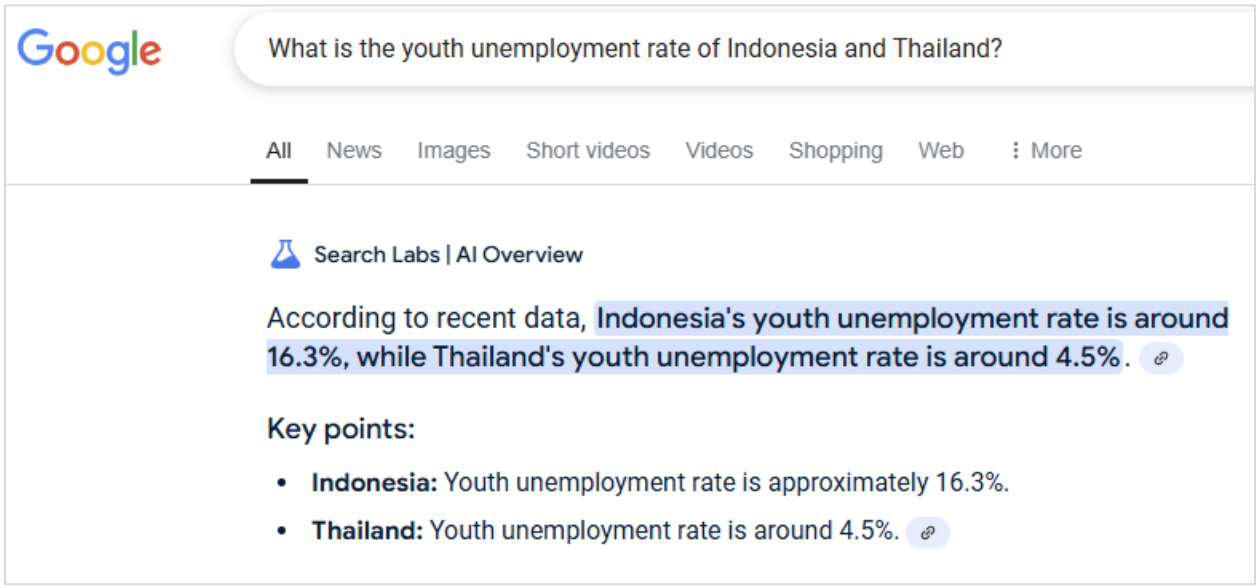

Figure 2 – Refreshed query "What is the youth unemployment rate of Indonesia and Thailand?" (as of 28 February 2025)

Figure 3 – Refreshed query "What is the youth unemployment rate of Indonesia and Thailand?" (as of 28 February 2025)

ChatGPT 4o provided data that also differ from values from the World Bank's and Google (Figure 4). It warns users that "*ChatGPT can make mistakes*" and that they should "*check important info*". The process of checking information is however challenging, considering how multiple sources provide inconsistent information with limited metadata.

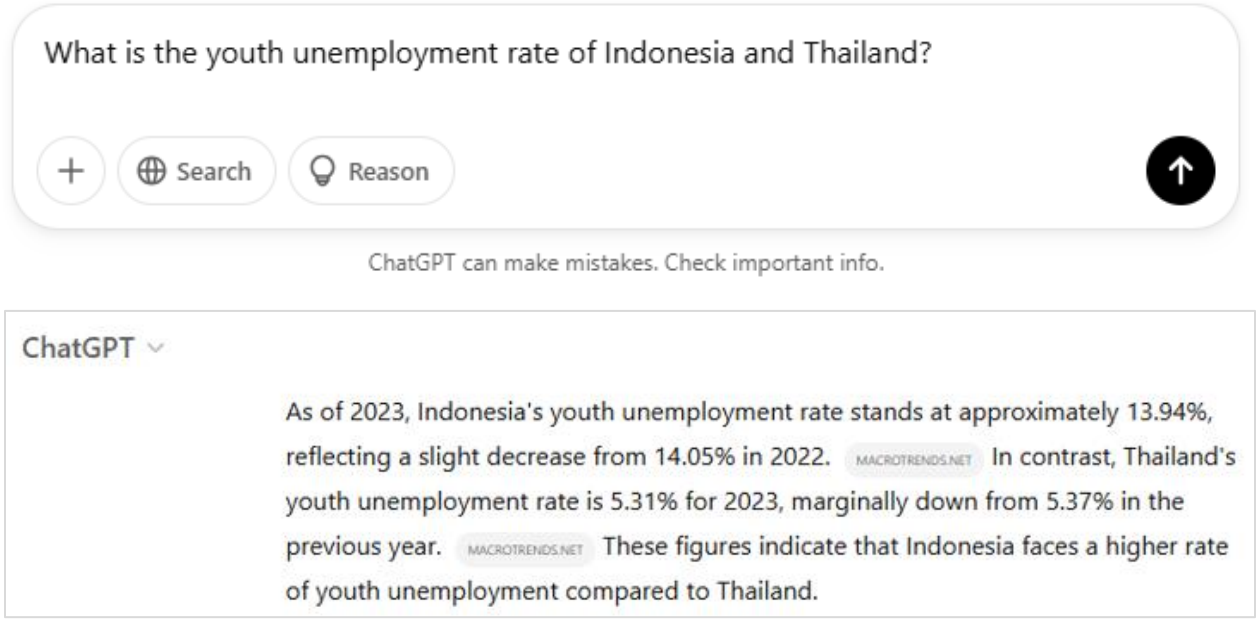

Figure 4 – ChatGPT answer to query "What is the youth unemployment rate of Indonesia and Thailand?" (as of 28 February 2025)

Perplexity.ai provided an additional set of estimates (Figure 5), with a value for Thailand that does not correspond to values in other sources. The footnote attached to the information provided on Thailand provides a link to the source of the data, a newspaper article.

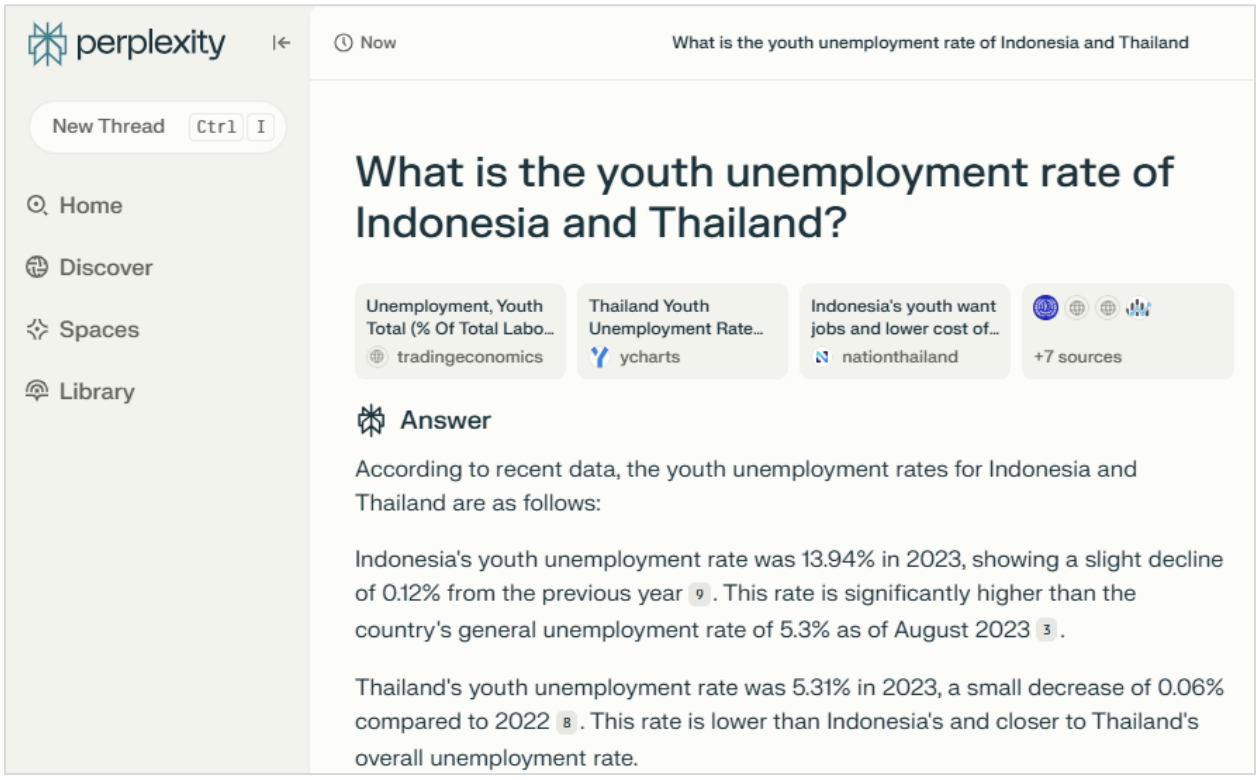

Figure 5 - Perplexity.ai answer to query "What is the youth unemployment rate of Indonesia and Thailand?" (as of 28 February 2025)

The examples given above highlight the necessity for official statistical agencies to take measures ensuring that their estimates, along with those from other reliable sources, are utilized by AI-enabled search engines and chatbots.

Applications utilizing generative AI include disclaimers and source citations to mitigate issues of hallucination and inaccuracies. However, this method may prove ineffective if users, as anticipated, do not engage in fact-checking. Additionally, generative AI-enabled search is frequently presented as a tool for swiftly answering data-related queries and making information accessible to a broader, less technical audience. This creates an inherent contradiction wherein users are expected to verify the accuracy of the information—an endeavor often requiring technical expertise—while simultaneously being presumed to lack such skills, which justifies the need for AI. The responsibility for ensuring the quality of the information reported should be attributed to the developers of these applications rather than relying on the users' ability to validate the outputs. For users seeking authoritative sources, being provided with the most relevant URLs may be more beneficial than AI-generated answers.

Official data producers can address these issues by ensuring their data and metadata are compatible with AI systems and by supporting the improvement of AI-enabled data dissemination systems. The official

statistics community should engage in this process by collaborating with the private sector. We suggest the following two actions accordingly.

Action 1: Standards and tools to modernize data curation and dissemination

Action 1 prioritizes the advancement and implementation of global public goods, with the following priorities.

- **Adopt common metadata standards**: The more systematic adoption and implementation of common metadata standards by official data producers would ensure consistent documentation of indicators, microdata, geographic datasets, tables, documents, and scripts. Open-source tools must be created and published to facilitate the production of standard-compliant metadata and support the optimization of metadata using AI. This approach will enhance the completeness and usability of metadata, improve data management platforms' interoperability, and foster generative AI's ability to interact with structured data.
- **Adopt common data exchange standards**: Supporting the implementation of SDMX can boost efficiency in transferring data and metadata across platforms with open APIs. This will facilitate the automation of data exchange, reduce the burden on statistical agencies, and enable efficient data acquisition and redistribution by third parties, including AI-powered systems.
- **Promote open data licenses**: Promoting the adoption of standard open data licenses, such as CC BY 4.0, will help ensure the responsible use of data.
- **Enhance data discoverability tools**: Developing advanced open-source data discoverability tools, including semantic search and recommendation systems that can be applied to various data sources and types, will enhance the discoverability and visibility of official data published by statistical agencies.
- **Maintain a central metadata catalog of accredited sources**: A central metadata catalog of accredited sources could significantly benefit data users by distinguishing authoritative and reliable data sources from others. This distinction would help identify fit-for-purpose data, enable search engines to rank results more appropriately, and assist generative AI applications in formulating more relevant answers. Establishing such a catalog as a global public good would necessitate a governance mechanism, accreditation criteria for sources, and an efficient system for metadata sharing and discoverability. This would be facilitated by adopting common metadata standards and API-based systems like SDMX.
- **Build the capability of large language models (LLMs) to interact with structured data**: The examples provided above highlight the current challenges LLMs face in effectively interacting with relevant structured data sources. Developing AI systems that can interact with databases via SQL, SDMX, or other structured queries can address this issue. One example of such a solution is the IMF's StatGPT application. Implementing this solution involves adopting SDMX and metadata standards. Tools like the World Bank Metadata Editor and dissemination platforms such as the

African Development Bank's Open Data Platform or the OECD's DotStat suite[2] complement StatGPT.

- **Develop AI tools for data quality and enhancement**: Generative AI is employed to retrieve and interpret data and statistics. Ensuring proper quality control and augmenting the metadata associated with the data can enhance its interpretation accuracy, whether by humans or large language models (LLMs). AI solutions can aid in detecting anomalies and classifying them as errors or "unusual but true" values (e.g., sudden variations in time series). Figures 7 and 8 illustrate examples of such anomalies. Adding explanations for true anomalies to the datasets (in the form of cell-level or series-level annotations) would provide data users and LLMs with relevant information to interpret the data accurately.

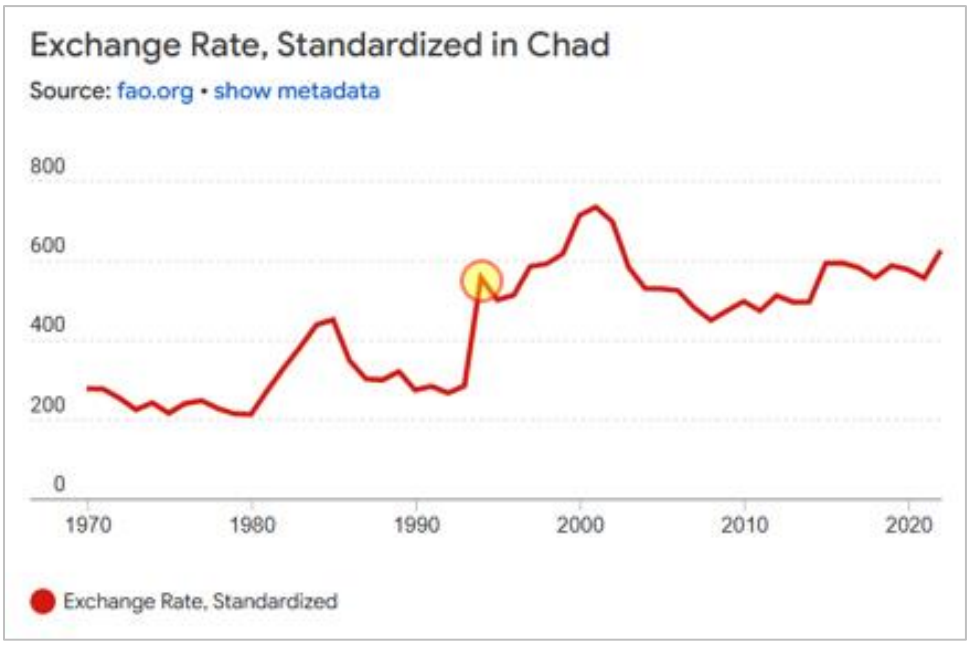

Figure 7 – Exchange rate in Chad; 1994 peak corresponds to 50% devaluation of CFA (source: FAO, reported in Google DataCommons, as of 15 December 2024)

[2] See https://siscc.org/stat-suite/

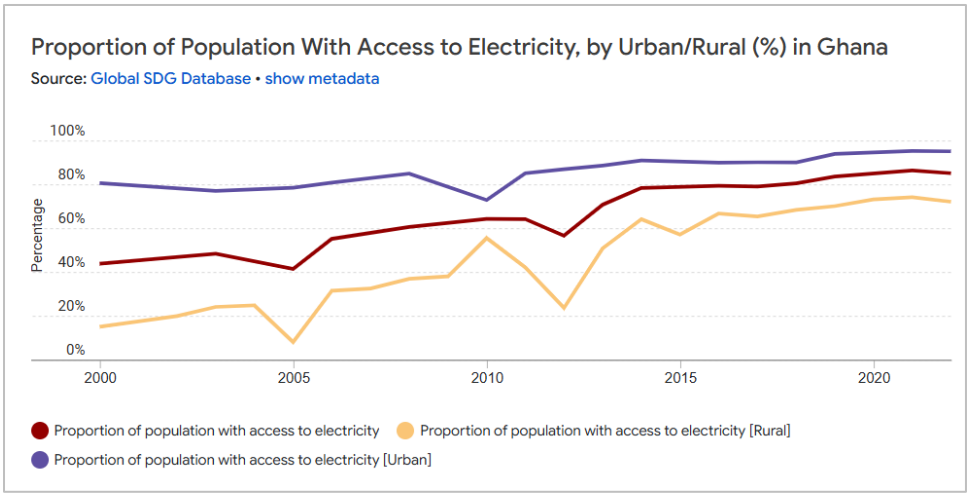

Figure 8 – Access to electricity in Ghana: unexplained reason for 2010 decrease in urban areas / increase in rural area, followed by 2012 major decline in rural areas / unchanged rate in urban area (source: UN SDG Database, reported in Google DataCommons, as of 15 December 2024)

Action 2: Support data producers in low and middle-income countries

Action 2 involves designing and implementing a coordinated program of technical support for official data producers in low and middle-income countries. The objective is to aid the adoption of common metadata standards and best practices in data documentation, implement API-based data dissemination, and improve the discoverability of their data dissemination systems.

Fostering responsible redistribution of official data

Government agencies and international organizations increasingly release data under open licenses like the Creative Commons Attribution 4.0 International (CC-BY 4.0)[3] or open government licenses. These permissive licenses typically allow third parties to copy, modify, and re-publish data for any purpose, often but not always including commercial use[4], as long as they adhere to certain conditions:

- Attribution: Properly credit the data provider and its data providers.
- No endorsement: Do not imply that the data provider sponsors or endorses dataset use.
- No association: Do not use the data provider's name or logos without consent. No implied affiliation is allowed.
- No warranties: The data provider can modify or discontinue datasets at its discretion, with or without notice.
- Exclusion of liability: The data provider may develop competing products or services using the datasets without liability.

[3] See https://creativecommons.org/licenses/by/4.0/deed.en

[4] Prohibiting commercial use could pose a legal obstacle to data redistribution by certain third parties, potentially reducing the visibility of the data on key platforms.

The conditions tied to open data licenses can be challenging to enforce and may offer limited protection against improper redistribution. This can lead to data being redistributed without adhering to the original quality and transparency standards, potentially resulting in misleading content. These issues in current implementation are illustrated in Figures 9 and 10.

Figure 9 illustrates the "Gini Index of Economic Activity of a Population in Nigeria," as reported by Google DataCommons, [5] citing the World Bank as the source. It is important to note that the World Bank catalog does not include an indicator under this precise label; it uses the term "Gini Index" instead. Any alteration in the designation of an indicator constitutes a significant modification, which, in accordance with the CC BY 4.0 license, must be adequately documented.

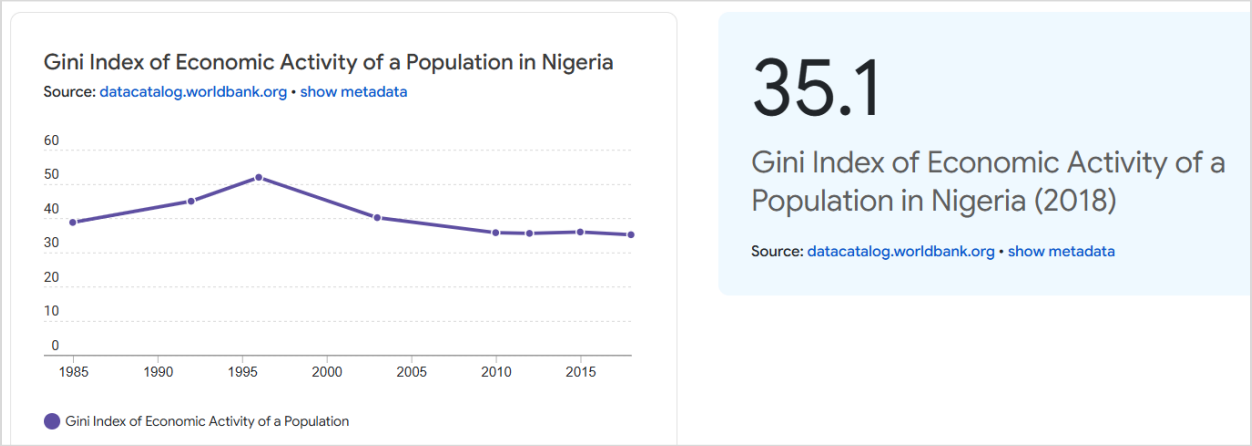

Figure 9 - World Bank Gini Index as reported by Google DataCommons (as of 15 December 2025)

Figure 10 is derived from the Worldometer data platform. It presents population information as of the query date, noting that the estimates are "based on the latest United Nations estimates." While this statement is not incorrect, it may lead some users to incorrectly attribute these estimates directly to the United Nations Population Division. In reality, the estimates are projections generated by Worldometer, which utilizes United Nations data as an input source.

Figure 10 is sourced from the Worldometer data platform, [6] after submitting a query for "*Urban Rural population of West Africa*". It provides population information as of the query date, specifying that the estimates are "*based on the latest United Nations estimates*." While this statement is not fully inaccurate, it may lead some users to mistakenly attribute the reported population estimates to the United Nations. In fact, the estimates are projections created by Worldometer, which uses data from the United Nations Population Division as one of its input sources.

[5] From: https://datacommons.org/explore#q=gini+nigeria&client=ui_query
[6] https://www.worldometers.info/world-population/western-africa-population/

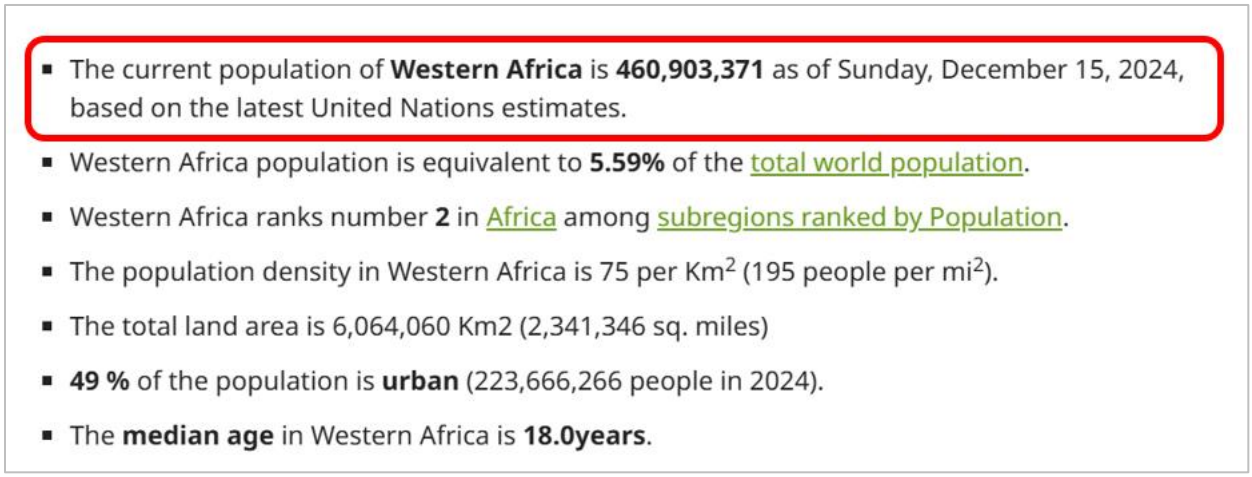

Figure 10 – Worldometer population estimates with reference to "latest United Nations estimates" (as of 15 December 2024)

Given the benefits of open data distribution, it is advisable to maintain or enhance open data practices while implementing measures to address risks and ensure data integrity. Actions 1 and 2 outlined earlier aim to address these challenges by providing redistributors with more convenient methods for acquiring and re-packaging open data and related metadata. Additionally, third parties could be encouraged to adopt improved data dissemination practices to ensure clarity and prevent misuse. Republished data should clearly present statistics, include relevant metadata, and inform users of any modifications transparently. The official data community could assist third parties in adhering to such best practices. The following action is proposed.

Action 3: Establish guidance for data redistribution, and collaborating with redistributors

The third proposal entails formalizing a collaborative initiative between the official statistics community and major data redistributors and search engines. The primary objective is to ensure adherence to open data usage terms and best practices, as well as to implement improved practices in data dissemination. This initiative would require:

- **Develop technical guidance for redistributors.** The guidelines would outline the expectations of official data producers and suggests practical actions for their implementation. At its meeting of 10 October 2024, the CCSA members agreed to "*jointly finalize a document defining core principles and recommended practice for the redistribution of official data by third parties*" and to "*Share and discuss the document with private sector leaders in a collaborative spirit, with the intent to encourage the adoption of statistical data dissemination good practice by third parties.*"
- **Provide systematic feedback to third parties.** Principle 4 of the FPOS states that "*statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics*." In this context, these agencies should also address misleading or sub-optimal re-publishing of their data. Official data producers should proactively offer feedback to data redistributors. Additionally, CCSA organizations should collectively encourage internet search engine companies and data

redistributors to systematically involve data specialists to assist IT experts in testing data dissemination applications.

- **Metadata mapping**. The lead developers of internet search engines have created their own metadata standards, optimized for information discovery on the internet. The main standard in use is schema.org.[7] To facilitate collaboration with these IT companies, it is recommended to map the specialized metadata standards used by the statistics community with schema.org (and possibly Croissant,[8] a new standard developed to complement schema.org and designed to facilitate the documentation of datasets intended for use by machine learning applications). This will facilitate the exchange of information and interoperability of official and third-parties data dissemination platforms.
- **Develop a mechanism for third parties to provide usage insights to data producers**. The value of data is often shown through its use. When third parties redistribute data, the usage metrics captured by official data producers through web analytics in their own dissemination systems become less representative and useful, reducing their ability to assess the relevance of their data. To address this issue, the official statistics community can collaborate with the lead data redistributors to establish mechanisms for sharing distribution insights and usage metrics.

[7] See https://schema.org/
[8] See https://research.google/blog/croissant-a-metadata-format-for-ml-ready-datasets/