german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

In cooperation with
Global Partnership for Sustainable Development Data

Citizen-generated data in Kenya: a practical guide

Understanding and generating quality citizen-generated data to improve policy and decision-making

Acknowledgements

This guide was produced by the Global Partnership for Sustainable Development Data team: Karen Bett, Jenna Slotin and Davis Adieno. Stephen Okiya led the technical writing of this work. Colleen Wile managed the process to develop this guide.

This work would not have been possible without the contributions and support of external contributors.

Our thanks to the staff of the civil society organizations in Kenya who provided input into the development of this guide. Specifically, we are grateful to Florence Syevuo from the SDGs Kenya Forum, Al Kags and Benjamin Charagu from the Open Institute for reviewing the guide and providing useful comments.

We thank the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) for the funding that enabled this work. We especially recognize Kim Schultz, Paul Mungai, Edward Murimi, and Mark Irura Gachara.

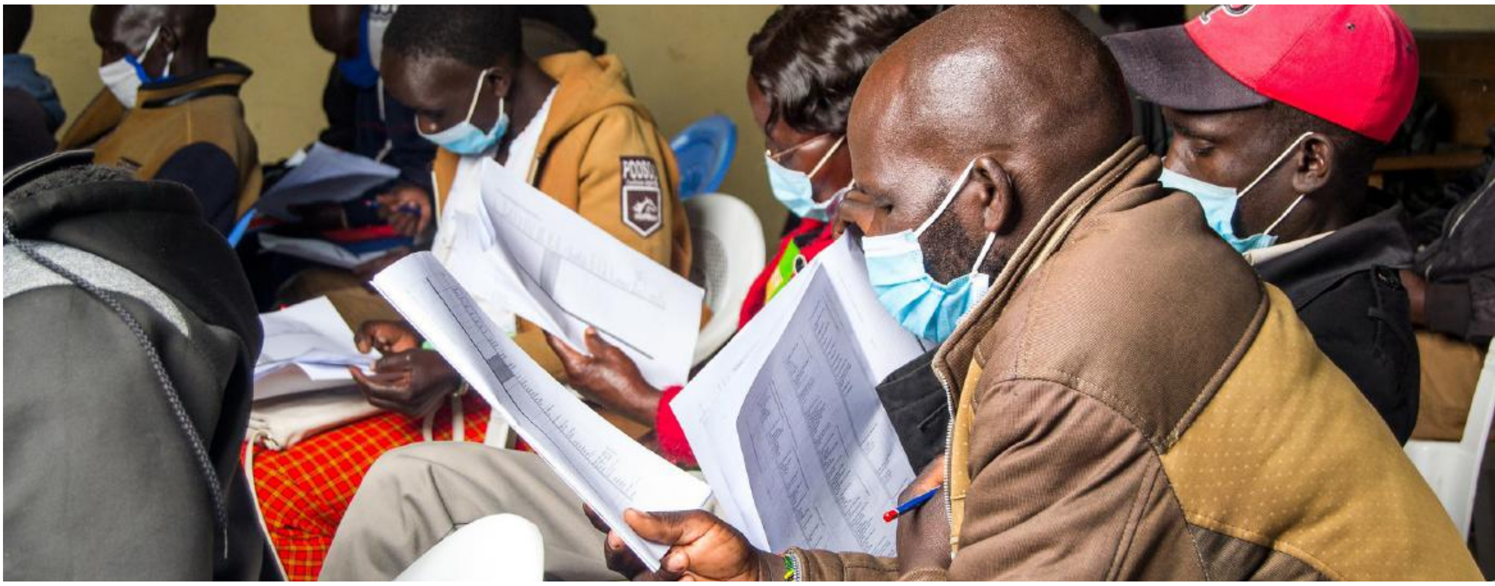
Finally, we are indebted to the Kenya National Bureau of Statistics (KNBS) for their support throughout this project. We thank the KNBS senior management and quality working group team for taking the time to review this guide over three workshops, and for their valuable inputs in strengthening this guide. Special thanks to Benjamin Avusevwa, Mutua Kakinyi, Sarah Omache, John Bore, and Linah Ngumba for leading and organizing the review workshops.

Data Entry in Eldoret Town, EMC. Credit: Open Institute

Contents

List of acronyms

CGD	Citizen-generated data
CSO	Civil society organization
GSBPM	Generic Statistical Business Process Model
GSS	Ghana Statistical Service
GPSDD	Global Partnership for Sustainable Development Data
IMF	International Monetary Fund
ISO	International Organization for Standardization
KNBS	Kenya National Bureau of Statistics
NSO	National Statistical Office
NSS	National Statistical System
SDGs	Sustainable Development Goals
UNECE	United Nations Economic Commission for Europe
UNICEF	United Nations International Children’s Emergency Fund
UN NQAF	United Nations National Quality Assurance Framework
UNSC	United Nations Statistical Commission
UNSD	United Nations Statistics Division



Training of CHWs in Kapyego Ward, EMC. Credit: Open Institute

Chapter 1: Introduction

1.1 Background

The adoption of the Sustainable Development Goals (SDGs) in September 2015 injected a new sense of urgency to the role of data in accelerating both monitoring and achieving the SDGs (UNDP, 2015). The case for alternative data sources was amplified by the Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda, which called for "a data revolution for sustainable development, with a new international initiative to improve the quality of statistics and information available to citizens." The report recommended harnessing new technology, crowdsourcing, and improved connectivity to empower people with information on progress towards the targets (United Nations, 2013).

Official statistics have traditionally taken the role of underpinning the success to the Sustainable Development Goals. However, with advances in technology alternative data sources such as citizen-generated data (CGD), mobile data, geospatial data, and big data have become increasingly relevant. The changes to how and where data is being generated have allowed citizens to generate great volumes of data on issues that affect them. This is known as CGD, defined as "data that is produced by organizations or people to monitor issues around them and to drive or demand change on issues that concern them" (CIVICUS, 2015). This approach platforms the opinion of citizens and can help to inform tailored policies that are responsive to the needs of citizens. CGD has been used for years to tackle issues that matter to citizens around the world.

CGD offers an important complement to official data produced by national statistical offices (NSOs) in driving forward a data revolution for sustainable development. CGD can provide timely and granular data on issues that affect citizens. CGD empowers citizens by engaging them in one or several stages of the data value chain—from production to analysis, dissemination, and use—on issues that matter to them. It allows people to collaborate and contribute to public decision-making by ensuring inclusivity and responsiveness, although challenges remain with ensuring representation.

Using alternative data sources, including CGD, was new to many NSOs and official statistics teams when the SDGs were adopted in 2015. Over time, the official statistics stakeholders have come to see the value of CGD, but there is still a need to build mutual understanding of the different approaches and methods for using CGD.

CGD's potential to support sustainable development is widely accepted. Yet challenges relating to access, quality, production, and use remain. By following more rigorous standards, methods, and classification, CGD can be more easily used by NSOs and government decision-makers and systematically implemented by all data providers (UNSD, 2020).

1.2 Rationale

This guide responds to a call from civil society organizations in Kenya for guidance on producing quality CGD that adheres to rigorous standards and draws on key concepts related to official statistics. It reflects on case studies from CGD actors across the country and documents their approaches to generating data.

This guide was produced with the Kenyan context in mind, but the issues addressed are cross-cutting in nature. Furthermore, this work is inspired by a guide published in 2019 by the Global Partnership's CGD task team[1] to help stakeholders better understand whether CGD is suitable for their proposed project, alongside identifying what type of data is appropriate for their needs. It was designed for governments and non-governmental organizations interested in developing, engaging with, and supporting CGD initiatives. It presents a list of distinct criteria between CGD methods, highlights the benefits and pitfalls of CGD, and provides a basis for strategic engagement with CGD (GPSDD, 2019). The guide drew from an analytical framework that revolves around three aspects: workflows to generate data, participation, and data's fitness for purpose (GPSDD, 2019).

[1] The task team on CGD has three focus areas: produce recommendations on CGD initiatives for different purposes, actors, and needs; produce guidance on how to navigate and engage with different types of CGD initiatives and provide a forum to share experiences, challenges, and learning related to CGD. See more details: https://www.data4sdgs.org/initiatives/citizen-generated-data-task-team.

The structure of this guide is as follows:

- Chapter 1 sets the scene on CGD, provides examples of CGD in the country, and introduces official statistics in Kenya;
- Chapter 2 presents best practices on producing quality data learning from official statistics;
- Chapters 3–8 walk through the data value chain, from the identification of needs to evaluation and provide guidance on each stage;
- Chapter 9 speaks on data privacy as a key priority in data production;

At the end of each chapter, additional reading material is provided. The Annex includes a checklist of the guide, definition of key statistical terms and a roadmap to institutionalizing CGD among CSOs and NSOs.

1.3 Methodology

This guide was developed using a human-centered approach, listening and learning as much as possible from CSOs and NSOs. The approach consolidates learning from the following:

- **A multi-stakeholder peer-to-peer exchange between Ghana and Kenya held in 2019**. One discussion session focused on how to integrate CGD into official statistics or apply it for official purposes. Participants held group discussions (led by representatives from the Kenya National Bureau of Statistics (KNBS) and the Ghana Statistical Service (GSS) on how civil society can produce high-quality standards of CGD. The Lanet Umoja Pilot project was used as a case study (GPSDD, 2019).
- **A session on CGD at the 2019 Data Tamasha event hosted by Tanzania Data Lab.** This session brought together the Tanzania National Bureau of Statistics, KNBS, the county government of Vihiga, Open Institute, Humanitarian OpenStreetMap and Twaweza to share experiences in working with CGD[2].
- **A co-creation workshop with CSOs in early 2020**. The workshop aimed to share knowledge on existing data guidelines within KNBS and among CSOs and gather insights on priority areas to be covered in this guideline.
- **A side event at the United Nations Statistical Commission (UNSC) in 2020**. This brought together NSOs and CSO representatives across the globe to discuss ways of building trust in CGD.
- **A workshop with teams from the KNBS in the second half of 2020**. This workshop aimed to sensitize KNBS on CGD and receive their reviews of the guide as well as forging a way to adopt CGD as an alternative source of statistics.

[2] Data Tamasha (Data Festival) brings together data enthusiasts, leaders and practitioners across industries, the public sector, and academia. It involves dialogues, panel discussions, presentations, demonstrations, interactive sessions, exhibitions and public outreach activities. See more details: https://datatamasha.dlab.or.tz/.

The learning from each of these activities contributed to the content and structure of this guide, as outlined below in Table 1.

Table 1: Learning from partners - challenges and solutions

Challenges	Lessons and possible solutions
CGD is not recognized by government officials because it is not official statistics.	Mutual quality standards can ensure that quality data is acceptable even if it is not generated by a government agency. Trust can be built between communities and government data agencies. As custodians of data, NSOs can provide data stewardship to guide production and use of CGD. However, there is a need to have clear guidelines governing the process of data production. Such guidelines will ensure compliance by other data stewards, such as CSOs.
The tools and capacity to collect and use CGD comprehensively is currently limited.	CSOs and citizens can be empowered on data collection and management skills, and tools can be co-created to help in making data useful.
Bureaucratic red tape is a barrier to acquiring information or seeking guidance on CGD from the government.	A good working relationship can be developed with the government and the issues that lead to the bureaucratic red tape can be addressed. A partnership between the government and non-state actors would ensure CGD would be easily understood and used to complement official statistics. Furthermore, it would help provide much-needed guidance, reducing the gap between the government and its citizens.
The scope and purpose of CGD is not clearly defined.	Clarity is a challenge when it comes to CGD. For example, it is not always clear what CGD is, how the data is collected, who the citizens are, and what the purpose of the data is. Not all CGD is meant to be applied for official purposes or government statistics, emphasizing that communities and non-state actors could also run independent CGD initiatives and enhance the quality of their data. By improving the quality of CGD, it can be more easily used by NSOs and government decision-makers and systematically implemented by all data providers.
Data is not available to the public, or data collection among government agencies is done in silos.	Multi-agency and mutual data protection policies can commit government and non-state actors to safeguard data gathering and sharing.
There is low funding for CGD and data-driven initiatives.	Funds could be allocated to CGD and data-driven initiatives.

1.4 Citizen-generated data as an alternative data source to official statistics

1.4.1 What is citizen-generated data?

Citizen-generated data is defined as "data that people or their organizations produce to directly monitor, demand or drive change on issues that affect them. This can be produced through crowdsourcing mechanisms or citizen reporting initiatives, often organized and managed by civil society groups" (Wilson and Rahman 2016, p. 16). It is actively given by citizens, offering a direct representation of their perspectives, and is an alternative to datasets collected by governments or international institutions.

The past few decades have seen the rise of CGD projects across the globe. CGD is critical to fostering collaboration among data producers and users and achieving the SDGs. It primarily fosters relationships through citizenship as well as supports monitoring and implementation of priorities, as outlined in Table 2. Beyond education, community engagement, and community-based problem solving, these actions include baseline research, planning and strategy development, allocation and coordination of public and private programs, and improvement to public services.[3]

Table 2: How CGD can help achieve the SDGs

Link to the SDGs	Role of CGD
Citizenship: Creating new relationships and public spaces	• As a venue for local development and education strategies. • Providing the frame for outreach and engagement with special interest communities. • Enhancing and innovating governments' engagement strategies with citizens. • Providing localized statistics that have a greater level of detail. CGD is detailed because citizens are often more knowledgeable on the issues that surround them day-to-day.
Monitoring: Informing, expanding, and improving SDG monitoring	• Producing data in regions otherwise not reachable by producers of official statistics. • Identifying patterns hidden behind averages. • Improving the capacity of states to detect issues. • Cross-verifying government data with CGD and vice versa. Example: Since 1999, Afrobarometer has conducted public opinion surveys on democracy, governance, the economy, and society in more than 30 countries on the continent and repeated on a regular cycle.
Implementation: Informing public policy goals and community-driven problem solving	• Providing baseline data for research and test assumptions. • Enabling planning, strategy development, and resource allocation. • Monitoring performance of public facilities. • Identifying root causes around problems. Example: CGD has proven useful during responses to natural disasters, such as the April 2015 earthquake in Nepal. Volunteers for the Humanitarian OpenStreetMap Team provided rescuers with much-needed local data.

[3] See examples of CGD from across the world and in different sectors that have influenced change data4sdgs.org/sites/default/files/services_files/Advancing%20Sustainability%20Together%20CGD%20Report_1.pdf and data4sdgs.org/resources/choosing-and-engaging-citizen-generated-data-guide

Examples of CGD projects across Kenya are outlined below:

CASE STUDY 1

Quality assurance in CGD: Usawa Agenda

Usawa Agenda was established in Kenya, growing from the UWEZO project. It is an initiative that seeks to improve literacy and numeracy skills among children aged 6–16 years across three countries in East Africa: Kenya, Uganda, and Tanzania.

The organization's innovative approach enables it to conduct large-scale, citizen-based, and household-based assessments of the levels of literacy and numeracy competencies among children. It has done this for the last 10 years.

The initiative's stringent measures in quality assurance for its citizen-generated data include:

- Ensuring proper sampling per the KNBS guidelines.
- A diverse network of actors to the subnational level.
- Developing data collection tools with government organizations. It is currently supported by a technical team of 30 specialists from the Ministry.
- Ensuring that enumerators are well-trained and that staff monitor data collection.
- Ensuring data is validated, protected from unauthorized access, rechecked, and audited to ascertain compliance with standards.

CASE STUDY 2

Quality assurance in CGD: Open Institute (OI)

OI has worked with communities for the last 10 years. It helps the community understand what the power of data can do for them. Communities are guided by OI to collect their own data and to interpret the data. These communities design the tools that help them collect data relevant to them.

For example, the farming community would gather data that would inform them about their activities, produce inputs and markets, while women groups would want to know about economic opportunities they could participate in.

The community then uses the findings to make better and targeted decisions, creating conversations with other stakeholders including government.

The organization's steps in quality assurance for its citizen-generated data include:

- Citizens are encouraged to participate in the design and execution of the data activities.
- Once they have been trained, they collect their own data. This promotes comprehensiveness and accuracy of the data.
- The data is analyzed (community members are trained to analyze) and presented back to the community. At this stage, the community also verifies the findings.

1.4.2 Producing official statistics

Official statistics are developed, produced, and disseminated as a public good by the members of National Statistical System's (NSS).[4] These statistics must comply with international classifications and methodologies, such as the United Nations Fundamental Principles of Official Statistics and accepted quality frameworks, such as the United Nations National Quality Assurance Framework (UN NQAF), as well as other internationally agreed statistical standards and recommendations (UNSD, 2014; UNSD, 2019; UNSD, 2020.

Kenya has domesticated the fundamental principles of official statistics in the 2019 Statistics Amendment Act (see Table 3). Integrating these principles in the production of CGD could significantly enhance its quality.

1.4.3 Producing official statistics in Kenya

The KNBS is the principal government agency for collecting, compiling, analyzing, publishing, and disseminating statistical information for public use. It also coordinates, monitors, and supervises the NSS, as guided by the 2006 Statistics Act (GoK, 2006). Under its various directorates, the KNBS produces data and statistical information (statistical services) that conform to the principles outlined in Table 3 and other international standards.

Annex 1 defines the key concepts and definitions used by the KNBS and gives some concrete examples. These examples will help CSOs gain a better understanding of different concepts and definitions as they implement data initiatives. A full list of the definitions and concepts can be found in the KNBS' compendium of general concepts and definitions.[5]

Data Collection in Tambach Ward, Elgeyo Marakwet. Credit: Open Institute

[4] See The national statistical system is the group of statistical organizations and units within a country that jointly collect, process, and disseminate statistics on behalf of the government. See more details: oecd.org/sdd/na/1963116.pdf

[5] By the time of publishing these guidelines, KNBS was finalizing its Kenya National Quality Assurance Framework (KeNQAF), which includes further guidance on CGD as well as a quality criteria for CGD.

Table 3: United Nations Fundamental Principles of Official Statistics and its application in Kenya

Principle	Explanation	How the principle is applied in Kenya
Principle 1: Relevance, impartiality, and equal access	The UN Fundamental Principles of Official Statistics stipulate that "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." Thus, they should "be compiled and made available on an impartial basis by official statistical agencies to honor citizens' entitlement to public information."	Official statistics that meet the test of practical utility should be compiled and made available on an impartial basis by the KNBS to honor citizens' entitlement to public information
Principle 2: Professional standards, scientific principles, and professional ethics	The production of official statistics ought to be ethical, impartial, and based on high professional standards to maintain public trust.	To retain trust in official statistics, KNBS shall apply strictly the professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage, and presentation of statistical data.
Principle 3: Accountability and transparency	The public has the right to be informed about all the statistical works of the statistical organization by presenting information in accordance with scientific standards, methods, and procedures.	To facilitate a correct interpretation of the data, the KNBS shall present information according to scientific standards on the sources, methods and procedures of the statistics.
Principle 4: Prevention of misuse	The statistical agencies have the right to comment on erroneously interpreted statistical information to maintain integrity and trust of official statistics by the public.	The KNBS is entitled to comment on erroneous interpretation and misuse of statistics.
Principle 5: Sources of official statistics	Various data sources may be used as sources of official data since production of official statistics can be laborious and cost intensive.	Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. The KNBS shall choose the source with regard to quality, timeliness, costs, and the burden on respondents.
Principle 6: Confidentiality	Statistical data should be strictly confidential and only be used for statistical purposes or for the purposes mandated by the law.	Individual data collected by KNBS for statistical compilation, whether they refer to natural or legal persons, shall be strictly confidential and used exclusively for statistical purposes.
Principle 7: Legislation	Laws and regulations governing the production of official statistics and under which the statistical systems operate are supposed to be made public.	The KNBS shall cooperate with other national statistical agencies within their countries to achieve consistency and efficiency in the NSS. (Principle 7 & 10)
Principle 8: National coordination	Coordination of national statistical programs is very important to ensure consistency and efficiency across the various statistical entities at the national level.	KNBS shall apply the international concepts, classifications, and methods to promote the consistency and efficiency of the NSS at all official levels.
Principle 9: Use of international standards	The use by statistical agencies in each country of international concepts, classifications, and methods promotes the consistency and efficiency of statistical systems at all official levels.	
Principle 10: International cooperation	The improvement of statistical systems occurs from bilateral and multilateral cooperation between the national official statistics in the different countries globally.	

Chapter 2: Data quality assurance and the data value chain

2.1 International best practices

This chapter introduces guidelines on producing quality data which guides KNBS. By the time of publishing this guide, KNBS was in the process of finalizing its Kenya National Quality Assurance Framework (KenQAF). KenQAF aims to domesticate the international guidelines shared in the following sections.

2.1.1 United Nations National Quality Assurance Framework (UN-NQAF)

The UN-NQAF was adopted in 2019 by the United Nations Statistical Commission (UNSC). The UN-NQAF aims to guide countries in the implementation of a national quality assurance framework, including for new data sources such as CGD.

UN-NQAF recommends that the national quality assurance framework is implemented at the level of NSOs and throughout the entire NSS. Kenya is currently finalizing its Kenya National Quality Assurance Framework (KeNQAF), which benefits from the UN-NQAF. The national quality assurance framework should also be applied to all data and statistics produced outside of the NSS that is disseminated with the help and support of a member of the NSS or that is used for government decision making, as deemed appropriate and required (UNSD, 2019).

Table 4: Principles guiding the UN-NQAF

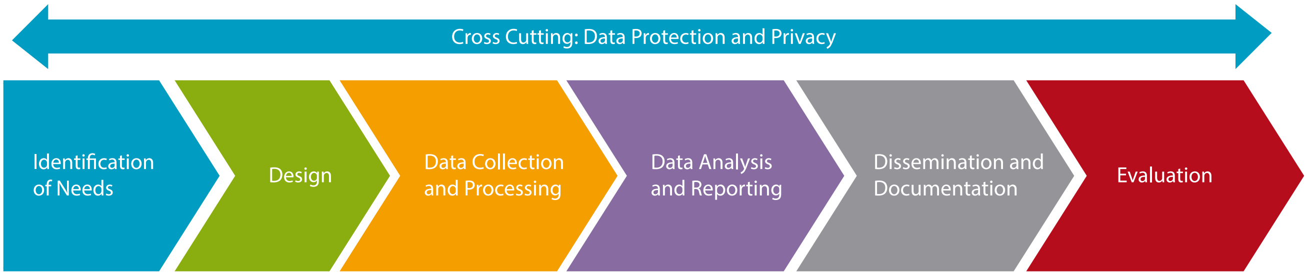
Levels	Quality principles and requirements
Level A: Managing the statistical system	Principle 1: Coordinating the NSS Principle 2: Managing relationships with data users, data providers, and other stakeholders Principle 3: Managing statistical standards
Level B: Managing the institutional environment	Principle 4: Assuring professional independence Principle 5: Assuring impartiality and objectivity Principle 6: Assuring transparency Principle 7: Assuring statistical confidentiality and data security Principle 8: Assuring the quality commitment Principle 9: Assuring adequacy of resources
Level C: Managing statistical processes	Principle 10: Assuring methodological soundness Principle 11: Assuring cost effectiveness Principle 12: Assuring appropriate statistical procedures Principle 13: Managing the respondent burden
Level D: Managing statistical outputs	Principle 14: Assuring relevance Principle 15: Assuring accuracy and reliability Principle 16: Assuring timeliness and punctuality Principle 17: Assuring accessibility and clarity Principle 18: Assuring coherence and comparability Principle 19: Managing metadata

Exposing non-state actors to the principles and frameworks of the UN-NQAF will help them better understand how this relates to the production of their own data and how that process can be improved.

2.1.2 Generic Statistical Business Process Model

The Generic Statistical Business Process Model (GSBPM)[6] provides a roadmap of the various set of business processes required to produce official statistics. The model offers a standard framework and consistent definitions that can be adopted by different data producers with an aim of ensuring production of high-quality data. The model outlines best practices for data and metadata management.

Figure 1: Steps in data production process


Cross Cutting: Data Protection and Privacy
Identification of Needs
Design
Data Collection and Processing
Data Analysis and Reporting
Dissemination and Documentation
Evaluation

The GSBPM consists of eight phases, with further sub-processes identified within these phases. It is not a rigid framework, so is not mandatory to follow the different phases of the GSBPM in any order and can be iterative at times.

Table 5: Phases of the Generic Statistical Business Process Model

Specifying needs	This phase identifies the purpose for which data will be produced. It establishes the objectives based on the needs, clarifies the concepts from the user's point of view and prepares a business case upon which the data production is carried out. For this phase, it is important to involve key stakeholders and to develop a data production process that is as cost-effective as possible.
Design	This phase describes the design and the development activities to be carried out and any other practical research work that may be needed to define the data production outputs, concepts, and methodologies. All relevant metadata is also defined under this phase.
Build	This phase includes building and testing of the instruments to collect and disseminate the data, to the point where they are ready for use in the actual data collection. For cases where statistical outputs are generated regularly, this phase is done after the first iteration.
Collect	In the 'collect' phase, different collection models are used to collect all the necessary data and place them in the appropriate data environment.
Process	The 'process' phase involves the cleaning of the data records and preparing them for analysis. Here data is checked, cleaned, and transformed. This is an iterative process for statistical and non-statistical sources of data.
Analyze	The 'analyze' phase involves scrutinizing produced data and preparing it for dissemination. It encompasses the activities and sub-processes that help analysts to understand the produced data.
Disseminate	The 'disseminate' phase consists of managing the release of the outputs to the users.
Evaluate	Evaluation takes place at the end of a data production process or can be done on an ongoing basis as per the organization's policy or rules. Evaluation relies on all the inputs collected during all the different phases as above. Instances within this process are evaluated for success based on the quantitative and qualitative inputs, identifying weaknesses and making improvements where necessary.

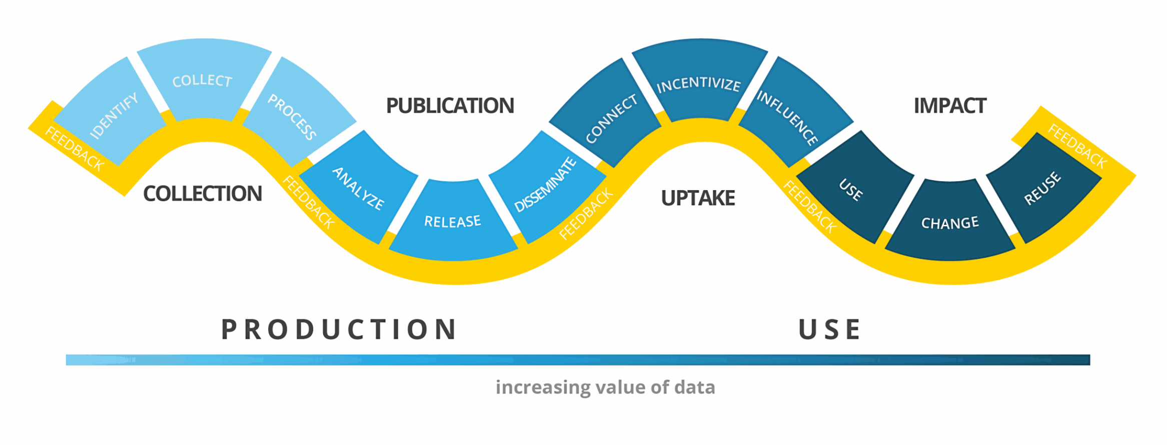
[6] The GSBPM was developed by the United Nations Economic Commission for Europe (UNECE) to provide a standard framework on which business processes seeking to produce official statistics can be anchored on. It can also be used for integrating data and metadata standards, as a template for process documentation, for harmonization of statistical computing infrastructure and to provide a framework for process data quality assessment and improvement. See more details: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/bur/2009/mtg1/20.add.2.e.pdf

2.2 The data value chain

Data value chains describe the process of data creation and can be used from first identifying a need for data to its final use and possible reuse. The Open Data Watch and Data2X data value chain has four major stages: collection, publication, uptake, and impact, as outlined in Figure 2 below. These four stages are further separated into twelve steps: identify, collect, process, analyze, release, disseminate, connect, incentivize, influence, use, change, and reuse.

Throughout the process, there should be constant feedback between producers and stakeholders. The data value chain can be used as a guiding tool to show the complex set of steps from data creation to use and impact or as a management tool to monitor and evaluate the data production process (Open Data Watch & Data 2X, 2018).

Figure 2: Data Value Chain



2.3 Why is the data value chain important for CGD?

Each section of the data value chain involves a set of processes and activities, on data production and use, which can help non-state actors to significantly improve the quality of their CGD data.

These guidelines narrow the focus to the data production stage: design, data collection and processing stages, data analysis and reporting and finally dissemination and documentation as seen in Figure 1 and guided by the GSBPM (UNECE, 2014). These were identified by stakeholders as areas with critical gaps. Users should be encouraged to consider completing the data value chain by promoting use of the data they produce.

Chapter 3: Identifying the purpose for which data will be produced

This chapter describes the importance of collaboration and engaging key stakeholders throughout the data production process as well the steps required in the identification of needs.

3.1 Engaging National Statistical Offices

The KNBS (like other NSOs) is central to Kenya's production of statistics through their supervision role. It coordinates national statistical systems, ensures production of reliable statistics, and has extensive experience on best practices and standards. KNBS has in the past partnered with various non-state actors in its data-driven initiatives.

KNBS can play a critical role in a CGD data study[7] in the following ways:

- Offering technical inputs in the design of study instruments to ensure the data collection tools are reliable and accurately measure the desired indicators.
- Providing a sample from its sampling frame, the National Sample Survey and Evaluation Program, to ensure estimates are representative of the target population.
- Supporting the training of enumerators or trainers.

3.1.1 Contacting local authorities to establish a local presence

Most CGD studies are at the sub-national level, and so have relatively limited coverage. It is advisable to let local authorities know you intend to conduct a study in their geographical region.

Contacting a local authority can be useful because:

- CGD is generated and used with the communities, including by local authorities. This builds data ownership, strengthens local data skills, increases data uptake, and gives communities a voice.
- CGD studies may generate data that are useful for local authorities to fill data gaps and therefore strengthen government systems.
- Local authorities can be key respondents and amplifiers of the study through barazas (village meetings) and other opportunities.

3.2 Identifying needs

This phase begins by identifying the demand or need to produce CGD or to review previously existing data to establish if it meets the user's needs.

This phase is divided into six stages:

1. **Determining the need for the information** – Investigating what statistics are needed and what the requirements are. Additionally, reviewing best practices for producing the statistics based on methodologies from other organizations.

2. **Consulting and confirming the need** - Consulting stakeholders on their needs to confirm the need for the statistics. This gives the organization involved a clear picture of what is required of them in terms of delivery.

3. **Establishing objectives and outputs** - Expected outputs are established after consulting and confirming the needs with the users and relevant stakeholders.

4. **Identifying concepts** - identifying the concepts to be measured and how to measure their quality.

5. **Checking data availability** - Available data is checked and reviewed according to whether it meets user requirements and the conditions under which it is to be made available. When all existing sources have been critically reviewed, a decision is made on how to fill any gaps that may exist.

6. **Preparing a business case** - The findings for all the other stages above are documented to get approval for a new business case or for review of an existing one and the best method for implementation.

[7] While KNBS can help CGD data producers to generate high-quality data, producers need to be aware that KNBS can be stretched. This guide is part of efforts to provide the data collection tools that align with NSO standards. As good practice, it is also important to map out other producers of data to ensure there is no oversaturation in data-driven activities and to minimize respondent burden.

Chapter 4: Designing data collection

The design phase is the backbone of the entire data collection process, identifying the roadmap to be followed. It covers the statement of objectives, instrument design, and sample design of the study.

See the design phase and sub-processes below:

A. **Design outputs** - Detailed design of the statistical outputs, services, and products that will be produced. Processes that relate to confidential outputs are designed here. Existing standards should be followed, such as international ones.

B. **Design variable description** - Defining variables to be collected and any other variables that may arise in processing and analysis. Statistical classifications used are also defined here. This sub-process may run together with the design collection.

C. **Design collection** - Determining appropriate methods of data collection and collection instruments. Design of collection instruments, templates and mechanisms to monitor data is done here.

D. **Design frame and sample** - Identifying the population of interest, sampling frame, sampling criteria, and methodology is done here. This only applies where sampling is employed in data collection.

E. **Design processing and analysis** - Designing a statistical processing methodology. It also includes design specifications for data integration, validation, and estimation.

F. **Design production systems and workflows** - Determines the workflow from the collection of data to its dissemination. It also considers how staff will engage with the systems and what responsibility each will bear.

4.1 Study formulation and planning

4.1.1 Specify the study objectives

When conducting a data study, the statement of objectives must be clear. These objectives guide all subsequent aspects of the study. Objectives should be well-defined to ensure that the study meets its intended purposes. It outlines the data that is required, the operational definitions to be used, primary users and uses of the data, specific topics to be addressed, and a plan for analyzing the data.

TOP TIP

What should you include in the statement of objectives?

- **Information needs**: This is determined by stating the problem in broad terms. What are the underlying issues and in what context have they been raised?

- **Users and uses of the data**: The users of the data need to be known, since their input is very important in the planning phase. The uses of data must be identified to specify more precisely what the information needs are.

4.1.2 Developing data collection instruments

When collecting data through a study, it is important to develop tools that ensure information is obtained in a structured manner and is of a high quality. The kinds of tools to be used depend on the context of the study. Examples of data collection instruments include questionnaires, manuals, physical devices, and equipment.

It is also imperative to determine whether the data to be collected is quantitative or qualitative and structured or unstructured. The study may adopt both qualitative and quantitative approaches as an effective and complementary means of collecting data. Quantitative methods include: survey observations, experiments, and interviews, in which numerical data is collected. This means that quantitative data can be measured, counted, and expressed in numbers. For instance, the number of children in each school. On the other hand, qualitative approaches include interviews, focus group discussions, case studies, and discourse analysis, where information on ideas, perceptions, and experiences is collected. The study objectives, indicators of interest, and available technical capacity determine the appropriate instrument(s) to be used.

Questionnaires are the most tried-and-true tools for collecting information from respondents. Questionnaire design is critical to ensure that the resulting data is of acceptable quality and adequately addresses all study objectives. Questions need to be framed exactly as they are to be asked of the respondents. They should be clear, concise, and appropriate to the cultural context. The ordering and wording of questionnaire items should be carefully considered since it enhances data accuracy. The questionnaire should also have clear instructions to either the respondent or the interviewer, depending upon the questionnaire's focus. This is usually captured in an instruction manual.

A good questionnaire should:

- Enable the collection of accurate information to meet the needs of potential data users in a timely manner.
- Facilitate the work of data collection, processing, and tabulation.
- Ensure economy in data collection; that is, avoiding collection of any non-essential information.
- Permit comprehensive and meaningful analysis and purposeful utilization of the data collected.

TOP TIP

Principles for questionnaire design

- Questions can either be open- or closed-ended. Closed-ended questions confine the respondent to predestined answers (such as "yes" or "no"), while open-ended questions allow the respondent to give his/her own answer to a question.

- The questions should be clear, precise, and unambiguous. The respondent should easily understand questions asked by the interviewer to avoid non-sampling errors.

- The questions should not lead a respondent to answer in a certain way. The questions should not be biased in favor of a certain answer.

- The questions should be relevant to most respondents.

- The questions should have logical connections to one another. The question order should motivate the respondent to answer them. Best practices recommend that the first questions be easy, interesting, and non-sensitive. Sensitive questions should come at the end and the approach of asking sensitive questions should be explained/observed.

4.1.3 Developing study manuals

Manuals and learning materials should be developed for interviewers and field supervisors that incorporate different data collection processes. The interviewer's manual helps interviewers maintain consistency in their responsibilities, such as recording respondents' responses to questions. The manual includes information such as study background, interviewing methods, field operations and procedures, data collection requirements, ethics, and safety while conducting interviews. The field supervisor's manual contains clear instructions on how to handle field activities.

4.2 Sample design

4.2.1 Choose a sampling design that is context responsive

Sampling is the foundation of inferential statistics, since it must ensure that the selected sample is representative of the whole population. It is important to identify the appropriate sampling frame(s) and sampling technique to infer the relationship between the target population and the unit of selection.

TOP TIP

Questions to consider while designing a sample

- What is the target population?
- What is the unit of selection?
- What is the effective sample size?
- Is the sample frame complete, accurate, and current?
- Can the KNBS sampling frame be of help?
- Is the probabilistic or non-probabilistic sampling method appropriate?
- What is the appropriate sampling technique (random, systematic, purposive, etc.)?

Sampling Frame: The sampling frame is used to identify and select sampling units into the sample and is also used as a basis for making estimates based on sample data. This implies that the population from which the sample has to be selected must be represented in a physical form. The frame ideally should have all sampling units belonging to the population under study with proper identification particulars. Frames should be exhaustive and preferably mutually exclusive.

Sampling frames should be complete, accurate, and current. In some cases, a listing of households or establishments may be an appropriate way to design a sampling frame.

4.2.2 Sampling methodologies

Sampling is the foundation of inferential statistics, since it must ensure that the selected sample is representative of the whole population. It is important to identify the appropriate sampling frame(s) and sampling technique to infer the relationship between the target population and the unit of selection.

Sampling Methods:

Sampling methods can either be **probabilistic** (a scientific method of selecting a fraction of items or individuals from the population in a random manner such that each item in the population has a known non-zero probability of being selected) or **non-probabilistic** (a non-scientific method of selecting a fraction of items or individuals from the population using subjective or personal experience).

In most surveys, the use of probability sampling methods is recommended.These methods include (see Annex 1 for detailed definition of these terms):

- Simple random sampling
- Systematic sampling
- Cluster sampling
- Stratified sampling

Sample size:

Sample size is central to the design of a sample. There are several factors that must be considered in determining the sample size. These include key estimates desired, target population, desired level of precision for key indicators, estimation domains, whether measuring level or change, clustering effect, allowance for non-response, and available budget.

Sample weights:

Household surveys are based on complex sample designs. The resulting samples have limitations, such as selection of units with unequal probabilities, non-coverage of the population, and non-response. This might lead to bias and inference challenges between the sample and target population. Sample weights are needed to correct these limitations and thus derive appropriate estimates of characteristics of interest.

Sample weighting is used to:

- Compensate for unequal probabilities of selection.
- Compensate for (unit) non-response.
- Adjust the weighted sample distribution for key variables of interest (for example, age, race, and sex) to make it conform to a known population distribution.

Sample documentation:

The sample documentation should describe in detail all procedures and include the following:

- Target population
- Expected sample size
- Sampling frame
- Stratification
- Sampling procedures
- Household listing results
- Sampling weights
- Results of the survey implementation and response rates
- Sampling errors and limitations of survey implementation.

Additional Resources

- Sampling frames and master samples
- Designing Household Survey Samples: Practical Guidelines

Chapter 5: Data collection

This chapter discusses the best practices in collecting CGD and processing including recruitment and training of field staff, testing data application systems, informed consent, data collection, data quality control, data editing, and coding.

The basis for good quality data is good data collection. Data collection is the process of gathering information on defined variables of interest in a systematic manner to answer research questions and test hypotheses. Accurate data collection is critical in maintaining the integrity of the research process. It is therefore important that appropriate data collection instruments are selected, and their correct way of use clearly stated to reduce the errors. Following a detailed plan for the collection of data, data may be obtained from individuals or organizations or indirectly from other sources. Proper procedures should be employed to encourage participation in the data collection process and to improve the quality of the data.

CGD can be obtained through research, social audits, crowd-sourcing online platforms, mobile phone and text surveys, phone calls, reports, storytelling, social media, and community radio. The data collection method(s) can be interviewer assisted, self-administered, or a combination of the two. Interviewer-assisted methods involve interviewers administering questionnaires to selected respondents by clarifying question wording and by probing for deeper responses. Such methods include face-to-face and telephone interviewing. In self-administered methods, respondents complete the questionnaire themselves without an interviewer. An example of the self-administered method is self-administered paper questionnaires or interactive voice recording. A combination of both interviewer-assisted and self-administered methods is another approach.

5.1 Recruitment and training

5.1.1 Recruitment

It is essential to recruit personnel for data collection who have the appropriate abilities and personal attributes. Important factors to consider when hiring personnel include academic qualifications, interpersonal skills, fluency in local languages, organizational skills integrity, etc. Familiarity of the local area may also be a factor to consider in hiring because it taps into the local knowledge and capacity.

5.1.2 Training of personnel

Training enables personnel to perform their duties effectively, enhancing the generation of high-quality data. The purpose of training is to:

- Clarify the study's rationale and study protocol.
- Ensure the study materials are administered in a standardized manner.
- Give practical suggestions on the data collection process, such as ethical issues.
- Explain duties and the way they are supposed to be undertaken.

5.1.2.1 Training of trainers

Training of trainers is a learning process intended to engage subject matter specialists in teaching other trainers for the study. The training should be delivered consistently to all the trainers.

5.1.2.2 Training of field personnel

The personnel to be involved in actual data collection must be well trained. They should be trained to administer the questionnaire and ensure that the targeted respondents participate. Different training methods can be used, including home study exercises, classroom training sessions, and on-the-job training.

Topics likely to be covered in training field personnel include:

- Confidentiality
- Personnel role
- Receiving, checking, and accounting for material
- Definition of terms
- Sequence guides
- Procedure at the doorstep
- Procedure for interviewing
- Handling unresponsive respondents
- Training on technological systems and procedures
- Checking and editing completed material
- Health and safety

5.2 Testing the data collection process

Data collection forms or questionnaires should be systematically tested before starting data collection. Testing of the data collection instruments is an important step in data collection systems. Testing procedures are put in place to address all issues that relate to the questionnaires and the accompanying supporting infrastructure. This ensures that the data collection exercise is effective, and the resultant data is of high quality. Challenges identified relating to the programmed questionnaire and/or the supporting infrastructure should be addressed as early as possible. In addition, it is necessary to have a test plan that covers the specification of the testing subject, necessary resources for its implementation, testing process and its scope, and the testing schedule and products.

5.2.1 Pre-field tests

Pre-field tests are usually not carried out in the field but are used in the initial stages of questionnaire testing. The tests are also important because they test the effectiveness of the instrument while it is still under development. Pre-testing can assess respondents' ability to understand the questions and respond accurately. It can also assess the flow of the questions.

5.2.2 Pilot surveys

Pilot surveys involve testing the data collection instruments, such as a questionnaire using a small sample of intended respondents in a similar manner to the actual survey. These surveys are important in identifying problems before the actual survey begins. They may also collect useful information, such as the length of the interview time, adequacy of the questions, and comparisons with different versions of the questionnaires.

5.3 Informed consent in data collection

Informed consent is an inherent characteristic of responsible data collection practices. It is the process where participants are informed about a study and they can willingly choose whether they wish to participate or not. In the era of greater awareness of data misuse, clear explanations are needed for what the data is going to be used for. Legal requirements are built around ethical standards, respect of persons, and the moral authority to safeguard the vulnerable from harm and exploitation. In many forms of data collection, such as surveys and interviews, informed consent is the basis of ethical research.

TOP TIP

Understanding informed consent

- **Information**: Participants must be notified of all the relevant information, including risks, benefits, and data use and protection.
- **Understanding**: Participants should be made to fully understand the information by means of clear communication.
- **Volunteering**: People should not be coerced or manipulated into participating in any data exercises.
- **Decision-making capacity**: Respondents decide whether they want to participate or not through assessing risks and benefits.

TOP TIP

Best practices for seeking consent

- There should be a statement detailing that the individual provides the data voluntarily and consent can be withdrawn at any time.
- There should be a statement that indicates that failure to provide data will not lead to denial of services or affect the ability to receive the services being delivered.
- The purpose of the data should be explained in a language that can be clearly understood.
- Information should be provided on what data is being collected, why it is being collected, who it will be shared with, and what part of the data will be shared and with whom. This should be explained in a clear, understandable language.
- Information should be provided on how and for how long the data will be stored.
- A description should be provided of predicted benefits to the community or the individual.
- A sensible description should be provided of anticipated harm to individuals and communities in case privacy is contravened and how such risks will be mitigated.

Extra caution is needed in some social-cultural, legal, political, and religious contexts that may significantly influence consent. These may include cultural differences between those who obtain consent and those who give it or sensitive communities and/or groups that are at risk due to their demographic or health characteristics, such as those with HIV.

5.4. Organizing and supervising data collection

When data collection activities are properly supervised, it significantly contributes to high-quality statistics. Supervision in the field can include such things as accompanying, random spot checks, auditing (back checks), scrutiny, and non-response analysis. Random spot checks require supervisors to make frequent, unannounced visits to the survey locations. Back checks involve revisiting some respondents to ask them a few questions from the questionnaire and comparing their responses with the original responses. This offers a second set of responses to check the quality of enumerators' work and the reliability of the data. Scrutiny is where a supervisor checks for blank fields, inconsistent answers, or responses that are inaccurately categorized. This process may also be automated using software.

5.5 Quality control during data collection

Quality control is the use of procedures to ensure that data being collected is accurate and of high quality. This aligns with the UN NQAF level C on managing statistical processes. Procedures should be put in place to monitor the quality of statistical production processes. Quality control is used for questionnaires, IT infrastructure, management systems, interviewing, and other processes of data collection that may affect data quality. Data quality control must be done at all stages of the data collection process.

Quality control can be achieved through:

- Proper design of the data collection instruments, which is the first stage of quality control and is central to the data collection process and testing of these tools.
- Recruitment of competent enumerators and their proper training.
- Data quality checks on collected data.
- Management and sharing of metadata and documentation of methods and different statistical processes throughout the processes, as appropriate.

TOP TIP

Guidelines for ensuring data protection during data collection

- The methods and procedures of data collection, use, and dissemination should be done through lawful, legitimate, and fair means.
- Any form of interaction with the data should be governed by laws, moral and ethical conduct, and uttermost confidentiality.
- Survey data that can be identified individually should be protected and information anonymized.
- Encryption during data collection.
- Procedures should be established to protect confidential data.
- Access to data should be controlled. Only authorized people should be allowed to access, write, and read the data.

More information on data protection is outlined in Chapter 9.

Additional Resources

- Handbook on the Management of Population and Housing Censuses
- UN National Quality Assurance Framework (NQAF) Manual for official statistics

Hands in action recording data at Nairobi West Prison Clinic. Credit: Elphas Ngugi

Chapter 6: Data processing and analysis

This chapter discusses data processing and data analysis. Data processing describes the process of inputting data in preparation for analysis. Additionally, it includes the process of calculating weights and aggregates. The output for this process is the finalized data files that can be used in analysis. On the other hand, data analysis involves preparation of draft outputs, validation, interpretations, explanation, and finalization of outputs. Moreover, data analysis includes application of disclosure control.

6.1 Data processing

The data processing phase includes inputting data and preparing it for analysis. It is made up of eight sub-processes that classify, integrate, check, clean, and transform data so that it can be analyzed and disseminated in statistical formats. The process phase and the analyze phase may run together where analysis reveals aspects of data that may require further processing.

This phase is divided into five sub-processes:

1. **Integrate data** - Data from one or more sources is integrated. The results from the collection phase are combined here. Data integrations involves combining data from multiple sources, combining statistical data with other data such as geospatial and non-statistical data, data pooling, matching or linking records, data fusion and prioritizing.

2. **Classify and code** - Data is classified and coded in this sub-process.

3. **Review and validate** - Examining data to reveal problems, errors, and discrepancies. It is also referred to as input data validation.

4. **Edit and impute** - New values may be inserted or outdated data may be completely removed in this sub-process. A rule-based approach is employed in this sub-process. The steps include, determining if data should be added or removed, selection of method to use, adding/changing the values, writing in new data, producing metadata for this process.

5. **Derive new variables and units** - Variables and units for data are derived in cases where they are not explicitly provided. Arithmetic formulae are used to derive new variables.

6. **Calculate weights** - Weights are created for units of data records based on the methodology developed in the design phase.

7. **Calculate aggregates** - Aggregated data and population totals from microdata are created in this sub-process. Data is summed for records having certain characteristics, measures of average and dispersion are determined and weights are applied from the sub-processes.

8. **Finalize data files** - The results of all other sub-processes in the process phase are brought together here in a data file.

6.2 Data analysis

During the analysis phase, statistical outputs are produced and examined in detail. In this phase there are also some processes and activities that help analysts to understand the data and the statistics produced. The output of this phase can be used as an input for other sub-processes.

The following are sub-processes of the data analysis phase:

1. **Prepare draft outputs** - Data is transformed into statistical output such as indexes seasonally adjusted statistics and coefficients of variation.

2. **Validate outputs** - The quality of the outputs that is produced is validated. Validation may include, performing macro-editing, investigating inconsistencies in the statistics, checking for consistency in the data, and checking for the presence of metadata guidelines.

3. **Interpret and explain outputs** - This is where a deeper understanding of the outputs of the data is gained by the statisticians. This understanding is thereby used to interpret and explain the statistics.

4. **Apply disclosure control** – This involves ensuring that data meant for dissemination does not go against the rules on confidentiality based on rules and or policies of the organization.

5. **Finalize outputs** - The statistics are evaluated, and checks are made to ensure that they are of the desired quality and thus ready for use.

6.2.1 Data editing

Data editing is the process of detecting and correcting data errors. It involves checks to identify missing, invalid, or inconsistent records, and flagging data records that are potentially erroneous. It uses available information and assumptions to impute values for inconsistent values in a data file. The development of edit rules relies largely on results from analysis of data and input from subject matter specialists. Therefore, it is advisable to use these two approaches in developing the editing rules and parameters. Proper documentation of editing rules is critical.

TOP TIP

Guidelines for data editing

- Edit rules should be developed by staff with expertise in the subject matter, questionnaire design, and data analysis.

- Editing should be done at several stages (collection, processing, and analysis).

- Edits performed at each stage should not contradict edits at other stages. In other words, edits done in collection and processing should be consistent with each other.

- Editing should be used to provide information about the survey process, either in quality measures for the current survey or to recommend improvements for future surveys.

- Information on the types of edits performed and the impact of editing on the survey data should be well documented and be part of metadata.

- Quality assurance and control procedures should be implemented to minimize and correct errors because of editing.

Imputation is a technique to determine and replace values to resolve problems of missing, invalid, or inconsistent data. It is achieved by changing some of the responses and all the missing values on the record being edited to ensure that the record is consistent. Imputation is generally used to ensure that key variables, such as gender, age, and marital status, have no missing values. Imputation is a powerful tool in handling problems that cannot be resolved either by contacting a respondent or by manually studying the questionnaire.

6.2.2 Data coding

It is important to add codes to collected data to identify aspects of data quality, such as missing data. This will allow users to appropriately analyze the data. Data coding can either be manual or automated. In manual coding, the coder reads, interprets, and manually converts a written response to an open-ended question into a numeric code. Alternatively, numeric codes can be assigned through an automated process.

TOP TIP

Guidelines for data coding

- Use codes that clearly identify missing data and cases, where an entry is not expected (e.g., skipped over by skip pattern).
- Avoid the use of blanks and zeros as codes to identify missing data, since they tend to be confused with actual data.
- Use standardized codes if they exist when converting text data to codes to facilitate easier analysis.
- Create a quality assurance process when using manual coding to convert text to codes. This minimizes errors due to manual coding and therefore maintains data quality standards.

6.2.3 Data analysis

Developing a plan to analyze data before the start of a specific analysis to ensure that appropriate statistical analysis is used and there are adequate resources to complete the analysis. It is advisable to prepare a preliminary set of proposed tabulations and other desired statistical outputs. The analysis plan determines results relevant to the study objectives. It helps limit the analysis to what is needed rather than dozens of irrelevant analysis tables. The proposed tabulations should state each variable to be presented in a table and its categories. Where the data will be presented in a dashboard, the variables and visualizations can be pre-designed.

6.2.4 Turning data into information

Data analysis is the process of transforming raw data into statistics and statistics into actionable information. This information may be presented in numerous forms, such as numbers, tables and graphics, and dashboards. The information depicts trends or patterns in the data and is a solid foundation of data-driven policymaking. Data analysis involves organizing, summarizing, and interpreting the data in a way that provides clear answers to policy-relevant questions.

Good practice requires that data be disaggregated by income, sex, age, education level, race, ethnicity, economic status, geographical location, disability, or other characteristics. There are internationally agreed dimensions and categories of disaggregation and a defined minimum disaggregation set. Dimensions are the characteristics by which data are disaggregated (such as age, sex, and level of education) whereas categories are the different characteristics under a certain disaggregation dimension (such as male/female under the sex dimension).

Additional Resources

- Data Disaggregation and SDG Indicators: Policy Priorities and Current and Future Disaggregation Plans.

Chapter 7: Metadata

This chapter covers metadata as a good practice in dissemination and documentation of CGD. Often data collected may be released to the public for use, hence the need to have proper documentation to help users easily understand the data.[8]

Metadata provides additional information to users about the data file. Detailed metadata is needed for the data to be appropriately used and accurately interpreted. Metadata provides information on data collection, file format, sampling design, unit of analysis, relationships among records, reference period, aggregation of records, restrictions on the use of the data, indicators of data quality and names, and definitions of all variables on the file, including derived variables that are essential for replicating the key survey outputs.

Metadata consists of the following:

- Structural metadata provides information on the structure of data sets, such as the nature of files (tabular or otherwise), variable names, and hierarchical relationships.
- Reference or descriptive metadata gives more general information, from single data values to the whole data collection process. It can be categorized as:
- Conceptual metadata that describes the different concepts used in the study.
- Methodological metadata that details methods used in the data production (e.g., sampling, data collection methods, and data editing processes).
- Quality metadata that describes the different quality dimensions of the resulting statistics (e.g., timeliness and accuracy).

Metadata has the following benefits:

- It is a crucial element in the dissemination of all statistics. Metadata equates to transparency, which aligns with the UN Fundamental Principles of Official Statistics. From the metadata, end users can make informed decisions on the data's relevance to their purpose.
- Metadata allows integration of individual statistical areas, such as social, economic, and demographic data.
- In the internet era, where data is available from online resources, metadata helps end users of statistical data to interpret, analyze, and understand the data as well as retrieve additional resources, if possible.

Additional Resources

- Dublin Core Metadata Initiative
- Metadata registries (ISO/IEC 11179)
- Common warehouse metamodel (CWM) specification
- Enhanced General Data Dissemination System-e-GDDS
- Geographic information metadata (ISO 19115-1)
- Statistical Data and Metadata Exchange (SDMX; ISO 17369)

[8] Dissemination also includes sharing of various knowledge products generated from the data collected, such as briefs, reports, fact sheets, etc.

Chapter 8: Evaluation

Evaluation takes place at the end of a data generation process or can be done on an ongoing basis as per the organization's policy or rules. Evaluation relies on all the inputs collected during all the different phases of the data production process. Instances within this process are evaluated for success based on the quantitative and qualitative inputs, identifying weaknesses and making improvements where necessary. Where data is produced regularly, evaluation should be done by assessing each recurrence to determine if future recurrence should be done, and whether improvements are required. However, for well-established processes, evaluation is not always done for each recurrence of the process. In such instances, phases are viewed as providing decisions about the next phase. For example, whether to specify needs again or start from another phase such as the collection phase. The evaluation phase consists of three sub-processes: gathering evaluation inputs, conducting the evaluation, and development of an actual plan.

8.1 Gathering evaluation inputs

Evaluation material takes many forms. These include feedback from the users, metadata from the processing phase, metrics of the system and suggestions made by the staff working on the data. Progress from an action plan outlined in a certain phase may form the basis for input for evaluation in another phase. In this sub-process, all the inputs gathered are compiled into quality indicators which are then availed to the team responsible for evaluation. Material collection should be automated and continuous throughout the process, as defined by the quality framework. However, for some processes, it becomes necessary to carry out some activities such as small surveys to determine the extent of the success of a process. For example, post-enumeration surveys which are run after a survey to determine how many people were not counted or how many may have been counted more than once.

8.2 Conducting an evaluation and developing an action plan

Inputs in the evaluation process are analyzed and compared to the target results or the expected results if they are available. They are then synthesized into an evaluation report. This can be done at the end of the statistical business process or for certain activities, it can be done throughout the process, fixing issues and making improvements where required. The report herein generated should make note of all the issues such as the quality and highlight all the deviations that do not match the expected or target outputs. Appropriate recommendations should be made for all these issues. These may include changes to be made to specific phases or the omission of some of the processes all together.

An action plan is agreed upon based on the findings of the evaluation report by bringing together all the people responsible for decision making. It should include a way of monitoring the impacts of the actions to be taken which may provide an input for subsequent evaluations.

Important issues to consider in the evaluation phase:

- Ongoing statistical activities should be reviewed periodically and adapted according to changing needs in the survey and data analysis program.
- Get feedback from the data users about the quality and relevance of the data that is produced and put their suggestions into consideration.
- Quality issues should be prioritized based on their overall impact on the aggregate data. Focus should be made on the issues with the greatest impact.
- Programs/phases that can be made more efficient should be determined for subsequent iterations.
- User feedback that relates to data accessibility should be reviewed. Possible improvements should be made to improve the user experience.
- Survey practices should be pitched against best practices from other statistical organizations, following quality protocols that will improve comparability of the outputs with those from other organizations.
- The need for change should be balanced with the need for consistency.

Additional Resources

Generic Statistical Business Process Model

Chapter 9: Data privacy and protection

Data privacy and protection are closely interconnected. Data privacy defines who has access to data and who defines the access, while data protection is about securing the data from unauthorized access. The chapter first discusses the need for data protection, then offers some guidelines on data protection and privacy.

9.1 Legal frameworks on data protection internationally and in Kenya

The European General Data Protection Regulation (GDPR) is one of the international legal frameworks governing consent, privacy, and data protection in the European region. The regulation offers provisions and requirements connected to processing and handling of individuals' personal data. It requires organizations to ensure personal data is collected lawfully and under strict conditions. In addition, organizations are obliged to protect data from both misuse and exploitation.

In Kenya, the production and usage of statistics is governed by the 2006 Statistics Act and the 2019 Statistics (Amendment) Act. The Act stipulates establishment, objectives, and functions of KNBS. KNBS is mandated with collection, compilation, analysis, publication, and dissemination of statistical information. In addition, it is responsible for coordination of the national statistical system and related purposes (GoK, 2006; GoK, 2019).

The 2019 Data Protection Act governs data collected from persons. The act offers both legal and institutional framework for the protection of personal data. The legislation outlines the rights of data subjects and obligations of data controllers, data processors, and third parties who handle personal data. The Act regulates collection, processing, storage, and transfer of personal data with the aim of guaranteeing the right of an individual to privacy and their information and/or their family be protected from unauthorized persons (GoK, 2019).

The Data Protection Act seeks to:

A. Regulate the processing of personal data.

B. Ensure that the processing of personal data of a data subject is guided by the principles set out in the act.

C. Protect the privacy of individuals.

D. Establish the legal and institutional mechanism to protect personal data.

E. Provide data subjects with rights and remedies to protect their personal data from processing that is not in accordance with the act.

9.2 Need for data protection

The discretionary and unregulated use of data has raised genuine concerns among the general population about the control and use of such data. CGD mirrors the same concerns over privacy and protection as other types of data. The vast amount of personal data collected daily by different organizations requires carefully considering how such data is managed. Data protection is a crucial element in maintaining public trust among the data collection entities. Further, the right to privacy is enshrined in the Kenyan constitution. The recognition of privacy as a key human right internationally makes protection and privacy of data essential. Therefore, it is critical to adhere to certain guidelines in ensuring protection of personal CGD.

CGD, like any other data, should be safeguarded against loss, compromise, or corruption. Data protection laws are aimed at shielding personal data from unauthorized access. For citizens and users to have confidence in agencies that collect data, the best data protection practices should be combined with effective legislation to mitigate against data exploitation. Such concerns have led to the development of data protection principles. These are important starting points for putting in place proper safeguards to keep data protected.

The need to respect privacy goes beyond acknowledging that it is the respondents who "own" the information about themselves. It also includes the need for individuals to decide what information should be disclosed, to whom it should be disclosed, and when it should be disclosed. Examples of personal data include name, home address, email address, identification card number, location data (for example, the location data function on a mobile phone), and Internet Protocol (IP) address.

9.3 Guidelines for data protection

The following guidelines for ensuring privacy and the protection of CGD are informed by the UN Fundamental Principles of Official Statistics (UNSD, 2014) and Kenya's domestication of the principles as outlined in the fourth schedule of the Statistics (Amendment) Act, 2019 (GoK, 2019).

- All legal limits to protect privacy in the jurisdiction where CGD is being collected should be observed.
- Data should only be disclosed, used, or retained for the original purpose (i.e., the purpose at the time of collection). Where data is to be reused, privacy protection should be factored in.
- Precautionary measures should be taken to protect the data from modification, loss, destruction, disclosure, or unauthorized access, such as defining who has access to data and what level of access is granted.
- Data should not be processed in secrecy. Individuals should be made aware of the data processing and who is responsible for its processing. For example, organizations collecting data must be solely responsible for its handling, processing, analysis, storage, and dissemination. The inclusion of other parties in the handling of such data should be made public.
- A valid explanation should always be provided to respondents where data that touches on family and/or private affairs is collected. For instance, when doing a study on poverty levels aimed at improving community livelihoods, it should be explained why the source of income is an important aspect of the survey.
- Inaccurate personal data should be deleted or corrected immediately. A description should always be provided of measures to be taken to maintain data integrity and privacy. For instance, it is important to notify the respondents that their personal information will remain confidential and will not be disclosed to anyone.

9.4 Confidentiality

Confidentiality involves limiting data access and disclosure to authorized users and preventing access by or disclosure to unauthorized ones. Confidentiality is also related to the broader concept of data privacy: limiting access to individuals' personal information. Principle 6 of the United Nations Fundamental Principles of Official Statistics relates to confidentiality (Chapter 1 Table 3): "Individual data collected by statistical agencies for statistical compilation, they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes."

Good practices include the following:

- Implementing measures to prevent either direct or indirect disclosure of data on persons, households, individual respondents, etc.
- Developing a framework describing methods and procedures that adhere to confidentiality requirements for users (such as researchers) interested in further analysis.

9.5 Data anonymization

Data anonymization is the process of removing or modifying the identifying variables in the dataset. Identifying variables can either be direct identifiers, such as names, addresses, or registration numbers, or indirect identifiers, which are characteristics that may be shared by several respondents, and whose combination could lead to the re-identification of one of them. Anonymization is a principle of data protection and privacy since it minimizes the risk of identifying the statistical units.

The following are two major steps in anonymizing the data:

- Determine variables that are potential identifiers.
- Modify these variables' level of precision to minimize the risk of re-identification to an acceptable level.

Data anonymization techniques include:

- Data reduction: This is a form of aggregation that condenses data into more meaningful summaries that are easier to communicate or understand and that yields descriptive statistics.
- Data perturbation: This involves adding random "noise" to confidential, numerical attributes, thereby protecting the original data.

Additional Resources

- Responsible Data for Children
- Data Ethics Canvas
- A handbook on data protection in humanitarian action
- Privacy by Design: The 7 Foundational Principles

Conclusion

This practical guide aims to enable non-state actors to produce high-quality CGD that meets the quality standards documented by NSOs such as KNBS and international standards.

With continued collaboration and trust-building among actors, the value of CGD to complement and supplement official statistics will continue to build.

To improve CGD in a context like Kenya, we recommend the following:

1. **Collaboration between stakeholders**: There should be collaboration between multiple stakeholders, NSOs, and other data producers to enable sharing of lessons, methodologies, and templates where possible. This could be through regular convening of the producers of CGD and users beyond the National Statistical System (NSS) to allow conversations that lead to constructive collaboration. Producers of CGD should also collaborate to ensure data collection is efficient and filling data gaps instead of duplicating efforts. It is recommended that a multistakeholder technical working group on CGD is established with leadership of KNBS and CSOs.

2. **Encouraging government involvement**: Producers of CGD should ensure the involvement of KNBS as the government agency tasked with developing standard guidelines for official data in the production of CGD.

3. **Standardizing practice and strengthening capacity within institutions**. Producers of CGD (and other data producers of data who are not from civil society) should adopt these guidelines as standard practice for producing quality data. This should be used while also applying existing government guidelines and standards from KNBS (KeNQAF) and the United Nations Fundamental Principles of Official Statistics.

   Where producers of CGD may not be able to follow the guidelines entirely, we have produced a checklist in Annex 2 to make this process faster and achievable. Individual organizations can use the guidelines, but the use of umbrella bodies such as the SDG Kenya forum or the NGO Board can further promote wider use.

   Not all CGD is meant to be applied for official purposes or government statistics. As such the use of these guidelines does not qualify it to be official statistics but improves the quality of the data generated. Institutions should strengthen internal data capacity and skills in data production to enable the use of these guidelines.

4. **Activities to institutionalize CGD**: data gaps assessment, quality criteria and checklists guided by KNBS. Through a collaborative effort, KNBS should further develop criteria and checklists for assessing the quality of all non-official data produced in the country. These tools should align with the Kenya National Quality Assurance Framework that is currently under development. KNBS should also lead a detailed inventory of CGD approaches and the SDGs data gaps it fills organized by priority and which CGD can be used to fill the gaps. This can align with existing sector technical working groups, but also through establishing the multi-stakeholder technical working group for CGD. KNBS can borrow from other NSOs that have gone through this process such as Statistics Canada and the Philippine Statistics Authority (Statistics Canada, 2017; PARIS21, 2020).

Annex 1: Definitions and concepts

Below is a series of definitions relating to the process of data collection:

Population is the entire group of items or individuals of interest in a study at a time and area.

Sample is a fraction of items or individuals selected from the population with the intention of studying or investigating them to approximate or estimate information about the population.

EXAMPLE

Suppose KNBS is conducting an investigation or study to determine the unemployment rate in Kenya for people aged 25–50 years in 2019. Given that the agency has limited resources for this study, it decides to select 1,000 individuals of this age group from each of the 47 counties in Kenya.

From this hypothetical example, every person who was in Kenya in 2019 and aged 25–50 years forms the population, whereas the 1,000 individuals selected from the 47 counties make up the sample drawn from this population. A sample is therefore used to generalize the characteristics of a population.

A **target population** is the entire group of items or individuals of interest for which the information is wanted, and estimates required.

Parameter is a number that describes population characteristics. **Statistics** is a number that describes sample characteristics.

EXAMPLE

In 2019, KNBS conducted a complete count of all individuals living in Kenya. The results indicated that the total population of Kenya was 47,564,300. This number describes a characteristic of the Kenyan population—the total—which is an example of a population parameter.

It is not always feasible to obtain a parameter because of various reasons, such as cost. As a result, parameters are often approximated or estimated using statistics.

Suppose a study is conducted to determine the unemployment rate in Kenya in 2019 among individuals aged 25–35 years. Suppose the study is done on a sample of 4,000 individuals and the rate computed is 9.3 percent. This number describes a characteristic of a sample. The characteristic being described is the rate and is an example of statistics.

A **census** is the collection of information from entire items or individuals of interest in a study at a time and area. A **sample survey** is the collection of information from a fraction of items or individuals selected from the population to approximate or estimate information about the population.

EXAMPLE

In 2019, KNBS conducted a complete count of all individuals living in Kenya and collected information by asking questions to all individuals living in Kenya. This process is called a census.

Suppose the Government of Kenya seeks to collect the opinion of Kenyans on the housing project it intends to carry out. It then selects 1,000 individuals from each of the 47 counties and collects their opinions on the housing project. The process of collecting opinions from the selected individuals who represent all Kenyans is an example of a sample survey.

Sampling is the process of selecting a fraction of items or individuals from a population. The selected items or individuals are used to obtain information that approximates the population characteristics. Sampling methods can either be **probabilistic** (a scientific method of selecting a fraction of items or individuals from the population in a random manner such that each item in the population has a known non-zero probability of being selected) or **non-probabilistic** (a non-scientific method of selecting a fraction of items or individuals from the population using subjective or personal experience).

Sample size is the number of items or individuals to be selected from the population to constitute the sample.

Sampling frame is a complete list of unique and identifiable items or individuals in the population. This list is used for selecting samples.

Sampling methods

Simple random sampling: Simple random sampling (SRS) is a probability sample selection method where each element of the population has an equal chance/probability of selection. Selection of the sample can be with or without replacement. This method is rarely used in large-scale household surveys because it is costly in terms of listing and travel. It can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. SRS is attractive by virtue of its being simple in terms of selection and estimation procedures (for example, of sampling errors).

Systematic sampling: Systematic sampling is a probability sample selection method in which the sample is obtained by selecting every kth element of the population where k is an integer greater than 1. The first number of the sample must be selected randomly from within the first k elements. The selection is made from an ordered list. This is a popular method of selection especially when units are many and are serially numbered from 1 to N. The sample comprises the first unit selected randomly and every kth unit, until the required sample size is obtained.

The interval k divides the population into clusters or groups. In this method, one cluster of units is selected with probability 1/k. Since the first number is drawn at random from 1 to k, each unit in the supposedly equal clusters has the same probability of selection, l/k.

Cluster sampling: In cluster sampling, the higher-level units of selection, for example, enumeration areas contain more than one elementary unit. In this case, the sampling unit is the cluster. For example, a simple method of selecting a random sample of households in a city could entail having a list of all households. This might not be possible, as, in practice, there may be no complete frame of all households in the city. In order to go around this problem, clusters in the form of blocks could be formed. Then a sample of blocks could be selected, subsequently a list of households in the selected blocks could be created. If need be, from each block, a sample of households, say, 10 per cent, could be drawn. This is what is called cluster sampling.

Stratified sampling: In the stratified sampling method, the sampling units in the population are divided into groups called strata. Stratification is usually carried out so that the population is subdivided into heterogeneous groups that are internally homogeneous. In general, when sampling units are homogeneous with respect to the auxiliary variable, termed the stratification variable, the variability of strata estimators is usually reduced. It should also be noted that there is considerable flexibility in stratification, in the sense that the sampling and estimation procedures can be different from stratum to stratum.

Multi-stage sampling: This is a two-stage sample design used for household surveys. The first stage is stratification of the sample frame by regions and by urban-rural classification using the most recent census of population and housing. The primary sampling units are selected systematically according to probability proportional to size within each stratum. The second stage involves selection of households from each primary sampling unit based on the household listing. Therefore, there is a need to ensure the sampling frame is updated and reliable.

EXAMPLE

Suppose the Ministry of Health in Kenya wishes to study the prevalence of pregnancy among girls under the age of 18 years living in the informal settlement, Kibra. The Ministry decides to select 3,000 households as a sample from which to conduct a survey. The process of selecting the 3,000 households from all the households in Kibra is called sampling, while the number 3,000 is referred to as the sample size. The Ministry requests KNBS to provide a list of all households in Kibra based on the 2019 census. The list provided has 8,000 households. This list of all households in Kibra is an example of a sampling frame.

The Ministry of Health then writes down the unique numbers representing all households on small pieces of paper. These small pieces of paper are then put in a basket and mixed up thoroughly. Then, 3,000 Ministry employees are asked to pick only one piece of paper from the basket, one employee at a time. The process of selecting 3,000 households is an example of simple random sampling. This is because each household will have 3/8 (the sample size of 3,000 divided by the population size of 8,000) chance of being selected.

In this fictitious example, suppose the Ministry decides to group the list of households provided by KNBS based on the house structure. After grouping, two lists are created: one list contains households whose wall structure is made of stone, while the other list contains households whose wall structure is made of materials other than stone. These two groups are an example of strata.

Now suppose the Ministry writes down the unique household numbers for each list on small pieces of paper. These small pieces of paper are then placed in two separate baskets based on group and mixed up thoroughly. The Ministry then splits its employees into two groups each of 1,500. The employees then select one piece of paper at a time. This whole process of sample selection is an example of stratified sampling.

A **household** is an individual or group of individuals who live in, compound, or homestead; share cooking arrangements; and answerable to a common household head.

A **head of household** is a person who is in charge of making day-to-day decisions in a household. The person’s authority is respected by all members of the household. The person can be a father, a mother, grandparent, a child, etc.

A **household member** is a person who lives in the household and is either present or temporarily absent from the household for a period less than six months, as at the time the survey is being conducted.

Annex 2: Checklist for improving the quality of citizen-generated data

Study objectives

[ ] State the problem in broad tasks.
[ ] Identify the uses of the data.
[ ] Create an analysis plan for analyzing and presenting data.
[ ] Identify specific themes to be covered by the study.

Specifying needs

[ ] Ensure that important users and stakeholders are involved.
[ ] Elicit feedback from the group being surveyed.
[ ] Review available data to see if it meets user's needs and the condition under which it can be made available.

Design

[ ] Design questionnaire using a clear and concise language
[ ] Test possible frames for suitability and quality
[ ] Develop and test the imputation method before implementation

Data collection

[ ] Engage different stakeholders, e.g., KNBS.
[ ] Contact local authorities for support and establish a local presence.
[ ] Engage sectoral players for collaboration.
[ ] Seek informed consent from respondents.
[ ] Recruit field staff and train the trainers.

Processing Phase

[ ] Control data quality through survey management, data capture, data review, and data adjustment.
[ ] Establish a procedure for information protection for confidential data.
[ ] Ensure adherence to legal frameworks and policies.
[ ] Identify, analyze and correct extreme values to avoid detrimental impact on survey estimates.

Analysis Phase

[ ] Check for potential errors and validate data before release
[ ] Add codes to data to identify aspects of data quality information.
[ ] Develop an analysis plan for analyzing the data.
[ ] Present data in a clear, easy-to-understand format.
[ ] Determine whether the data are "fit for use".

Documentation and dissemination (and factor in data privacy)

[ ] Provide metadata for the data.
[ ] Ensure confidentiality.
[ ] Anonymize the data.
[ ] Observe privacy and data protection.

Evaluation

[ ] Review ongoing statistical activities periodically and adapt to evolving needs.
[ ] Gather feedback from data users about the relevance and quality of the data
[ ] Determine if any processes in the survey program could be made more efficient

Annex 3: Methodological note

This practical guide was developed using a human-centered design approach.[9] Stakeholders were given an opportunity to share thoughts on areas they believe the guide should cover. Below is an outline of the methodology.

Literature review

Relevant information on international and national standards and principles for both official and non-official statistics was reviewed. This was critical in developing this guide. Such information included the fundamental principles of official statistics and KNBS documents on data quality assurance, among others. In addition, the current legal framework for statistics in Kenya was analyzed and case studies of CGD work implemented in different contexts, both locally and internationally, were reviewed with the aim of identifying lessons learned.

Workshops, regional and global events

Co-creation workshops guided by the human-centered design approach were organized. The workshops' aim was to engage and gain inputs from different stakeholders and identify the areas of priority for this guide. CSOs such as Open Institute, Twaweza, Africa's Voices Foundation, and Map Kibera shared lessons from their previous CGD initiatives. Side events at Data Tamasha and the United Nations Statistical Commission were also hosted, bringing together NSOs and civil society to discuss ways of building trust in CGD.

Survey

An online survey was sent to CSOs to gather information on their data-driven initiatives, the barriers they face and potential solutions to these barriers.

Key informant interviews

KNBS and civil society organizations that have been involved in producing CGD were interviewed. KNBS is the constitutionally mandated body that produces official statistics in Kenya, the need to have active and consistent engagement throughout the process. CSOs are both users and producers of CGD, hence the need to seek their input throughout the process.

Meetings with various stakeholders

The existing works in CGD were mapped in order to identify any gaps and areas for cooperation between stakeholders. Consequently, meetings with various organizations such as UN Women, and PARIS21 were held.

Workshops with KNBS team to review the guide

In the last quarter of 2020, KNBS team was part of a five-day workshop to have an in-depth review of the guidelines and benefit from their expertise. This was an opportunity to discuss with KNBS on the recommendations and how to institutionalize CGD as well as a review mechanism to update the guidelines (Annex 4).

[9] This is an approach to creating solutions for problems and opportunities by emphasizing the needs, contexts, behaviors, and emotions of the people whom the solutions target.

Annex 4: A review mechanism to update the guidelines

1. It is recognized that changes may be required to the guidelines. Consequently, it would be necessary to institute a revision mechanism that accommodates change mechanisms either on an ad-hoc basis or periodically.

2. Change requests may be triggered from within KNBS or from external stakeholders as needs or issues emerge. This should outline:
   a. What revision is desired

   b. Justification or rationale for the revision

   c. Desired timelines for the revision

3. This will trigger internal processes which may review the request. This may entail convening subjects matter experts (if the issues are sectoral, they can be advised from technical working groups; if they involve broader practice, a meeting of the proposed CGD technical working group can be convened).

4. The periodical review of the guidelines is set at 12 months (from date of launch of the guidelines).

References

CIVICUS. (2015). Citizen-Generated data and governments- towards a collaborative model. Civicus. Retrieved from http://civicus.org/images/citizen-generated%20data%20and%20governments.pdf

GoK. (2006). Statistics Act. Kenya. Retrieved from kenyalaw.org/kl/fileadmin/pdfdownloads/Acts/StatisticsAct_Cap112.pdf

GoK. (2019). Data Protection Act. Retrieved from http://kenyalaw.org:8181/exist/kenyalex/actview.xql?actid=No.%2024%20of%202019

GoK. (2019). The Statistics (Amendment) Act. Retrieved from http://kenyalaw.org/kl/fileadmin/pdfdownloads/AmendmentActs/2019/StatisticsAmendmentAct2019.pdf

GPSDD. (2019). *Advancing sustainability together? Citizen-generated data and the Sustainable Development Goals.* Retrieved from https://www.data4sdgs.org/resources/advancing-sustainability-together-citizen-generated-data-and-sustainable-development

GPSDD. (2019). Choosing and engaging with citizen-generated data: A guide. GPSDD. Retrieved from https://www.data4sdgs.org/resources/choosing-and-engaging-citizen-generated-data-guide

GPSDD. (2019). Ghana and Kenya Peer-to-Peer Learning Exchange on SDG Monitoring., (p. 19). Retrieved from https://www.data4sdgs.org/sites/default/files/services_files/Ghana%20Kenya%20Peer%20Exchange_Summary%20Report_0.pdf

Open Data Watch, Data 2X. (2018). *the data value chain moving from production to impact.* Retrieved from https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact/

PARIS21. (2020). Use of CGD for SDG reporting in the Philippines: A case study. Retrieved from https://paris21.org/sites/default/files/inline-files/PSA-report-FINAL.pdf

Statistics Canada. (2017). Data quality toolkit. Retrieved from https://www.statcan.gc.ca/eng/data-quality-toolkit

UNDP. (2015, July 1). *Sustainable Development Goals*. Retrieved from UNDP: https://www.undp.org/content/undp/en/home/sustainable-development-goals.html

United Nations. (2013). Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda. New York: United Nations. Retrieved from https://www.un.org/sg/sites/www.un.org.sg/files/files/HLP_P2015_Report.pdf

UNSD. (2014). Fundamental Principles of Official Statistics. Retrieved from https://unstats.un.org/unsd/dnss/hb/E-fundamental%20principles_A4-WEB.pdf

UNSD. (2019). UN-NQAF Manual. Retrieved from https://unstats.un.org/unsd/dnss/docs-nqaf/UN_NQAF_Manual-Unedited_manuscript_of_3_May_2019.pdf

UNSD. (2020, March). *Report of the Friends of the Chair group on the Fundamental Principles of Official Statistics.* Retrieved from https://unstats.un.org: https://unstats.un.org/unsd/statcom/51st-session/documents/2020-21-FPOS-E.pdf

UNSD. (2020). *Supplementing the United Nations Fundamental Principles of Official Statistics: Implementation Guidelines.* Retrieved from https://unstats.un.org: https://unstats.un.org/unsd/statcom/50th-session/documents/BG-Item3b-FPOS-Implementation-guidelines-E.pdf

Wilson, Christopher and Zara Rahman. 2015. Citizen-generated data and governments: Towards a collaborative model, DataShift, p. 16. civicus.org/images/citizen-generated%20data%20and%20governments.pdf