Mapping education data in sub-Saharan Africa [1]

This is a preliminary draft. Please do not circulate

Overview

The Unlocking Data initiative, pioneered by ESSA, Zizi Afrique Foundation and EdTech Hub, aims to explore the ecosystem of education data in sub-Saharan Africa (SSA), to raise awareness on the importance of sharing education data and to increase access to the latter. Increasing access to data, especially in low resource settings, will help researchers generate policy-relevant empirical evidence, providing policymakers with a comprehensive picture of challenges and feasible policies.

This document presents a methodological approach to map existing education data. The approach, combining an opportunistic data search with a systematic approach and stakeholder consultations, is expected to be applied in country case studies. These case studies will help understand what exists, where there are data gaps and draw out the commonalities and contextual challenges surrounding availability and access to education data in SSA.

Introduction

In the Agenda 2030 for Sustainable Development, the United Nations adopted 17 Sustainable Development Goals (United Nations, SDGs) and 169 targets that balance the economic, social, and environmental dimensions of sustainable development. Specifically, the SGD-4 aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (United Nations, SDG-4). For the SDG on Education to be effective, numerous education-related targets and indicators need to be regularly assessed, justifying data-based research on the education system, educational attainment, and policies, especially in developing countries.

In sub-Saharan Africa (SSA), where little contribution to the world's research is generated (Blom et al., 2016), even so in the field of education, micro-level data (microdata) on education are fundamental in telling the story of education.[2] Yet, if they exist at all, such

[1] We are grateful to our partner organisations such as eBase Africa (R. Pambe), EdTechHub (T. Adam and A. Kreimeia), JET Education Services and NORRAG for their useful comments and suggestions which help improve this document.

[2] Microdata in our context refers to the *statistical term* for individual responses in surveys.

datasets are difficult to access and for researchers and policy analysts, the collection of primary data is costly and often beyond their financial resources.

As previously noted by Hallberg Adu (2014), this fact creates a vicious circle where the lack of data leads to limited relevant research, and less local research being considered for policy. Unlocking the patchwork of data that has been collected, among others, for empirical research, baseline evaluations, landscaping studies and feasibility assessments would give decision-makers a better picture of the state of education in their countries. Giving African scholars (HEIs, researchers, PhD candidates…) access to this data will increase the generation of relevant knowledge and be used to grow local capacity for analysis, enabling African researchers to produce outstanding policy-relevant studies.

Mapping the landscape of education data available in Africa and sharing the latter to increase access to and use of data are complex and ambitious tasks, and work is needed not just on assessing the ecosystem of data but also ensuring that the identified datasets satisfy certain quality criteria and can be used. In the same vein, the Webinar on Unlocking Data to Tell the Story of Education in Africa (Adam et al., 2020) raised a series of questions about, among others, the barriers to data sharing, quality and accessibility of qualitative versus quantitative data. Moreover, in a survey of education researchers based in SSA (whose research is featured in the AERD, 2020) and in recent breakout group discussions, concerns such as stakeholders' misunderstanding of the importance of sharing data, the lack of coordination in the data sphere, low research capacity and the absence of harmonised and comparable indicators are mentioned among the barriers to education research in SSA. Despite all these barriers, to make a change, we must first know where we stand.

Knowing where we stand in terms of data availability, challenges surrounding access to data and gaps, requires a methodical assessment of the landscape of education data in SSA. To the best of our knowledge, the literature on data mapping in social sciences is almost non-existent, whether as a methodological contribution or mapping existing data. Contrary to computer scientists, who largely developed methodologies for data mapping (schema mapping), there seem to be very few contributions from researchers in education, sociology, and economics relating to education data mapping. For their part, social scientists in education have devoted considerable efforts to understand the current state of knowledge on specific research questions relating to education. These efforts, among others, have led to systematic evidence reviews as observed in Pittaway and Cope (2007) on entrepreneurship education, Crompton and Burke (2018) on the use of mobile learning in higher education and Haßler et al. (2020) on TVET in SSA. Specifically on data mapping, while Masefield et al. (2020) highlight the potential of microdata collected by NGOs and private stakeholders, the only paper mapping datasets used in empirical research is the work by Koepke & Paetzold (2020).This concept note, as far as we know, is the first contribution introducing a strategy for education data mapping.

This document describes a methodological framework for education data mapping, whether it concerns primary, secondary, tertiary education or vocational training. It is a starting point in understanding the ecosystem of education microdata in SSA, its actors, and gaps, and is developed as part of the Unlocking Data initiative piloted by ESSA, Zizi Afrique, eBASE Africa, EdTechHub, Open Burkina and the Esme Kadzamira (University of Malawi). Once the data mapping strategy is deemed satisfactory, our approach is to start incrementally in a handful of selected countries in which we have partners and in specific areas where there is interest in mapping education data e.g., TVET in Kenya, EMIS data in Malawi and private school data in Burkina-Faso.[3] Based on these early pilots, lessons will be learned and shared specifically on:

- the implementation of the mapping strategy (e.g., success and limitations of the search strategy, modes of stakeholder engagement, resources required).
- the potential of private school data and data privately owned for education policy.
- the use of the outputs (e.g., do the data maps help to promote more useful research?).

The remainder of this document is organised as follows: Sections 2 and 3 present some definitions and overview some relevance and quality criteria. Section 4 describes the data mapping strategies, and Section 5 concludes.

1. From a systematic review to data mapping

1.1. Definitions

A systematic review: This is a procedure that identifies and selects existing studies, evaluates their contributions, analyses, synthesizes, and reports evidence in such a way that allows reasonably clear conclusions about what is and is not known (Buchanan & Bryman, 2009). Concerning data mapping, data scientists generally define it as the process of identifying and linking multiple datasets into a centralized database. Hence, education data mapping, in our context, involves a landscape review of data collected, among others, on primary, secondary, and tertiary education, as well as technical and vocational education and training (TVET). In so doing, rather than linking different datasets, we aim to identify and centralise education data in order to improve researchers' access to quality data and favour the generation of high-quality evidence to inform education policy in Africa.

Data mapping: Data is "information, especially facts or numbers, collected to be examined and used to help decision-making" (Cambridge Dictionary). In our context, it is educated-related qualitative or quantitative information. Several sources can be identified for data mapping as intended here. A non-exhaustive list includes, among others, i.) data from national institutes of statistics, ii.) micro-level or disaggregated data from international

[3] These specific areas and countries are where interest has been manifested so far. Lessons and challenges identified will be integrated into this document to serve as example for future data mapping exercises.

institutions (ILO, World Bank), iii.) data from research institutions, HEIs, private initiatives and NGOs, and iv.) microdata used in empirical papers on education in SSA.

Mapping the landscape of data is comparable to a systematic knowledge review in social sciences. However, contrary to systematic reviews, where researchers summarise available empirical evidence (from research papers, reports, theses) that matches pre-specified relevance criteria, mapping data focuses on the microdata collected (open or closed) and/or used in the research papers, reports and evaluations... As this contribution will also rely on the approach of a systematic review, it is necessary to briefly overview the technique of the latter.

1.2. The procedure of a systematic review

A systematic review is a research technique primarily devoted to answering research questions that have been largely discussed in empirical analyses but characterized by various and contrasting conclusions. Therefore, it identifies and gathers evidence on the specific research question, ensuring that the evidence collected matches certain criteria. Finally, it screens the evidence, extracts relevant knowledge that will be analysed and synthesised in a report. The methodological contribution introduced by this paper for mapping education microdata proposes to follow a similar procedure at its second stage.

2. Mapping education data: Relevance and quality

Mapping data implies identifying the datasets, the relevance and quality of the latter, as well as the actors involved in the data collection/production and use.

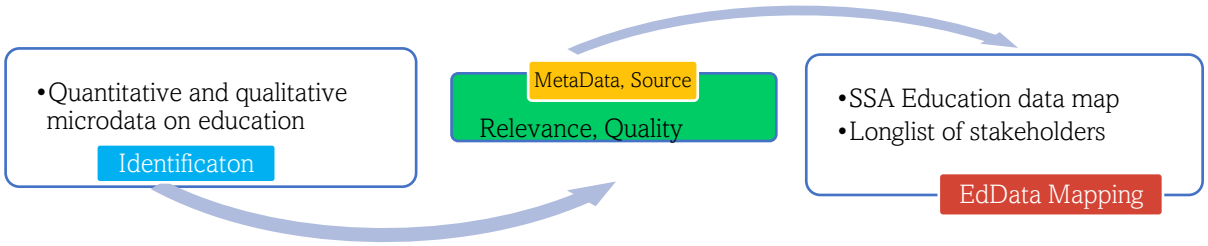

Figure 1: Assessment of education data ecosystem

1.3. Relevance criteria

Like systematic reviews, our approach assumes that the precise scope of the data mapping exercise has been clearly defined and documented by the researcher. This might be, for example, TVET data, data on public or private schools, and on primary, secondary and/or tertiary education data in a specific country. Globally, microdata on education collected at institution, region and country level will be considered, as long as it is useful to produce locally relevant knowledge.

To be considered in the present data mapping exercise, each dataset identified must fulfil certain criteria. To be relevant, the dataset must be:

- **Related to education**: Datasets are mostly constituted by several variables or layers. This criterium implies that, independently on the other layers (age, gender, income…) which could be exploited for empirical analysis on topics related to gender, inequalities, among others, education is the core variable of our data mapping exercise.
- **Collected in sub-Saharan Africa:** This criterium implies that the countries we are interested in when applying this data mapping methodology are sub-Saharan African countries.
- **Recent** i.e., collected between 2010-present. However, for periodical (repeated) surveys, waves of surveys before 2010 can be considered, as this might be very useful for researchers interested in panel data and any type of time-series analysis.
- **The statistical unit is the individuals (or households):** To generate locally and policy-relevant knowledge, disaggregated and preferably individual or household level data is needed. However, datasets in which educational institutions and regions are the statistical units will be considered, since such data is useful for between institutions or regions comparative analysis.
- **Usable** i.e., stored in a way that allows the use of the data to inform education policy. This criterium also implies that the data satisfy certain quality criteria that need further discussion.

1.4. Quality criteria

Poor-quality data can be a source of inaccurate analysis and ill-conceived policy strategies. Thus, in addition to the relevance criteria, data identified through this mapping strategy needs to satisfy certain quality norms. Besides issues related to data accuracy, which seems less in the scope of a data mapping exercise and concerns more data analysts, anonymity and completeness are some dimensions of data quality to consider.[4]

- **Anonymity:** There are concerns around anonymity if individuals can be identified by the specificity in the data. As a rule, it is the data producer's responsibility to anonymise the information during the data collection process before sharing the data. We recommend this mapping exercise avoids datasets that give away the respondent's identity. This mostly concerns qualitative data (e.g., videos).
- **Completeness:** Datasets largely constituted or dominated by missing values should be avoided.[5] As incomplete datasets, datasets with large missing values, may be of less utility for research purposes, we recommend the researcher applying this mapping methodology to gauge the dataset for completeness.

[4] Data accuracy implies that the values reflect what they are supposed to reflect. Considering, for instance, a dataset with a variable "age", it is obvious that "age" can neither take negative values nor be out of a certain range. The data mapping exercise mainly aims to identify what data exists.

[5] Missing values occur when no data value (or information) is stored for the variable in an observation.

Finally, regarding availability and accessibility, data identified through the different steps is assumed to be available but not automatically accessible. As the object of the mapping is to identify what education data exists, regardless of whether it can be accessed, the researcher is not expected to discriminate among datasets based on challenges surrounding accessibility. Instead specifics on the accessibility of the data are introduced in the metadata using the supporting document: Country_EdDataMapping.

2. A strategy for mapping data

This section describes the critical stages of mapping data. The first step, which is an opportunistic approach, is anticipated to be quite limited in identifying data collected in SSA countries. In addition, we propose a systematic approach. The data and related metadata such as the data sources, countries/regions, years, phase of education, funders, purpose of the data collection and stakeholders producing the data are expected to be collected and introduced into a spreadsheet specially designed for this exercise.

2.1. STEP 1: Online search for education datasets

The online search for education datasets aims to identify data collected, among others, by public institutes of statistics, national and international organisations, as well as data collected by HEIs, NGOs and researchers.

2.1.1. Education data from national institutes of statistics

Accessing microdata on education from national institutes of statistics seems to be the first and obvious part of identifying education data. This involves visiting the websites and data repositories of the respective institutes of statistics and searching for datasets on education that match the relevance criteria. Datasets derived from demographic/household surveys that contain variables (questions) related to education are to be considered, as long as they can be used to generate locally relevant knowledge on education.

The principal limitation of this source of education data is that most national institutes of statistics record and document only data generated by surveys they have been involved in. To deal with this, data generated by international organisations and private stakeholders also need to be identified and accessed.

2.1.2. Microdata on education from International Organisations

Contrary to national institutes of statistics, listing and visiting the website of every single organisation for education data mapping seems unfeasible. A straightforward way to proceed is searching and accessing data stored in Microdata Libraries such as The World Bank Microdata Library, Afrobarometer, DataFirst, HDX, UK Data Service, and Harvard Dataverse, among others. These data repositories/libraries centralise hundreds of survey data and datasets available on the internet, but sometimes with no comprehensive description or

metadata. When possible, it is recommended that the researcher accesses every single dataset for relevance and quality.

2.1.3. Micro-level education data from NGOs, researchers and others

Besides funding data collection activities and several studies, NGOs also produce large amounts of potentially rich data, most of which are not available online and used for empirical research. Thus, mapping education microdata of NGOs seems an important but largely unrealised opportunity to help researchers and policy analysts provide insights into critical education issues.

Capturing education data collected by NGOs requires the identification at the country-level of organisations (NGOs) working on any aspect of education and screening their last reports for new evidence on education. OpenAFRICA and NGO Advisor provide a non-exhaustive list of organisations working and concerned with education, which may serve as a starting point in identifying education actors (NGOs) at a specific country level. At this step, the researcher is expected to establish a list of education NGOs and to access their reports for potentially new education datasets. [6] Metadata relating to the datasets retrieved as well as the organisations identified through this process will be introduced in the spreadsheet for country-level stakeholders' consultations.

Despite the efforts to extend the reach of this mapping strategy, every single dataset collected by private (research) institutions and organisations will hardly be captured. Thus, to complete Error! Reference source not found., discussed so far, the next stage, Error! Reference source not found., proposes to systematically capture microdata used in empirical research on education.

2.2. STEP 2: Screening empirical research for education data

This step, screening empirical research on education for datasets, follows the procedure of a systematic review. For this to be comprehensive, we recommend the evidence research to exploit existing academic databases such as Web of Science (WoS), as the latter extensively feature research papers, books, and book chapters that can be filtered using keywords. To ensure the reproducibility of this process, elements relating to the database (WoS) and search queries need to be very detailed and precise.

2.2.1. The database (Haraldstad & Christophersen, 2015)

Haraldstad and Christophersen (2015) define WoS as an "interdisciplinary database with records from several bibliographic databases". It can be seen as a unifying platform, enabling

[6] Although every NGOs concerned with education in a country is not expected to hold education data, listing the latter will be very helpful in mapping out the current education initiatives and actors for the stakeholder consultations.

access to multiple databases, and allowing navigation through the latter in a timely manner. Figure 2 below depicts the database as it appears after logging in.

We recommend the search for research papers (documents) to use the WoS All Database, instead of the Core Collection, as the output of the formal option will always be greater or equal to the one obtained by using the latter (WoS, webpage). Consecutive to specifying a search in "All Databases" and "All Collections", the search query and the strategy to refine search results deserve a discussion.

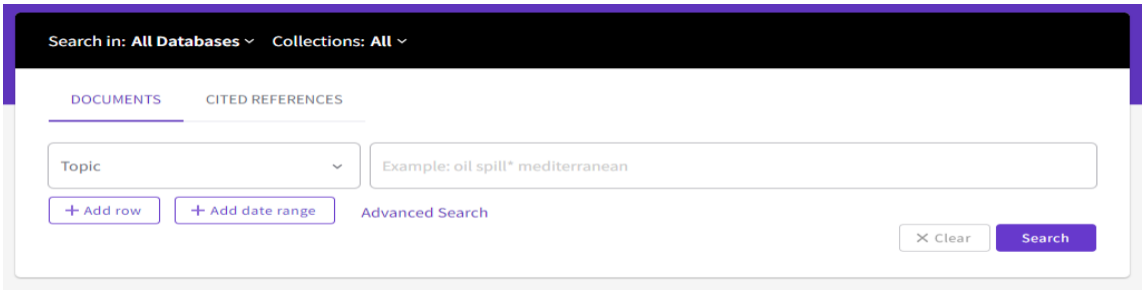

Figure 2: Web of Science search platform

2.2.2. Search query

Together with the **Error! Reference source not found.** related, among others, to years (2010-present) and geographical area (country), the search terms need to be very specific. For the present exercise to map out education data at country level, we recommend the following search terms: "education", "school", "skills development", "training", "learning", "TVET" and "Data" "Empirical analysis".[7] Figure 3 illustrates how the search terms can be used in WoS. The latter process delivers a longlist of studies containing the search terms in their title, abstract or author keywords, which needs to be refined.

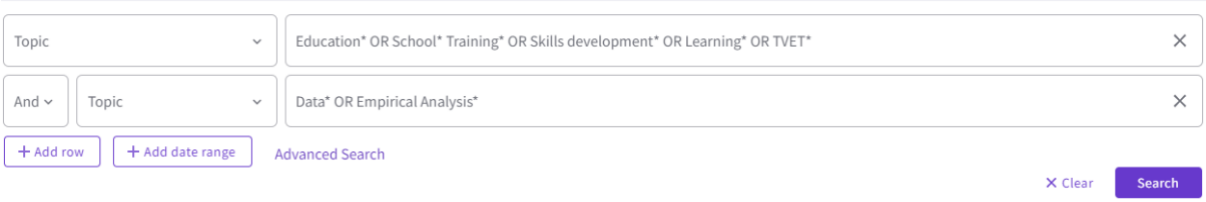

Figure 3: Illustration of a search in Web of Science

2.2.3. Refining the search result

This process consists of specifying the country, the publication years, and the research domains, among others.

- **Countries:** The country the researcher focuses on, mapping education data.
- **The publication years:** As previously mentioned, we recommend 2010-present.

[7] The search is conducted in the "Topic" as this covers title, abstract, author keywords, and Keywords Plus.

- **Research domains:** We suggest selecting "Social Sciences". It is to note that the "research domains" differ from "research areas" in WoS, and is constituted by *Science Technology*, *Life Sciences Biomedicine*, *Social Sciences*, *Physical Science*, *Technology* and *Arts Humanities.*

Figure 4: Web of Science-based search for research papers

Adopting the search strategy as described in WoS for Cameroon delivers a list of 329 documents.[8] Next, the search result will be screened and reduced to papers discussing any aspect of education, and in which real data have been used (no simulated data). Once the search process is completed, the researcher will consider the papers individually to identify and access the dataset used and collect metadata such as the source of the data, the area of focus, the phase of education concerned, the name of the survey. The joint spreadsheet provides the full list of metadata.

2.3. STEP 3: Country-level stakeholders and consultations

This education data mapping strategy, as described in Error! Reference source not found. and Error! Reference source not found., is concerned with identifying, visualizing microdata (when possible) and collecting additional information about the data. The latter includes, among others, organisations, researchers, and institutions that have funded, collected or produced the data. Collecting the data-related information at country level and for each dataset will deliver a list of public and private organisation and institutions, researchers as well as NGOs involved in education data collection and use. Such a list will be of major importance as the next steps require engagement (discussions) with stakeholders and advocacy.

Consultations with local education actors are crucial in identifying gaps and missing data: data needed for a clearer picture of the education system, data that have been collected but are not publicly accessible and data needed but that have not been collected yet. Such consultations are also decisive in understanding why a specific data is not accessible and what strategies can be used to make data publicly available. For these group discussions, the country-level list of education data stakeholders will be used. Finally, including education authorities and regulatory bodies in these consultations will be vital not only in making useful data available for research. This is also vital in understanding some barriers faced by public authorities to share data and learn about the types of data that have more influence in policy-making.

[8] The researcher is expected to produce a short protocol describing his individual settings.

2.4. Metadata presentation

As mentioned throughout this document, subsequently to identifying an education data set, the host website of the institutions producing the data will be accessed or the researchers/authors will be contacted, if needed, for the purpose of metadata collection.

Table 1: propose an overview of the metadata matrix.

Metadata	Definition
Source of Data	Source where the dataset has been identified
Name of survey	Name given by data producers to the collection project
Year	Years covered by the survey, or the specific dataset identified
URL	URL leading to the website of the organisation, institutions, report, or research paper in which the dataset has been identified or used
Country	The country or countries where the data has been collected
Region	Region, if provided in the description of the dataset
Phase of education	Phase of education concerned: Primary, secondary, tertiary…
Type of education	General programme or TVET
Type of access	The researcher is expected to provide details on the accessibility of the data. This could be, among others, open, accessible upon request, not accessible.

Note: See the Country_EdDataMapping for a full list of metadata

Dataset-related information such as the source of data, type of organisation, area of focus and the URL to the data, country/region of focus, most current available data, the sample size, funders and the phase of education, etc. will be extracted and introduced into the Country_EdDataMapping spreadsheet. Table 1 presents the main metadata and their definition.

3. Concluding remarks

This document presents a data mapping strategy that requires that the data match some specific relevance criteria. The expected outcome of the different steps is a country-level education data map, associated with descriptive metadata and a longlist of stakeholders involved in education data collection and use, which will be of prime importance for the ongoing 'Unlocking Education Data Campaign'. Of course, this document can be adapted and modified to fit the context, if necessary.

References

Adam, T., Agyapong, S., Asare, S., Heady, L., Wacharia, W., Mjomba, R., Mugo, J., Mukiria, F., & Munday, G. (2020). *Unlocking Data to Tell the Story of Education in Africa: Webinar Summary & Synthesis*. Zenodo. https://doi.org/10.5281/zenodo.4279156

AERD. (2020). *African Education Research Database*. https://essa-africa.org/AERD

Blom, A., Lan, G., & Adil, M. (2016). *Sub-Saharan African Science, Technology, Engineering, and Mathematics Research*. 115.

Buchanan, P. D., & Bryman, P. A. (2009). *The Sage Handbook of Organizational Research Methods*. SAGE Publications Ltd.

Cambridge Dictionary. (n.d.). *Cambridge Dictionary*. Retrieved August 6, 2021, from https://dictionary.cambridge.org/de/worterbuch/englisch/data

Crompton, H., & Burke, D. (2018). The use of mobile learning in higher education: A systematic review. *Computers & Education*, *123*, 53–64. https://doi.org/10.1016/j.compedu.2018.04.007

eBASE in Africa, E. B. (n.d.). *EBASE in Africa*. Effective Basic Services. Retrieved August 6, 2021, from https://www.ebaseafrica.org/programs

EdTech Hub. (n.d.). *EdTech Hub*. EdTech Hub. Retrieved August 6, 2021, from https://edtechhub.org/

ESSA. (n.d.). Retrieved August 6, 2021, from https://essa-africa.org/about_ESSA

Hallberg Adu. (n.d.). *Describing the Elephant: Data for Higher Education*. Retrieved August 6, 2021, from https://essa-africa.org/node/1101

Haraldstad, A.-M. B., & Christophersen, E. (2015). Chapter 5—Literature Searches and Reference Management. In P. Laake, H. B. Benestad, & B. R. Olsen (Eds.), *Research in Medical and Biological Sciences (Second Edition)* (pp. 125–165). Academic Press. https://doi.org/10.1016/B978-0-12-799943-2.00005-7

Haßler, B., Haseloff, G., Adam, T., & Others. (2020). *Technical and Vocational Education and Training in Sub-Saharan Africa: A systematic review of the research landscape* (Version 1.0). Bundesinstitut für Berufsbildung.

Koepke, R., & Paetzold, S. (2020). Capital Flow Data—A Guide for Empirical Analysis and Real-Time Tracking. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3686036

Masefield, S. C., Megaw, A., Barlow, M., White, P. C. L., Altink, H., & Grugel, J. (2020). Repurposing NGO data for better research outcomes: A scoping review of the use and secondary analysis of NGO data in health policy and systems research. *Health Research Policy and Systems*, *18*(1), 63. https://doi.org/10.1186/s12961-020-00577-x

Open Burkina. (n.d.). *Open Burkina – Savoir pour décider ensemble*. Retrieved August 6, 2021, from https://www.openburkina.bf/

Pittaway, L., & Cope, J. (2007). Entrepreneurship Education: A Systematic Review of the Evidence. *International Small Business Journal*, *25*(5), 479–510. https://doi.org/10.1177/0266242607080656

United Nations, SDG-4. (n.d.). *Goal 4 | Department of Economic and Social Affairs*. Retrieved August 6, 2021, from https://sdgs.un.org/goals/goal4

United Nations, SDGs. (n.d.). *THE 17 GOALS | Sustainable Development*. Retrieved August 6, 2021, from https://sdgs.un.org/goals

Unlocking Data. (n.d.). Retrieved August 6, 2021, from https://unlockingdata.africa/

Zizi Afrique Foundation—Children and youth learning and thriving. (n.d.). Zizi Afrique Foundation. Retrieved August 6, 2021, from https://ziziafrique.org/