Botswana Data Quality Assessment Framework User's Manual

Private Bag 0024,
Gaborone
Tel: 3671300
Fax: 3952201
Toll Free: 0800 600 200

Private Bag 47,
Maun
Tel: 371 5716
Fax: 686 4327

Private Bag F193,
City of Francistown
Tel. 241 5848,
Fax. 241 7540

Private Bag 32 ,
Ghanzi
Tel: 371 5723
Fax: 659 7506

E-mail: info@statsbots.org.bw
Website: http://www.statsbots.org.bw

STATISTICS BOTSWANA

Botswana Data Quality Assessment Framework User's Manual

Botswana Data Quality Assessment Framework Training Manual

Published by

Statistics Botswana
Private Bag 0024, Gaborone

Contact: Quality Assurance Unit
Tel: 3671300/ Ext 377/ 415
Fax: 3952201/3926419

E-mail: info@statsbots.org.bw
Website: www.statsbots.org.bw

March 2021

Acronyms/abbreviations

BDQAF	Botswana Data Quality Assessment Framework
CI	Confidence Interval
COICOP	Classification of Individual Consumption according to Purpose
CPI	Consumer Price Index
CPC	Central Product Classification
CV	Coefficient of Variation
DBA	Database Administrator
DQAT	Data Quality Assessment Team
DQAF	Data Quality Assessment Framework
EA	Enumerator Area
EMIS	Education Management Information System
IMPS	Integrated Patient Management System
GDDS	General Data Dissemination Standard
GDP	Gross Domestic Product
ICD-10	International Classification of Diseases (10th Revision)
ICT	Information and Communication Technology
ILO	International Labour Organization
IMF	International Monetary Fund
IT	Information Technology
ISIC	International Standard Industrial Classification
LFS	Labour Force Survey
MSE	Mean Square Error
NSO	National Statistics Office
NSS	National Statistics System
NGO	Non-governmental Organization
OECD	Organization for Economic Cooperation and Development
PES	Post-enumeration Survey
PPI	Producer Price Index
PSU	Primary Sampling Unit
QA	Quality Assurance
RAM	Random Access Memory
RMSE	Root Mean Square Error
SDDS	Special Data Dissemination Standard
SE	Standard Error
SNA 93	System of National Accounts 1993
SOC	Standard Occupational Classification
SB	Statistics Botswana
SVC	Statistical Value Chain
SITC	Standard International Trade Classification
UIS	User Information Service
UPS	Uninterrupted Power Supply
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNSD	United Nations Statistical Division
URS	User Request Service
USS	User Support Services
WTO	World Trade Organization

Table of Contents

PREFACE

The main purpose of BDQAF is to provide data quality criterion and clear procedures for designation of data as official statistics as prescribed by the Statistics Act. It is also utilized by producers of statistics to self-evaluate the quality of their data and produce quality declarations to guide users of their data.

BDQAF and training manual are mostly used by experienced data quality experts as well as subject matter specialists combined. Assessment using BDQAF is done through observations and interviews with the data owner.

Through the implementation of the BDQAF, statistical information that is generated form all sectors of the NSS will be assessed, among others, for quality so that any statistics that these sectors produce will be authenticated as official data and be used at large for various decisional purposes. The implementation of BDQAF is expected to identify the critical areas that contribute to the quality of data so that it assists to provide a reasonable response to improve data quality which in turn is expected to improve the evidence based decision making in the country.

This document is a Training Manual for BDQAF. It gives a number of examples stemming from the professional's own experiences as both users and producers of statistics as well as experiences accumulated through the independent assessment process.

Dr. Burton Mguni
Statistician General
March 2021

CHAPTER 1: INTRODUCTION

This document serves as a Training Manual for the Botswana Data Quality Assurance Framework (BDQAF). The purpose of the manual is to train the users and producers of statistics on the effective application of the BDQAF through the extensive elaboration of standards for each and every dimension. It also acts as a reference document.

The primary purpose of training in BDQAF is to create awareness of data quality among producers and users of data. It is also aimed at educating data producers on the advantages of using a framework for data quality and would assist in the provision of a common approach to data quality assessment in the country; it also raises awareness among producers around issues of standardization of data.

Training provides an understanding of key data quality elements and highlights areas of weaknesses that may exist which in turn would provide inputs on future improvement needed. This assists data producers in identifying priority areas of improvements and enables them to plan investments that can lend continuous improvements and result in the highest return in data quality. The training would also assist potential new data producers on the processes that need to be in place before embarking on costly data collection in order to gain high quality data.

A. The Need for BDQAF Training Manual

One of the main objectives for the development and implementation of DQAF is to address data quality issues within the National Statistics System. It is important in the process to ensure that users and producers of statistics are aware of these issues to facilitate the quality assessment processes such that the different data within NSS are declared as official statistics. It should also be noted that the process of assessment, identifies and highlights key areas of weakness and suggests improvements thereof.

This manual is structured that under every dimension, there would be its definition, key components, indicators and standards, thus following the DQAF

B. Definition of Data Quality and its Dimensions

ISO standard 9000:2005 defines quality as the "degree to which a set of inherent characteristics fulfills requirements". Statistics Botswana like other Statistical Agencies define data quality in terms of "fitness for use". Therefore, under this definition, the quality of statistical data can be determined by the extent to which they meet user needs. Data quality is further defined in terms of eight dimensions of quality including prerequisites of quality, and these are; relevance, accuracy, timeliness and punctuality, accessibility, interpretability, comparability and coherence, methodological soundness and credibility.

The following are definitions of each dimension as outlined in the BDQAF, including pre-requisites of quality;

Prerequisite to quality refers to the institutional and organizational conditions that have an impact on data quality. It defines the minimum set of necessary conditions that have to be met in order to produce quality statistics. It therefore serves as the foundation on which all other dimensions of data quality should be premised.

The credibility of statistical information refers to values and related practices that maintains users' confidence in the agency producing statistics and ultimately in the statistical product.

Comparability and Coherence: Comparability of statistical information is the ability to compare statistics on the same characteristic between different points in time, geographical areas or statistical domains; while coherence of statistical information reflects the degree to which it can be successfully brought together with other similar statistical information from different sources within a broad analytic framework and over time. It is the extent to which differences between two sets of statistics are attributable to differences between the estimates and the true value of the statistics.

Methodological soundness refers to the application of international, regional and national standards, guidelines, and good practices to produce statistical outputs. Application of such standards fosters national and international comparability.

The Relevance of statistical information reflects the degree to which the statistical product meets the needs of users.

The accuracy of statistical information is the degree to which the product correctly describes and or estimates the phenomena it was designed to measure. Accuracy also refers to the closeness of the values provided to the (unknown) true values.

Timeliness and Punctuality: Timeliness of statistical information refers to the time lag between the reference point to which the information pertains and the date on which the information becomes available. Timeliness also addresses aspects of periodicity and punctuality of production activities within the statistical value chain. **Punctuality** of statistical product is the time difference between the date the data are released and the target date on which they were scheduled for release, as announced in an official release calendar and laid down by regulations or previously be agreed with users.

The accessibility of statistical information refers to the ease with which it can be obtained from the agency. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed. The cost of the information may also be an aspect of accessibility for some users.

Interpretability of statistical information refers to the ease with which users understand statistical information through the provision of supplementary information (metadata and relevant supporting documents).

C. Structure of the Framework

The BDQAF covers the entire Statistical production cycle (statistical value chain) i.e. needs determination, design, build, collection, processing, analysis and dissemination, and all statistical processes. The framework certifies statistics as 'official' using one of the four levels:

Level Four: Very Good Quality Statistics- these are statistics that meet all the quality requirements as set out in the BQAF. They are designated as quality statistics to the extent that deductions can be made from them and are 'fit for use' for the purpose for which they were designed. Level four applies to highly-developed statistical activities with respect to the corresponding indicator.

Level Three: Good Quality statistics – These are statistics that meet most, but not all the quality requirements as stipulated in the BQAF. They are designated as acceptable to the extent that, despite the limitations, deductions can be made, and are 'fit for use' for the purpose for which they were designed. Level three refers to moderately well-developed activities with reference to a particular indicator.

Level Two: Acceptable Statistics – these are the statistics that meet few of the quality requirements as provided in the BDQAF. They are designated as questionable to the extent that very limited deductions can be made and they are therefore not 'fit for use' for the purpose for which they were designed. Level two refers to statistical activities that are developing but still have many deficiencies.

Level One: Poor Statistics – These are statistics that meet almost none of the quality requirements as provided in the BQAF. They are designated as poor statistics to the extent that no deductions can be made from them and are not 'fit for use' for the purpose for which they were designed. Level one refers to activities that are underdeveloped.

CHAPTER 2: DIMENSIONS OF QUALITY

This chapter introduces all dimensions of quality, including pre-requisites of quality. It highlights key components, indicators and standards definitions for every dimension, giving examples where possible for ease of reference and understanding. All these are as per the DQAF layout.

1 Prerequisites of Quality

The prerequisite of quality refers to the institutional and organizational conditions that have an impact on data quality. It defines the minimum set of necessary conditions that have to be met in order to produce quality statistics. It therefore serves as the foundation on which all other dimensions of data quality should be premised.

a. Key components

- Legal and institutional environment (Statistics Act and other Acts including Memoranda of Understanding (MoUs) or Service Level Agreements (SLAs)
- Confidentiality
- Commensurate resources
- Quality as the cornerstone of statistical work
- Risk management

b. Indicators and Standards

Legal framework for Statistics Production

Legal framework for Statistics Production

Quality Indicator	**1.1** The responsibility for producing statistics is clearly specified.
Standard	**1.1.1** A legal arrangement exists that explicitly mandates the production of statistics

Within the context of the framework, the legal concerns surrounding data products include enhancing data product safety and minimizing product liability. Any statistics producing agency should be able to demonstrate a clear mandate in one or more of the following ways:

a. An Act (e.g. statistics law or regulations, Statistics Act 2009);
b. Terms of Reference that gives the appropriate authority or mandate organizations with data collection;
c. Memorandum of Understanding; and/or
d. Service Level Agreement.

While the foregoing list may be inexhaustive, it is incumbent upon any data collecting agency or statistics producing agency to demonstrate that it has a legal basis for collecting information and that the basis for collection has been obtained from the necessary authority. The mandate should be clear on who should be producing which statistics or data and for what purposes, and thus there are no ambiguities on mandates and responsibility by each agency or department. The producer has a responsibility to act on the mandate given within a specified period of time by establishing arrangements that are consistent with this assignment of responsibility.

On the same vein Statistics Botswana has signed MOUs with a number of sectors covering data sharing aspects. However, there is need for the signing of service level agreements between data

producing agencies which will address the timeliness, format, etc. of the data to address the quality data issues. This will also encourage ownership and commitment to data quality issues among producing agencies to ensure development and maintenance of Information systems such as Integrated Patient Management System (IMPS) at the Ministry of Health and Wellness, Education Management Information System (EMIS) in the Ministry of Education and Skills Development, etc.

Statistics Value Chain Policies and Standards

Quality Indicator	**1.2** Standards and policies are in place to promote consistency of methods and results.
Standards	**1.2.1** A set of policies must exist which covers all aspects of the statistical value chain.
	1.2.2 A set of standards related to appropriate policies must exist.

Producing agencies are responsible for developing corporate policies which cover all aspects of the statistical value chain (SVC). These policies should be implemented and maintained to be consistent with other policies, especially other quality policies, e.g. Policy on Metadata Management, ICT Policy, Pricing Policy, Dissemination Policy, Data Quality Policy, etc.

Policies provide general rules in the form of a statement of principles that indicate how the data-producing agency will act in a particular area of its operation, e.g. data dissemination. These policies have to be agreed upon officially. The rules are derived and created from the Statistics Act, the Public Service Administration Act, Conditions of Service, Employment act, Statistics Botswana's mission and vision, BDQAF, and the Fundamental Principles of Official Statistics. Policies are used to:

a. set standards;
b. ensure compliance with legal and statutory responsibilities;
c. guide the agency towards the achievement of its strategic plan; and
d. improve the management of risk.

Comparability and Consistency of Standards

To be comparable over time, statistics need to be collected and analyzed according to coherent and consistent classifications, standards, concepts and definitions. Generally, standards are agreed principles for doing things in a common and consistent way. The use of standards promotes common understanding; ensures required performance, cost saving and compatibility; and enables communication. A statistical standard ensures the use of national, peer-agreed or international best practices. Statistical standards can vary from classification and measurement to concepts and definitions some of which need to be updated accordingly. Examples are Compendium of Concepts and Definitions; classifications such as International Standard Industrial Classifications (ISIC) for industries (Statistical Business Register), International Standard Classification of Occupations (ISCO), International Standard Classification of Education (ISCED),Classification of Individual Consumption according to Purpose (COICOP), etc.

An example from the International Standard Classification of Occupations (ISCO)

54 Protective Services Workers		
541		Protective Services Workers
	5411	Fire Fighters
	5412	Police Officers
	5413	Prison Guards
	5414	Security Guards
	5419	Protective Services Workers Not Elsewhere Classified

An example from International Standard Industrial Classifications (ISIC)

DIVISION 63: INFORMATION SERVICE ACTIVITIES		
631		Data Processing, Hosting and Related Activities, Web Portals
	6311	Data Processing, Hosting and Related Activities
	6312	Web Portals
639		Other Information Service Activities
	6391	News Agency Activities
	6399	Other Information Service Activities NEC

All data-producing agencies are encouraged to establish an infrastructure or functions that will be responsible for the development and implementation of appropriate standards for related policies. The development of these standards should cover all aspects of the statistical value chain. The use of statistical standards is fundamental in providing high-quality statistical products to users, and also in facilitating data sharing among data producers.

Data Sharing Policy

Quality Indicator	1.3	Data sharing and coordination among data-producing agencies are clearly specified
Standards	1.3.1	A legal arrangement must exist which allows for the timely and efficient sharing of data between the collecting agency and the user
	1.3.2	Regular meetings must occur between the data-producing agencies and users/agencies to resolve statistical issues

Arrangements or procedures should be in place to ensure the efficient and timely flow of data between the agency with primary responsibility for compiling the statistics, and other data- producing agencies. Data sharing and coordination should occur within a legal framework (e.g. Service Level Agreement, Memoranda of Understandings, and Act) that directs data producing agencies to share statistical information as well as addressing the timely flow of data as necessary.

Ideally, a legal agreement must exist both within an agency and between agencies, e.g.:

a. data producer and user,
b. two data producers,
c. data producers and support services

It is advisable for the agencies that share data (depend on each other for data) to have regular contact through meetings (e.g. user-producer committee meetings, sub-committees), workshops (stakeholder workshops for specific areas of specialty) and other forums to ensure proper understanding of data requirements. These meetings can serve a variety of purposes and some of these are listed below:

a. Promote the idea that statistical needs have to be built into administrative collections since those may not have been designed with statistical production in mind;

b. Address statistical issues that prevent the duplication of effort and reduce respondent burden;

c. Address user needs in a structured and formalized process that may enhance the relevance of the statistical product and

d. Assist in aligning operational and statistical definitions, facilitate comparability and coherence with other information within the same domain.

Confidentiality Measures

Quality Indicator	**1.4**	Measures are in place to ensure that individual data are kept confidential, and used for statistical and administrative purposes only
Standard	**1.4.1**	There must be a law or policy that ensures information collected is kept confidential and used for statistical and administrative purposes only

In the production of statistics, whether through administrative collections, surveys or censuses data collection agencies should put in place a process that ensures the respondent information remains confidential. The process should be backed up through law, policy or any other formal provision that clearly states individual responses are treated as confidential. The provision should state that responses shall not be disclosed or used for any purpose other than statistical purposes unless disclosure is agreed to in writing by the respondent. It is important that confidentiality be maintained throughout the statistical value chain, while preserving the usefulness of the data outputs to the greatest extent possible. For example, Statistics Botswana uses the Statistics Act 2009 clauses as an enforcement or obligation to data provision by respondents as well as its confidentiality as per the UN Fundamental Principles of Official Statistics (i.e. Section 29 (2g)) and sections 46 (1) as well as 49(1 and 2) highlighting related offences Section 46(1)

46 **(1)** "Where the statistical information collected from individual returns, worksheets, or any other confidential source under this Act is to be copied or recorded on tape, card, disc, wire, film, or any other method of recording and whether encoded or plain language or symbols are used for the processing, storage or reproduction of the information, the Statistician General shall take such steps as are necessary to ensure the security and confidentiality of the statistical information".

49 **(1)** "Any person who, in the execution of any function under this Act, or who at any one time executed any function under this Act, discloses or makes known any matter or thing in breach of any oath or affirmation made under this Act shall be guilty of an offense and be liable to a fine not exceeding **P50, 000.00** or to imprisonment for a term not exceeding five years, or to both".

49 **(2)** "Any person who, in the execution of any function under this Act, or by virtue of his or her employment, becomes possessed of any information which might exert influence upon, or affect the market value of any share, interest, product or article, before such information is made public, directly or indirectly uses such information for personal gain, shall be guilty of an offense and be liable to a fine not exceeding **P150,000.00** and to imprisonment for a term not exceeding 15 years, or to both".

Data collecting agencies should have rules and regulations, (e.g. Statistics Act, conditions of service) governing staff to prevent disclosure of information. These rules could include restricting access to unit record data to those who require the information in performing their duties. Steps should be taken to secure the premises of the data-producing agency and its computer systems to prevent unauthorized access to individual data. Whether in electronic or paper format, the confidentiality of data should be appropriately guarded during storage and during the process of destruction of records. When the duration for the archiving of forms or questionnaires has lapsed, they should be disposed off in a manner that complies with the applicable policy. Adequate dissemination policies should also prevent the release of information at levels that can identify respondents.

A way of avoiding this is to develop special aggregation/anonymization rules. Ideally, before information is solicited from the respondent, officials of the agency should seek to reassure respondents that their information will be treated as confidential and used for statistical purposes only. In order to achieve this, a message to this effect can be stated on the forms/questionnaires, website of the data collecting agency, terms and conditions pamphlet or through a variety of media forming part of a publicity campaign as well as data collectors being sworn to secrecy as per the dictates of the Statistics Act 2009 Section 45 (8) which states that **" Every person involved in the research or statistical project for which information is disclosed pursuant to this section shall take an oath or affirmation as set out in the first schedule"**.

Where data are to be exchanged with other agencies, procedures should be in place to adequately secure data while in transit and prevent the negligent loss of data. If the transfer is electronic, data can be encrypted to ensure that if intercepted the data are of no value to the interceptor. In the same way, necessary precautions should be taken in cases of manual transfer of data. Amongst others, these include a combination of password protection, use of courier services, and encryption of data files.

Respondents Obligation

Quality Indicator	**1.5**	Measures to oblige response are ensured through law
Standard	**1.5.1**	There must be a law or other formal measures that inform respondents of their obligation to provide information; and any sanctions which may apply if they fail to do so

In collecting data, respondents should be informed of their obligation under law with regard to the provision of information, as well as the consequences for failure to comply. For example, Statistics Botswana uses clauses from the Statistics Act 2009 to address obligatory issues relating to provision of information, be it at individual respondents or companies/establishments ; Statistics Act 2009 Section49 (6) … states that

(6) "Any person who -

a) refuses or neglects –

(i) to fill in and supply the particulars required in any return, form or other document which by this Act he or she is required to fill in and supply;

(ii) to answer any questions or enquiries put to him or her under this Act; or

b) contravenes the provisions of section 32, shall be guilty of an offence and be liable on conviction to a fine not exceeding P10,000.00, or to imprisonment for a term not exceeding one year or both".

Note that Statistics Act 2009 Section 32 states that;

" For the purpose of enabling statistics to be collected otherwise than by way of census, every person shall, when required by the Statistician General or any authorized officer, furnish, in the form (if any)supplied by that officer, or in the absence of such form or where the person concerned is unable for any reason to complete such form, verbally, the particulars or information required by the relevant regulations relating to the matter in respect of which such particulars are, or such information is required".

At a minimum when soliciting information from respondents they need to be provided with three pieces of information, viz.:

- the collecting agency's mandate to collect data;
- the purpose of the collection
- and who the data will be shared with

If possible, when collecting data, partially completed forms/questionnaires with information previously supplied and having remained unchanged, should be pre-populated. In this way, the data-producing agency demonstrates that it carefully considers respondent burden by shortening the time required in completing the questionnaire or form.

While punitive measures may be applied to secure data from respondents it is advisable for collecting agencies to create goodwill with those from whom they seek to obtain information. Collection agencies can establish a single point of contact through service centers to deal with requests from

respondents on how to complete questionnaires or forms. These service centers may provide support to respondents who prefer to walk into service centers, or it could be through a call center or both. Either way the collecting agency demonstrates its goodwill towards the respondent by making a service available that assists in the data collection process.

Resources

Quality Indicator	**1.6** Resources are commensurate with the needs of statistical programmes • Staff • Facilities • Computing Resources • Financing
Standards	**1.6.1** Have adequately skilled personnel within the statistical value chain
	1.6.2 There must be a Statistics Unit or component responsible for compiling statistics
	1.6.3 Facilities must have the infrastructure to manage the needs of statistical programmes (Facilities-office space, furniture)

Human Resources

Human resources should be commensurate with the needs of statistical programmes. Good recruitment processes, such as interviews and competency tests, etc., should be in place with appropriate remuneration and rewards aligned to skills and competencies of the staff they desire to attract. Overall, the staff complement should be adequate to perform the required tasks with none being overloaded with work. Recruitment processes should be aligned with the skills needs of the organization, for example, an initiative such as an internship programme which can attract young graduates.

Job satisfaction should be measured, with areas of concern addressed. Staff should be encouraged to attend conferences and seminars to be kept abreast of new developments in their line of work. Staff should be provided with both formal and on-the-job training in all aspects of the statistical value chain. Examples of trainings include methodology, data collection and processing, GIS issues in relation to data analysis, information technology (IT), analysis, communication skills and management, etc. to acquire and sharpen the skills necessary to perform their tasks. Sound management principles ensure the retention of core staff which includes methods such as career-pathing strategies, staff mentoring, and succession planning.

Establishment of a Statistical Unit

Agencies whose core business includes the provision of agriculture, health, education, justice, security, services etc. a separate unit responsible for the compilation of statistics should exist. This is because the adequacy of services provided can only be measured through the collection and compilation of statistics from administration record-keeping or sample surveys. This will require that statistical outputs to be planned and budgeted for, as well as staffed with capable and competent personnel.

Facilities

The buildings and other storage facilities used should be secure with restricted access control and have adequate infrastructure to manage the needs of the statistical programme.

Amongst others, this will include:

- **a.** designated warehouse/storeroom to keep questionnaires, maps, listing books, and other relevant documents secure;
- **b.** adequate office space for staff to work in;
- **c.** facilities for handling forward and reverse logistics of questionnaires efficiently;
- **d.** stores management system that monitors the allocation and return of questionnaires during processing;
- **e.** adequate working environment such as lighting, heating, clean air, adequate office space and the provision of office furniture for performing the required tasks;
- **f.** printing facility for dissemination; and
- **g.** adequate logistics (vehicles, telephones, fax machines, photocopiers, etc.) to ensure that statistical production occurs in an efficient way.

Computing Resources and Business Continuity Plan

Standards	**1.6.4 Computer hardware resources must be adequate in terms of** • data storage; • data backup media; • power supply (uninterrupted); • memory; and • other necessary equipment (notebooks, desktops, etc.).
	1.6.5 A disaster recovery and business continuity plan must exist

The demand for Information Communication Technology (ICT) resources is critical and they form part of a sustainable statistical programme. Hence, at a minimum, the agency should ensure that hardware for statistical programmes has sufficient Random Access Memory (RAM) and data storage capacity that are commensurate with the needs of the programme. Additionally, ICT equipment such as file and database servers are sensitive to changes in electrical current as well as electrical outages.

To mitigate the risks associated with the foregoing, it is advisable that servers have Uninterrupted Power Supply (UPS) units. The units will ensure the controlled shutdown of these machines when there is a power outage. Regardless of the precautions taken above, even the best systems experience unanticipated problems that result in complete systems failure. When this happens, the system will have to be restored to state at a given point in time. Therefore, there has to be regular scheduled backup procedures of databases that will allow restore points to be established.

Computing resources should be commensurate with their usage on behalf of the statistical programme. Data processing and analysis should be resourced with sufficient or powerful computers. All computing resources earmarked for statistical production should be used for such and not be diverted to accommodate systems and processes unrelated to statistical production. Research should take place in the usage of new and more efficient technologies such as use of Computer Assisted Personal Interviews (CAPI) for data collection and Smart Census Application for census pre-enumeration exercise, etc.

Data collected in a decentralized manner need to be moved from different geographic locations to a central data repository. The use of Local Area Networks (LAN), Wide Area Networks (WAN), Virtual Private Networks (VPN) and their associated carrying capacity can become major bottlenecks in ensuring the seamless transfer of regional data to a central repository. Special attention needs to be given to the maintenance and upgrading of communications infrastructure if this is used in collecting data from different geographic locations. Failure to do so will result in significant degradation of the network's performance.

This will include ensuring that there is sufficient bandwidth where electronic data transfer occurs between locations. Infrastructure such as network switches, and routers need to be maintained and upgraded to ensure the predictable transfer of data collected from various locations to a central repository. This becomes crucial in an environment where administrative data collections are conducted on a daily basis and in a variety of locations across the country. This will require the implementation of scheduled and documented maintenance procedures.

Business Continuity plan

A business continuity plan should be in place that will illustrate how to recover and restore partially or completely interrupted critical functions within a predetermined time after a disaster or extended disruption. A disaster recovery plan which serves as a subset of a business continuity plan must exist and include planning for resumption of applications, data, hardware, communications such as networking and other IT infrastructure. Backup systems should be in place and backup should be done every day. It is advisable not to backup data and information in the same premises. However, business continuity should not stop at IT continuity only, it should include provisions for building infrastructure and human skills.

Computing Resources and Budget Adequacy

Standards	**1.6.6 Computer software resources must be adequate in terms of** • Capturing systems • Editing systems; • Coding systems; • Statistical software; • Up-to-date licenses; • Virus protection; and • Appropriate access rights.
	1.6.7 Budgets must be adequate

Every data collection programme needs to have sufficient ICT hardware resources. The collection of administrative or survey data and the associated production of statistical products require the use of specialized software. Amongst others, these may include software such as SQL, ADEPT, CSPro, STATA and SPSS to perform statistical computations. Where these are used, the agency should ensure that they have up-to-date software licenses that will allow access to the entire suite of statistical functions.

The SVC includes activities related to data collection, coding, editing, capturing and analysis that may require additional specialized software. Even though many of these can be done manually, there has been a clear move in automating these by purchasing specialized application software. Specialized software has to be accompanied by a detailed specifications document and business case outlining the proposed benefits it will bring to the statistical programme. This will ensure that funds are not needlessly used to purchase expensive software that does not result in material benefits to the statistical programme such as improved efficiencies through economies of scale.

Even though statistical programs require specialized statistical software there continues to be a demand for generic business software. These include database software provided by major software dealers that allows for managing and administering databases. It is crucial that any data collection program has sophisticated database application software that allows it to manage the data in the associated databases efficiently.

Up-to-date licensing ensures the right to access and use of the software. Virus protection software must be used to prevent, detect and remove computer viruses. It increases the security of computers and network systems used in the data collection program. Appropriate data access rights must be put in place to prevent access to data by non-intended parties.

Budget Adequacy

The budget allocated to the data collection and processing programs should be adequate for all production and related activities so that lack of resources does not jeopardize the optimality of the activities in the SVC. Plans must be put in place to allocate budgetary resources to future statistical development based upon identified statistical needs.

Policies on Resource Use

Quality Indicator	**1.7**	Measures to ensure efficient use of resources in 1.6 are implemented (Processes to assess if the job profile and descriptions are correct)
Standards	**1.7.1**	Staff of a statistical programme must be employed in positions that are aligned with their skills profile
	1.7.2	Asset management policies that prevent the abuse of facilities (e.g. Vehicles, telephones, etc.) must be developed, adopted and implemented
	1.7.3	Policies, guidelines and procedures governing the use of ICT resources must exist, so as to maximize the return on investment
	1.7.4	Budgets must be reviewed and financials audited to ensure that resources are used in the best possible way

Skills Audit Exercise

The organization should regularly conduct a skills audit exercise to determine training, recruitment and staff re-assignment needs as well as develop a skills profile for each staff member. Staff, both skilled and semi-skilled, in a statistical programme should be employed in positions that are aligned with their skills profile with documented reasons for deviations.

Recruitment processes should be aligned with the skills need of the organization with an internship programme put in place to attract young graduates. Job satisfaction should be measured, with areas of concern addressed, and staff should be encouraged to attend conferences and seminars to be kept abreast of new developments in their line of work.

Asset Management Policy

Asset management policies should be in place to prevent abuse of vehicles, telephones, printers and other facilities. Policies regarding security of the facility should be in place, especially for areas where sensitive information such as data is stored

ICT Policy and Guidelines

Policies governing the use and abuse of computing resources should be in place. Amongst others, these policies could address software usage and include:

- Internet;
- Email;
- data dissemination;
- data storage;
- network resources; and
- data access policies that prevent abuse of computing resources.

An appropriate type of ICT should be used for the data collection programme. Failure of ICT systems can have a devastating impact on any statistical programme if any activities of the **SVC** are automated. It is therefore important to have policies governing the use of these assets and protecting them from abuse. For example, Statistics Botswana has ICT policy and guidelines to address usage and protection of ICT resources.

Periodic reviews of budgeting procedures should be undertaken to ensure that financial resources are best employed in addressing major data problems or meeting new data priorities.

Quality Measures

Quality Indicator	**1.8**	Processes are in place to focus on, monitor and check quality
Standards	**1.8.1**	The agency must have a quality management system in place
	1.8.2	The data-producing agency must have an independent data quality audit process
	1.8.3	Staff members in production areas must have a data quality management requirement as part of their performance agreements or job descriptions

Quality Management System

There is a need for all data producers to develop and implement a Quality Management System (QMS). This system will include tools for quality declaration and assessment, auditing, policies around quality management, metadata management framework, clearance and other interventions to improve the quality of products. The QMS will provide a structure to ensure that the process of managing quality during the production of statistics is carried out in a formal and systematic way.

Data quality awareness programmes should be conducted to communicate to the employees the aim of the data quality management system; the advantage it offers to employees, customers and the organization; how it will work; and their roles and responsibilities within the system. The awareness programme should emphasize the benefits that the organization expects to realize through its data quality management system. Since data quality management systems affect many areas in the organization, training programmes should be devised for different categories of employees. The data quality implementation plan should make provision for this training. The training should cover the basic concepts of data quality management systems, the standard to be adhered to, and their overall impact on the strategic goals of the organization, the changed processes, and the likely work culture implications of the system.

Data Quality Assessment Framework

Independent quality assessment in the NSS has as its goal to produce more official statistics to inform government on progress made in the implementation of its programmes and projects. This is performed by a data quality assessment team (DQAT) which is appointed by the Statistician-General (SG). Yet, as part of the statistical production process and in line with the SVC, there should be a practice of independently auditing the data from time to time. Furthermore, there should be a clear separation of functions of those involved in the statistical production process and those who conduct the quality audit, as this obviates the outcome where assessment and production functions are vested in the same individual. The quality assessment process requires the existence of quality standards and benchmarks that form part of a broader quality framework.

The assessment could be a combination of metadata-based audit analysis of data extracts as well as computation of some of the indicator values. Thus this assessment presupposes the existence of metadata that have been collected as part of monitoring the operations of the entire SVC. Poor monitoring of processes along with inadequate supporting documentation (i.e. metadata) makes it impossible to arrive at a favorable assessment as an outcome. The assessment process should be undertaken periodically to ensure that where quality standards have been met, they continue to do so in the future; and where this is not the case, incremental improvements are made in a given period of time. In terms of a DQAT assessment, a good performance can then lead to the SG declaring statistics as official.

Quality Management as Part of Individual's Performance Agreement

Managers and professionals should be made sensitive to issues of data quality. Also, data quality criteria should be built into the job descriptions or performance agreements of employees, making them accountable for data quality. Sanctions should be applied for failure to comply with all known quality regulations and standards. Each phase of the value chain should have built-in quality assurance (QA) processes. In turn, each QA process should have a checklist of activities to perform and standards to achieve. Each quality assurance process should be thoroughly documented and recorded, and should form part of the operational metadata.

Risk Measures

Quality Indicator	**1.9** Policies and frameworks are in place to manage risk in the statistical value chain
Standard	**1.9.1** A Risk Management Framework, policies and register which cover all processes in the statistical value chain must exist.

Enterprise Risk Management

An Enterprise Risk Management (ERM) entails Risk Management Policy, Framework, and Registers. It provides a basis for management to effectively deal with the uncertainty of associated organizational objectives implementation, thereby enhancing its capacity to build value. Value is maximized when management sets objectives to strike an optimal balance between growth and related risks, and effectively deploys resources in pursuit of set objectives, that is, the organization gets value for money (efficiency, effectiveness and economic benefits). A change in strategic or operational objectives will necessitate a re-assessment of the risks and mitigation. It is accordingly accepted by all stakeholders that an organization will manage its risks in an appropriate manner.

ERM is an enabler of the management processes. It is one of the principles of corporate governance. It provides Management and the Board with information on the significant risks and how they are being mitigated. Internal controls are an integral part of ERM; it helps an organization to ensure that objectives are achieved in an efficient and effective manner.

Below are some of the advantages of ERM:

a. Ensures the organization achieves its performance targets, and prevent loss of resources.

b. Ensure effective and informed reporting.

c. Ensures that the entity complies with laws and regulations, hence avoiding reputational damage and other consequences.

In general, the ERM acts as an overall driver to risk management process within an organization, by providing the details of the 'what, who, where, when and why of risk management activities.

Statistical Agencies must ensure that they have enterprise risk management policy and framework in place or any other processes of mitigating risks. These processes must also be integrated in the organization's strategy setting to ensure that risks associated with its statistical and related processes within the statistical value chain (SVC) are identified, assessed, prioritized/rated, mitigated and monitored accordingly.

2 Credibility

The credibility of statistical information refers to values and related practices that maintain users' confidence in the agency producing statistics and ultimately in the statistical product.

a. Key components

- Professionalism and ethical standards which guide policies and practices.
- Assurances that statistics are produced on an impartial basis.
- Ethical standards are guided by policies and procedures.

b. Indicators and Standards

Confidentiality Measures

Quality Indicator	**2.1**	The terms and conditions, including confidentiality, under which statistics are collected, processed and disseminated, are available to the public
Standard	**2.1.1**	A terms and conditions document must be available and accessible to the public and follow UN fundamental principles of official statisticst.

UN fundamental Principles of Official Statistics

A good system of national statistics must meet certain general criteria which can bring a level of trust in the quality of the data that are made available to the users. These criteria must be readily available to users. The United Nations fundamental principles of official statistics provide terms and conditions, including confidentiality, under which statistics collected, processed and disseminated are available to the public. These principles are given below:

Principle 1 deals with Relevance, impartiality and equal access to statistical products;

Principle 2 deals with the application of professional standards and ethics to ensure the use of scientific principles on the methods and procedures for the collection, processing, dissemination and storage of statistical data, to retain trust in official statistics;

Principle 3 deals with accountability and transparency to ensure availability of metadata for interpreting data;

Principle 4 deals with statistics agencies' right to comment on erroneous interpretation and misuse of statistics;

Principle 5 addresses cost effectiveness through the maximum use of data from various sources provided that quality, cost and burden on respondents are covered;

Principle 6 refers to confidentiality protection of individual data collected for statistical purposes only;

Principle 7 deals with legislation, i.e. the Stats Act and how it should be made public

Principle 8 looks at coordination amongst statistical agencies within countries for achieving consistency and efficiency in the national statistical system;

Principle 9 considers the need for using international concepts and definitions;, standards and methods for promoting consistency and efficiency of statistical systems at all official levels; and

Principle 10 promotes bilateral and multilateral cooperation among statistical agencies or organizations in order to share best practices and in turn, contributes to the improvement of systems globally and hence international comparability

Data producing agencies are encouraged to follow the UN Fundamental Principles of official Statistics. Alternatively, they can develop their own terms and conditions under which statistics are collected, processed and disseminated, building on the Fundamental Principles of Official Statistics and should cover at least the following items:

Purpose of collection

This will include the name of an agency producing data, and the purpose or objectives of collecting such data.

Confidentiality

This item is already covered in detail in Chapter 1: Prerequisites, and should include issues surrounding confidentiality and anonymization of data. For example, individual data collected by Statistics Botswana for statistical compilation will be strictly confidential and used exclusively for statistical purposes as stipulated in the Statistics Act 2009.

Dissemination

Data should be disseminated according to a data dissemination policy. If this policy exists, it should be in accordance with international best practices such as the GDDS, SDDS or national and peer-agreed standards. There should be a policy or procedure to ensure confidentiality during data dissemination. These include data embargo and adopting access control systems. Data should be disseminated in an impartial manner, that is:

a. The timing of the data release should not be in response to political pressure.
b. Released data should not be withdrawn in response to political pressure.
c. The design of the survey should be described in the metadata and should accompany the published results.

Certain general criteria can bring a level of trust in the quality of the data that are being produced. These criteria can be defined as terms and conditions, and could include the law regulating the production of the statistics and the impartiality of the statistics that are produced. The statistics should also be produced using adequate scientific standards which will permit agencies to defend them. In addition, the data that are produced should be readily available to all users as part of the citizen's entitlement to public information.

Data Access Measures

Quality Indicator	**2.2**	The conditions under which users have access to data are described and published.
Standard	**2.2.1**	A data dissemination policy detailing the conditions under which users have access to the data must be available

Data Dissemination Policy

Data should be disseminated according to a data dissemination policy and in an impartial manner. The policy should be in accordance with international best practice such as the GDDS, SDDS or national and peer-agreed standards as well as the UN Fundamental Principles of Official Statistics.

Notification Measures

Quality Indicator	**2.3**	Advance notice is given of major changes in methodology and source data
Standard	**2.3.1**	Advance notice of at least 6 months must be given of major changes in methodology and source data

Advance Notice

Information about revisions and major changes in methodology should be communicated well in advance of the statistical release and before the data are made public. Changes can be gazetted, that is, be made through public announcements, on the Internet (website), etc. For example, notices of a change in the CPI basket and National Accounts rebasing, upgrading from NSA93 to NSA2008, upgrades on classifications (ISIC Rev 3 to ISIC Rev 4; ISCO to ISCO 2008; ISCED to ISCED 2011; moving from traditional data collection (Paper Assisted Personal Interview (PAPI) to Computer Assisted Personal Interview (CAPI)) should be given well in advance through special communication or as per the statistical agency's communication strategy.

Details of major changes and/or revisions to published data should be described in the explanatory notes of the relevant publication and metadata. Information about statistical standards, frameworks, concepts and definitions, sources and methods can also be released in a range of information papers and other publications to ensure that the public is informed about changes to statistical processes. It is also worth publishing discussion or technical documents well in advance where the nature of any changes is discussed.

Independence Measures

Quality Indicator	**2.4**	Government commentary, when data are released, should be identified as such, and not be seen as part of the official statistics
Standard	**2.4.1**	Government commentary, when data are released, must be identified as such, and not be seen as part of the official statistics

Data producers should neither favour any particular group in society nor be influenced by those in power to act against the principles of the prevailing professional ethics, i.e. UN Fundamental Principles of Official Statistics, Code of Conduct. Politicians should not be able to suppress reports which reflect poorly on them. The independence of the statistical authority from political and other external interference in producing and disseminating official statistics should be specified in law. The head of the data- producing agency has the responsibility of ensuring that statistics are produced and disseminated in an independent manner. For example, the Minister may write a foreword/preface to statistical documents, but these must not be seen to influence any of the results that follow.

Statistical Considerations

Quality Indicator	**2.5**	Choice of source data, techniques and dissemination decisions are informed solely by statistical considerations.
Standard	**2.5.1**	The choice of source data, techniques and dissemination decisions must be informed solely by statistical considerations.

Statistics should be compiled objectively, scientifically and impartially. The choice of data sources for statistical products as well as activities in the statistical value chain should be informed purely by strict professional considerations, including scientific principles and professional ethics. All activities of the statistical value chain must be informed solely by statistical considerations. Politicians should not be allowed to suppress or manipulate results.

Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies should choose the source with regard to quality, timeliness, costs and the burden on respondents. The decision regarding the choice of techniques, methods, definitions and source data that are to be used must be left to the data producers.

Ethical Considerations

Quality Indicator	**2.6**	Ethical guidelines for staff behavior are in place and are well known to the staff.
Standard	**2.6.1**	A professional code of conduct must be in place providing ethical guidelines for staff behavior.

Code of Conduct

The data producer should have a code of conduct in place which guides the behavior of staff with access to data, and address conflict of interest situations. There should be clear rules that make the connection between ethics and professional work. Management should acknowledge its status as a role model and should be vigilant in following the code of conduct. New staff members should be made aware of the code of conduct when they join the organization. Staff members should be reminded periodically of the code of conduct.

3 Comparability and Coherence

Comparability of statistical information is the ability to compare statistics on the same characteristic between different points in time, geographical areas or statistical domains. The coherence of statistical information reflects the degree to which it can be successfully brought together with other similar statistical information from different sources within a broad analytic framework and over time. It is the extent to which differences between two sets of statistics are attributable to differences between the estimates and the true value of the statistics.

a. Key components

- The use of common concepts and definitions within and between series.
- The use of common variables and classifications within and between series.
- The use of common methodology and systems for data collection and processing within and between series.

b. Indicators and Standards

Use of Common Concepts, Definitions and Classifications within Series

Quality Indicator	**3.1**	Data within series and administrative systems are based on common concepts and definitions, classifications, and methodology, and departures from this are identified in the metadata.
Standards	**3.1.1**	All data (including source data, related frame data, and related survey data) within the same series must use the same concepts and definitions. Departures from common concepts and definitions must be identified, documented in the metadata and archived.
	3.1.2	All data (including source data, related frame data, and related survey data) within the same series must use the same classifications. Departures from common classifications must be identified, documented in the metadata and archived
	3.1.3	All data (including source data, related frame data, and related survey data) within the same series must use the same methodology. Departures from common methodology must be identified in the metadata and archived

Concepts and Definitions Considerations

Coherence within a dataset implies that the elementary data items are based on common concepts, definitions and classifications; and can be meaningfully combined. This requires the development and use of standard frameworks, concepts and variables, and classifications for all the subject matter topics that are measured within the same series or administrative systems. This aims to ensure that the target of measurement is consistent within series, e.g. consistent concepts and definitions are used within series so that concepts such as 'household' have the same meaning from year to year and quarter to quarter. Differences in concepts and definitions within the same series and administrative systems over time should be described and the reasons for and effects of these differences should be clearly explained and that the quantities being estimated bear known relationships with each other.

Standard Classifications Considerations

Standardized classification systems should be used for all categorical variables used in a survey or administrative system. Classifications not only help with the better management of data but also aid the understanding and comparability of data. The realization of this element is normally through the adoption and use of frameworks such as the System of National Accounts and standard classifications systems for all major variables. Some examples of such classifications follow;

a. International Standard Classification of Education (ISCED)
b. International Standard Industrial Classification (ISIC)
c. Broad Economic Classification (BEC)
d. International Classification of Disease (ICD10)

An Extract from International Standard Classification OF Education

DIVISION 4: SCIENCE PROGRAMMES		
46		**Mathematics and Statistics Programs**
	461	Mathematics and Numerical Analysis Programs
	462	Operations Research Programs
	463	Actuarial Science Programs
	464	Statistics and Other Allied Fields Programs.
	469	Other **Mathematics and Statistics Programs NEC**

Methodology Considerations

The data-producing agency should develop and adopt a common methodology for a series. These include the development and use of the common:

a. Frames and source data;
b. Sampling techniques;
c. Frameworks such as the System of National Accounts used for the compilation of National Accounts Statistics;
d. Data collection and processing methodology and
e. seasonal adjustment methodology, etc.

This can be achieved by establishing centers of expertise in certain methodology and technologies, exchanging experiences, identifying good practices, developing standards, and training. The use of a data quality framework also enhances the consistency of the methodology used. The use of the same methodology aims to ensure that the series consistently uses the same methodology so that comparability over time for the series is feasible and any deviation from the trend is attributable to a large extent to statistical causes. The following are some of the examples that could bring non comparability in the data; changes in sampling design, geographical boundaries, coverage, reference period, standards and classifications, etc.

Statistics Consistency

Quality Indicator	**3.2**	Statistics are consistent or reconcilable over time
Standards	**3.2.1**	Statistics must be consistent over time
	3.2.2	The statistics must follow an expected trend established over time. Any inconsistencies in the key variables must be reconciled and included in the metadata

Statistics are estimates for the unknown value of the characteristic of a population such as monthly income, quarterly employment rate, etc. Thus, comparability over time of these characteristics is expected in the absence of error. Inconsistency over time occurs if these characteristics (collected for a specific period) are not comparable with the data for the following period. This is considered a break in the time series. The data-producing agency should reconcile the inconsistencies in key variables by adjusting the estimates. These adjustments should be clearly explained in the metadata, i.e. rebasing effects.

The length of time series is the number of years that it remains unbroken. Note that the longer the time series, the more the confidence that the true trend is described. An unbroken series is more likely to be comparable with similar statistics in another unbroken series. Any break needs to be analyzed and understood before the next instance of the survey. A reasonable period of time for a large series is 5 to 10 years because it allows the observation of a sufficient number of components of a time series. For example, changes made from the 2008/09 Botswana Core Wealth Indicator Survey (BCWIS) to 2016 Botswana Multi Topic Household survey (BMTHS) to 2019 QMTHS must be identified and included in the metadata.

The statistics being produced should be compared against trends established by the results of previous years or data derived from other systems. This is sometimes referred as output editing. This is where errors that could not be detected at earlier phases of the value chain can be detected. These statistics must be consistent with accepted trends or must be compatible with the theory stemming from the subject matter. For example, changes in values for the key variable from the previous period (or the same period in the previous year) are inconsistent with the changes in other related variables. Such inconsistencies must be avoided or must be resolved. Correction of the statistics may require going back to the source data or looking into the editing and analysis processes.

Data Comparability between Series

Quality Indicator	**3.3**	Data across comparable series or source data are based on common frames, identifiers, concepts and definitions, and classifications, and departures from these are identified in the metadata.
Standards	**3.3.1**	Data across comparable series or source data must be based on common identifiers. Departures from common identifiers must be identified in the metadata and archived.
	3.3.2	Data across comparable series or source data must use the same concepts and definitions. Departures from common concepts and definitions must be identified in the metadata and archived.
	3.3.3	Data across comparable series or source data must use the same classifications. Departures from common classifications must be identified in the metadata and archived.

This section focuses on the comparison and integration of data from different sources. Coherence across datasets implies that the data are based on common concepts, definitions and classifications. Any differences must be explained and be accounted for. An example of incoherency across datasets would be if exports and imports in the national accounts could not be reconciled with exports and imports in the balance of payments. Data can be enriched by integrating data from multiple comparable sources, or data from comparable sources can be used to compare and improve the precision of estimates.

These sources can include both survey data and data from administrative systems. Some integration activities are regular and routine, e.g. the integration of data in the national accounts, benchmarking or calibration of estimates to more reliable control totals, or seasonal adjustment of data to facilitate temporal comparisons. This indicator encourages and promotes the collection of comparable data based on common frames, survey instruments, rules and methodology; and advances the notion of an integrated approach to data collection. The use of common frames, identifiers, concepts and definitions, and classifications also allows for the integration of these various sources in a consistent fashion.

Departures from common practices, procedures, and rules (including the use of common concepts and definitions, classifications, frames and identifiers) across all comparable data sources should be communicated to stakeholders to allow for the reconciliation of disparities due to said departures.

Statistics Consistency

Quality Indicator	**3.4**	Statistics are checked for consistency with those obtained through other data sources.
	3.4.1	Statistics must be checked for consistency with a comparable dataset. Inconsistencies must be reconciled.

It is good practice to verify statistical estimates using data from alternative sources. The alternative sources may in some instances serve as a benchmark. Examples of such exercises can be the reconciliation of survey data using administrative sources. The confrontations of data from different sources, and the subsequent reconciliation or explanation of differences, are necessary activities of the pre-release review of the data or certification process of the data. The differences in statistics should be quantified and the reasons should be described. Typically, some discrepancies may be attributed to differences in data collection, processes or differences in reporting units.

Common or Unique Identifier Considerations

Quality Indicator	**3.5**	A common set of identifiers (for the purpose of record matching) exist and have been agreed upon by data producers.
Standards	**3.5.1**	A common identifier must be agreed upon by the data producers.
	3.5.2	The common identifier must be unique in every dataset. Rules and practices must be agreed upon to ensure uniqueness.

An identifier should be able to uniquely and unambiguously identify a statistical unit. A common identifier can be a variable or a combined set of variables. This identifier can be a code with some meaning for example

- Omang number
- Enumeration Area (EA)
- TIN Number
- VAT Registration Number

Databases may have multiple identifiers. However there should be one unique identifier that is common to most data sources. The relevant stakeholders need to get together and decide on what the identifier will be, its future use, and what attributes would then be required for the identifier. Ownership of the identifier must be explicit and known to all stakeholders. This identifier should be used by various data producing agencies according to the agreed standards and related policy. If we can ensure that the data sources are using the same identifier, we know that comparability and coherence will be significantly enhanced.

This unique identifier is indispensable for the record matching between different databases to create statistical registers. The use of a unique identifier also allows for the confrontation of data from various sources that use the same identifier. Different administrative sources often have different population unit identifiers. The user can utilize this information to match records from two or more sources. Where there is a common identifier, matching (integration) is generally more successful.

4 Methodological Soundness

Methodological soundness refers to the application of international, regional and national standards, guidelines, and good practices to produce statistical products. The application of such standards fosters national and international comparability.

a. Key components

- Application of International, regional and national standards and methods.
- Data compilation methods employ acceptable procedures.
- Statistical procedures employ sound statistical techniques.
- Transparent revision policy and studies of revisions are done and made public.

b. Indicator and Standards

Compliance to Standards

Quality Indicator	**4.1**	Concepts, definitions, and classifications used follow accepted standards, guidelines or good practices (international, regional and national).
Standards	**4.1.1**	The concepts and definitions must satisfy accepted standards, guidelines or good practice in line with international, regional and national standards; and must be documented. Deviations from the standard must be formally approved, and be fully documented.
	4.1.2	The classifications must satisfy accepted standards, guidelines or good practice in line with international, regional and national standards; and must be documented. Deviations from the standard must be formally approved, and be fully documented.

Compliance to Concepts and Definitions

The conceptual basis for the statistics should follow international, regional, national or peer-agreed norms such as standards, guidelines, and agreed practices. Concepts are the subjects of inquiry and analysis that are of interest to users. They refer to general characteristics or attributes of a statistical unit or a population. Concepts are usually based on a theoretical or statistical frame of reference and are used to define a subject, the statistical units to be described and/or the population under study.

Compliance to Classifications

Classifications of items allow better management of records. Classifications schemes are usually hierarchical and allow systematic grouping of records together. These are usually done by transcribing text descriptions of records into numbers or alphabetical ordering. Concepts, definitions and classifications should follow internationally accepted practices. It is important to have concordances between organizational /national and international concepts, definitions and classifications.

The UN proposes some international classifications on: activity (ISIC); product (CPC, SITC, BEC); expenditure according to purpose (COFOG, COICOP, COPNI, COPP). Classifications that are used must reflect both the most detailed and the collapsed levels. For example, Statistics Botswana has adopted some systems of classifications of occupation, industry, education, products, geography etc.

Data-producing agencies should source their concepts, definitions and classifications from international agencies. These practices will enhance comparability and integration of data by users. Not all concepts and classifications sourced elsewhere should be adopted as such. Some international definitions might need to be adapted to local needs, e.g. the $1.25 a day money metric definition of poverty has been adopted in Botswana as a definition of poverty. It is recommended that agreed standards should be used for the compilations of such statistics. Since all categorical statistical data need to be classified for analysis, the classification criteria chosen to group data systematically need to be suitable for these analytical purposes. In a case where there is a need to deviate from the standards, reasons for such deviations should be documented in the metadata.

Scope of Study Considerations

Quality Indicator	**4.2**	The scope of the study is consistent with accepted standards, guidelines or good practices.
Standards	**4.2.1**	The scope of the study must be appropriate for the intended topic. The scope of the study must be consistent with accepted standards, guidelines or good practices in line with the survey constraints

Statistics must be sufficiently comprehensive in scope and in terms of conceptual development of concepts to adequately describe the subject area in question.

The scope of the study should also be aligned with the project constraints.

Methodology Standards Considerations

Quality Indicator	**4.3**	Methodologies used follow accepted standards, guidelines or good practices (national, international, peer-agreed), viz.: • Questionnaire design • Sampling methods • Sampling frame development • Frame maintenance • Piloting • Data Collection methods • Data editing, capturing, coding and imputation methods • Data analytical methods • Revision procedures
Standards	**4.3.1**	The designing of the questionnaire must follow accepted standards, sets of guidelines or good practice.
	4.3.2	The sampling methods must follow accepted standards, sets of guidelines or good practice.

Questionnaire Design

Collection instruments (questionnaires) play a central role in the data collection process and have a major impact on respondent behavior, interviewer performance, and respondent relations, all of which have a major impact on the quality of information collected. Questionnaire design should therefore follow international good practice and at the same time take into account user requirements, administrative requirements of the organization, processing requirements, as well as the nature and characteristics of the respondent population. All questionnaires must undergo pre-testing.

Sampling Methods

The choice of the sampling methodology has a direct impact on data quality. This choice is influenced by factors like the desired level of precision, availability of appropriate sampling frame, estimation methods used, objectives of the survey and budget. All these factors should be taken into account when deciding on the sampling method. Only scientific and statistically sound methods should be used. Scientific methods are the selection using probability-sampling methods. Examples of scientific methods include the following;

a. **Simple random sampling:** A sample is chosen from a population with each individual chosen at random entirely by chance, and having the same probability of being chosen for the sample as any subset of k individuals;

b. **Systematic Sampling:** Systematic sampling consists of taking every kth sampling units after a random start. For example, if your population size is N = 17 and the sample you need is 3, then k = 17/3 = 5.7~6 if the random start is 4, units 4, 10 and 16 will be included as your sample of n =3.

c. **Cluster sampling:** This refers to the methods of selection in which the sampling units, the unit of selection, contains more than one population element and hence the sampling unit is a cluster of elements. Each element must be uniquely identified with one (and only one), sampling unit.

d. **Stratified Sampling:** In broad terms, stratified sampling consists of the following steps: (i) the entire population of sampling units is divided into distinct subpopulations called strata. (ii) Within each stratum, a separate sample is selected from all the sampling units composing that stratum. (iii) From the sample obtained in each stratum, a separate stratum means (or other statistic) is computed. These stratum means are properly weighted to form a combined for the entire population. (iv) The variances are computed separately within each stratum and then properly weighted and added into a combined estimate of a population

Survey Frames

Survey frames should conform to the target population and contain minimal under-coverage and over-coverage. In the absence of standardized methods for sampling, methods should at least follow accepted guidelines and good practices outlined in academic or research texts, articles, posters, institutions and journals. Examples of academic texts are Kish (1995), Kalton (1983), Kendall and Babington Smith (1950), Neyman (1934). Journals are the Journal of the Royal Statistical Society, Journal of the American Statistical Association, Journal of Survey Methodology, and Journal of Official Statistics.

The frame should be well maintained and regularly updated.

Methodology standards Cont.

Standards cont.	**4.3.3**	The frame maintenance methods must follow accepted standard, sets of guidelines or good practice
	4.3.4	The piloting methods must follow accepted standard, sets of guidelines or good practice
	4.3.5	Data collection methods must follow accepted standards, sets of guidelines or good practice
	4.3.6	Editing, coding, capturing and imputation methods must follow accepted standards, sets of guidelines or good practice
	4.3.7	The methods of analysis used must follow accepted standards, sets of guidelines or good practice
	4.3.8	Revision methods used must follow accepted standards, sets of guidelines or good practice

Frames Maintenance

The quality of the frame must be monitored by periodically assessing its coverage. The statistical units within the frame must be maintained through regular updates from the source administrative register(s) or through regular fieldwork operations or as per the UN guidelines. For example, Statistics Botswana follows the UN guidelines on the maintenance of Statistical Business Register (SBR).

Piloting Processes

Data collection instruments, data capturing processes, computer codes used for data processing must be tested prior to the actual collection period. The purpose of a pilot is to refine the survey process and reveal any unanticipated problems. The outcome can thus improve the quality of the instrument as well as the other processes in the SVC for the actual survey.

Ideally, the pilot should cover all aspects or elements of the SVC and must be carried out in exactly the same way it will be administered in the main survey. It is good practice to pilot the questionnaire on respondents who have similar characteristics of the target population of the survey. In some instances a pre-test is conducted instead of a pilot. The pilot or pre - test exercises are intended to give feedback on the questionnaire along these lines:

How long it takes to complete the questionnaire;

a. Clarity of instructions;
b. Clarity/unambiguity of questions;
c. Questions they refuse to answer; and
d. Clarity of the layout of the questionnaire.
e. Applicability of the system, etc.

Data Collection Methods

Data collection methods should take into consideration the size of the survey, nature of the respondent population, and the type of information needed. For example, Statistics Botswana normally employs mailing method of data collection for business-based surveys, where respondents complete the questionnaire themselves. This method is usually supplemented by telephone, email or fax as follow-ups. On the other hand, the organization employs face-to-face interviews for household-based surveys, whereby an enumerator asks questions and completes the questionnaire for the respondent.

Data Management Methods

Data editing is done in order to take care of invalid/inconsistent entries or inconsistencies in the data so as to maintain credibility and to facilitate further automated data processing and analysis. The editing process cuts across the SVC (data collection, data processing and analysis). Caution should be exercised against overuse of query edit as these may create bias through the addition of data which may or may not be correct.

Imputation Methods

Imputation methods are employed to compensate for a unit or item non-response bias. A unit non-response is when there is complete non-response for a person, household or business; whereas an item non-response is when a record has partial information, i.e. certain questions not answered in a questionnaire.

Data Analysis Methods

Procedures and guidelines followed in carrying out data analysis must exist and must be well documented so as to unable replication of estimates where needed. For example, to be consistent with international standard of reporting, seasonally adjusted estimates must be reported.

Data Revision Methods

A clear and concise data revision policy for all published data must be produced for users to describe the main reasons why particular statistics are subject to revision, for example, GDP estimates are subject to revisions as well as preliminary estimates for any survey being carried out. These policies will explain how frequently and to what extent revisions will be made. All revisions should have taken place by a specified date in the survey with results clearly explained, analyzed and documented in the metadata. For example, in case where new and improved information is received, revision of estimates should be carried out. This will help on the improvement of the quality and reliability of the estimates

The data-producing agency should identify a set of standards and guidelines or accepted practices either from international or national communities such as IMF on SDDS, GDDS, etc. These usually stem from research conducted on methods used in specific fields of study.

Recommended accepted practice ensures consistent, cost-effective and repeatable methodologies. Thus for an agency to commit to using accepted practice in the area of statistics, it has to commit to using all knowledge, technology and resources available in the field to ensure success. The United Nations for example, has made available various documents on good practice in various topics of statistics e.g. civil registration and vital statistics; crime statistics, National Accounts, Trade, Employment, etc. These can be found at the web address: **http://unstats.un.org/unsd/progwork/pwabout.asp**. Hence the crucial part in the selection of good practice is the identification of the best ones from a pool of available ones. International organizations such as the OECD, WTO, ILO, IMF, Eurostat, UN, etc. and countries with relatively advanced statistics systems can serve as models of such practices.

Peer-agreed standards should have transparent and meaningful linkages to international standards. These should be developed by a community of experts in consultations with relevant stakeholders. The designing of the questionnaire, methods of sampling, sampling frame design, frame maintenance, piloting, data collection, editing and imputation of data, data analysis, and revisions of data methods should follow peer or nationally, regionally or internationally accepted standards, sets of guidelines or good practice. Data-producing agencies should describe the major methodological changes that have taken place during the reference period and how they affect the data quality.

Revision Schedule

Quality Indicator	**4.4**	Revision schedules are followed, regular and transparent.
Standards	**4.4.1**	A revision schedule must exist for statistical products, where applicable and must be publicly available, accessible and adhered to.

A revision schedule that represents the optimal revision cycle should exist for statistical products where applicable and should be followed. A pattern of revisions should be regular in order not to introduce bias. It is important that all users of data are at all times made aware of the revision policy relating to these data. Measures should therefore be taken to align patterns of availability of new data/ information with the intended revisions patterns. For example, in Statistics Botswana revisions of the estimates for all components of the national accounts is usually done every five years in conjunction with the rebasing of the estimates and constant prices. At such a time the results of the census that has become available in the meantime and any other additional information sources are incorporated in the estimates.

Preliminary and Revised Data Metadata

Quality Indicator	**4.5**	Preliminary and revised data are identified in the metadata.
Standards	**4.5.1**	Preliminary and revised data must be identified in the metadata. Metadata must contain an explanation of the changes.

Always indicate to the users the nature of the data (e.g. preliminary estimates) when publishing data that are likely to be subsequently revised. Studies should assess preliminary estimates against final estimates and identify all changes. These studies of revision should be documented in detail. Adequate documentation on revisions should be maintained and should include descriptions of causes of revisions, methods used to incorporate new data sources, the way data are adjusted, the frequency of revisions and the magnitude of the revisions. Both preliminary and revised data must be kept and made available to users and the metadata must clearly identify such revised data.

Revision Studies Considerations

Quality Indicator	**4.6**	Studies of revisions and their findings are made public.
Standards	**4.6.1**	Regular studies of revisions or upcoming revisions must be done and their findings must be made public.

Findings from revision studies are usually used to refine preliminary data and data collection programs for the subsequent periods. These studies also investigate the source of errors and fluctuations in the data and are used to make appropriate adjustments to the data. Although some of these findings may routinely be analyzed, they may mostly be used for internal quality control purposes. However, good practice suggests that these studies must also be published as they give insight in methods used and add to transparency in the production of statistics on respondents who have similar characteristics to the target population of the survey.

5 Relevance

Relevance of statistical information reflects the degree to which the statistical product meets the needs of users.

a. Key components

- Identify the need to conduct the survey or collect data
- Identify users of the statistics and their needs
- Determine whether users are satisfied with statistics produced
- Monitor user needs and incorporate their feed back into the design process

b. Indicator and Standards

User Database

Quality Indicator	**5.1**	The internal and external users of the data have been identified.
Standards	**5.1.1**	An up-to-date user database must exist.
	5.1.2	A documented process to identify user needs must exist.

Creation of User Database and Identification of User Needs

Measuring relevance requires the identification of user groups and their needs. Since needs evolve over time, a process for continuously reviewing programs in the light of user needs and making necessary adjustments is essential. The data-producing agency should ensure that processes are in place to ascertain user needs and what they use the data for. To ensure that the contents of a product reflect the needs of intended users, data producers should consider user needs early in the development process.

Data producers should have strategies in place to collect user feedback on the utility of their products and solicit recommendations for making data more useful. This strategy should include the development of a database containing a list of users classified according to different categories, their contact details, specific usage, classification of users by usage, and their needs. This will allow the data producer to categorize statistical products by the different classes of users. The user database should be kept up-to-date. The following guideline for the broad classification of users can be used:

a. government(central and local)
b. research and educational institutions;
c. private sector;
d. media;
e. international agencies;
f. non-governmental organizations (NGOs);
g. trade unions and professional associations; and
h. others not classified elsewhere.

The following are some mechanisms that may be used to identify user needs and obtain feedback from current users of products on their level of satisfaction, and to identify potential new markets for information:

- user feedback through User Information Service (UIS) and customer satisfaction surveys;
- user feedback on existing products and services;
- a statistics council/board can provide overall advice on policies and priorities for statistical programs;
- advice provided by the national statistics office (NSO);
- consultative groups e.g. through stakeholder workshops, user producer meetings;
- professional advisory committees in major subject areas e.g. GDP, trade, employment;
- participation of the head of the data-producing agency in Cabinet and High Level Consultative Council, also keeps the producing agency's management aware of current and emerging issues within government and private sector, e.g. Business Botswana, Ministry of Labour and Home Affaires or Bureau of Employment
- liaison through the NSS and consultation arrangements with districts or local government officials, e.g. Inter-Agency Statistics Committee (IASC);

- periodic liaison with business associations and labour unions that help to understand information needs and reporting preferences in the business sector, e.g. Business Botswana, Committees formed as a result of MOUs signed with sectors;
- ad hoc consultations with interest groups to provide input to the design of programs such as censuses and surveys, e.g. Technical Working Groups, Reference Groups, etc.
- bilateral liaison with major users and with foreign statistical agencies, and multilateral liaison through international organizations, e.g. international statistical institutes helps to identify information needs and emerging reporting obligations, e.g. meetings such as SADC Statistics Commission, UN Statistics Commission, SHASA, African Agenda 2063.

User Needs and Usage of Statistical Information Reporting

Quality Indicator	**5.2**	User needs and the usage of statistical information analyzed
Standards	**5.2.1**	A report containing the findings of the usage of statistical information must be available
	5.2.2	Findings of user needs must be assessed and availed to users

The analysis of user needs should be conducted with the intention of assessing and translating them into statistical objectives. This will allow the producer of statistics to define the statistical parameters of the survey (indicators, precise definitions of the indicators, domains of study, etc.). An analysis of the usage of the data provides feedback on information gaps, the adequacy of the data and their limitations. An active program of analysis should be put in place and should be nurtured through several mechanisms, viz.:

a. direct analysis of user needs through customer satisfaction surveys, focus groups;
b. number of queries, requests, complaints received. For example Statistics Botswana has put in place a system Customer Relationship Management System (CRMS) for this purpose, etc.;
c. published articles with external authors, often academics and peer review processes for these articles;
d. feedback in reaction to and commentary on analytical results;
e. decentralized analysis in subject-matter divisions such as demographic, economic or poverty analyses; and
f. contracts with external analysts to produce analytical reports for the data-producing agency.

This information provides the basis for management to decide on revising the data production program. After gathering user needs, they need to be translated into statistical requirements. The results of user needs analysis will provide information regarding estimates, scope, the period for data collection, geographical detail, data items to be covered, target population and other parameters.

Changes based on User Needs Assessments

Quality Indicator	**5.3**	Changes are made as a result of user needs assessments
Standards	**5.3.1**	The results of the user needs assessment must influence decisions on the statistical value chain of the survey or on administrative data collection systems, where feasible. Documented reasons for not implementing user needs must be provided as feedback to users.

After using a variety of mechanisms to keep abreast of users' information requirements, the data-producing agency should identify areas of weakness and gaps; and provide room for change. The results of the user needs assessments should be built into the data production processes, and

influence decisions on the design of the survey or series. However, costs, respondent burden, public sensitivities, and the agency's capacity and expertise need to be taken into account when changes are considered. Some of the features of the process that are particularly important to managing this part of relevance are as follows:

a. an annual strategic planning meeting to identify major corporate priorities to be addressed in the coming long-term planning round;

b. an invitation to program managers to submit new initiatives that would respond to user needs, especially in areas identified as corporate priorities such as Sustainable Development Goals (SDGs) indicators; and

c. a review and screening of proposals from key users.

Customer/User Satisfaction Survey

Quality Indicator	**5.4**	There is a process to determine the satisfaction of users with the statistical information.
Standards	**5.4.1**	A formal process to determine the satisfaction of users with the statistical information exists.

The evaluation of whether users' needs have been satisfied should be carried out using all efficient available means. User satisfaction surveys provide the best scenario, but failing which, proxy measures and substitute indicators of user satisfaction should be produced using auxiliary means such as publication sales, number of questions received, complaints, etc. The results of the assessments should always be built into the corporate processes and influence decisions on the design of the survey. The data-producing agency is required to put in place processes that monitor the relevance of its existing programs. These program should identify new or emerging information gaps that the current program is not filling.

A user satisfaction index must be calculated based on the survey results. The following serves as a guideline for the calculation of a user satisfaction index:

- the index should be based on the opinion of users of a statistical product;
- the index score runs from 100% (total satisfaction) to 0% (total dissatisfaction).

6 Accuracy

The accuracy of statistical information is the degree to which the product correctly describes and or estimates the phenomena it was designed to measure. Accuracy refers to the closeness of the values provided to the (unknown) true values.

a. Key components

- Assessment of sampling errors where sampling was used.
- Assessment of non-sampling errors including
  - Frame coverage errors (coverage of data collection in relation to target population)
  - Measurement errors (Data collection errors)
  - Processing errors (Data capture, coding, editing, …)
  - Non-response errors (response rates with a view to determine usability of the data)

- Assessment of the impact of imputation
- Assessment of source data accuracy or inconsistency problems with register based statistics

b. Indicators and standards

Measures of sampling errors

Sampling error is a statistical analysis arising from the sample design. It is the difference between a population value and an estimate based on a sample. It results in a deviation of the selected sample from the true characteristics, traits, behavior and qualities of the entire population. This covers a lot of aspects in the design such as unrepresentativeness of the sample taken and lack of randomness in the sample which could result in bias estimation of population parameter. This would ultimately make inference not plausible.

Measures of Sampling Errors

Quality Indicator	**6.1**	Measures of sampling errors for key variables are calculated. Amongst others these are: • standard error (SE) • coefficient of variation (CV) • confidence interval (CI) • mean square error (MSE) • design effect (DEFF) • Intracluster Correlation (ICC).
Standards	**6.1.1**	Measures of sampling errors must be calculated for the main variables. They must be available for other variables on request.
	6.1.2	Measures of sampling errors must fall within acceptable standards. At a minimum, the following must be calculated: standard error, coefficient of variation, confidence interval, mean square error. The low accuracy of variables (if this exist), is explained.
	6.1.3	Scientific Sampling Techniques must be used

Best practice dictates that sampling errors for key variables should be calculated and the expectation is that they should be within the acceptable level, failing which should be explained in the metadata. A scientific sampling technique must be used. For example in Statistics Botswana a commonly used sampling technique is a two stage stratified sampling technique in exception of prices collection, where purposive sampling technique is employed.

These are some of the measures of sampling errors that are needed to show the precision of data. The absence of these measures makes the results of the survey highly questionable;

a. Standard error (SE): This is the measure of variability. In sampling theory, the standard error is replaced by root square error (RMSE). The standard error gives users an indication of how close the sample estimator is to the population value: thus the larger the standard error, the less precise the estimator.

$$SE = \sqrt{Var(\hat{\theta})}$$

b. Coefficient of variation (CV): This is the measure of relative variability of an estimate; also called relative standard error (RSE). It is the ratio of the standard error to the sample mean.

$$CV = \frac{\sqrt{Var(\hat{\theta})}}{E(\hat{\theta})}$$

c. Confidence interval (CI): Confidence interval measures the range within which an estimate can be found a certain number of times, e.g. about 95% of the time the estimate will fall within the confidence interval. The narrower the CI the better the estimate.

d. Mean square error (MSE): This is the sum of the variance and the bias squared and it is used to measure the total effect of bias and variability in estimates

e. Design effect (DEFF): It is the ratio between the standard error using the given sample design and the standard error that would result if a simple random sample had been used. That is, it measures the deviation of the selected design or technique for a particular survey from simple random sampling method. Its calculation will depict the effect of complexity of the survey design. The larger it becomes, the lesser the accuracy on the survey design. For example a DEFF value of 2 indicates that the sample design is twice as efficient as simple random sample. A further increase in DEFF indicates an increase in the sampling error. In most cases the increase is mainly due to the use of more complex and less statistically efficient designs.

f. Intra- Correlation Coefficient (ICC): As complex designs involve clustering and stratification, the clustering effect has been controlled by fixing the number of households to be interviewed in each Enumeration Area. Due to homogeneity factor, it is assumed that the cluster size will serve the optimum purpose and increasing it will have a tendency of increasing the standard errors, hence the inter-correlation coefficient (ICC) calculation. High correlation of errors has a tendency of distorting the results of the survey. For example, for 2013 Botswana Literacy Survey, the ICC was reasonable at 5 percent level.

The vast number of different study variables or population characteristics and the different domains of interest in a survey make it impractical and almost impossible to calculate and publish standard errors for each statistic (estimated value of the population variable or characteristic) and for each domain individually. Best practice dictates to publish SE, CV, CI, and MSE for the key study variables on selected domains. It should be noted that although the standard states that measures of sampling error must fall within acceptable standards; the extent of the errors is survey-specific and has to be established during the planning of the survey. The sampling errors should always be there and included in the metadata.

Example of sampling errors calculated from the 2017 Botswana Demographic Survey

Variables	Mean	Standard error (SE)	Confidence Interval [95% C.I]	Coefficient of Variation (%)
P35m: How many MALE children have been born alive?	2.161371	.5812015	(0.58 ; 3.44)	65.11
P35f: How many FEMALE children have been born alive?	2.13698	.5795491	(0.85 ; 3.43)	65.22
P38m: How many of the children have died? MALE	1.42517	.3664921	(0.58 ; 2.27)	64.0
P38f: How many of the children have died? FEMALE	1.345035	.3281095	(0.50 ; 2.19)	53.23
P39m: How many children have been born alive by … (past 12 months)	1.010724	0.0212177	(0.74 ; 1.28)	9.83
P40m: How many of these children have died? (past 12 months)	0.0213904	0.0327501	(-0.1195 ; 0.1623)	32.27
D6: How old was … in completed years at the time of death?	59.17	4.04751	(51.11 ; 67.23)	45.50

Sampling Techniques

Given that most sample designs have one or more of the following three characteristics: unequal probabilities of selection, stratification, and clustering, it is important to ensure that appropriate scientific techniques for the estimation of variance in sample surveys are identified, implemented and documented. Variance estimates should be derived for all reported point estimates whether reported as a single, descriptive statistic, or used in an analysis to infer or draw a conclusion.

The selection of the Primary Sampling Units (PSUs) or Enumeration Areas (EAs) involves procedures and formulae to ensure a well-balanced representative sample. Therefore, the selection is done using probability proportional (pps) to a measure of size, being the number of households. In general, a wide variability in the number of households across PSUs has adverse effects on the fieldwork logistical planning.

Measures of Non Sampling Errors

Quality Indicator	**6.2** Measures of non-sampling error are calculated, viz • Frame coverage errors • Misclassification errors • Systematic errors • Measurement errors • Processing errors.
Standards	**6.2.1** The extent of measures of non-sampling errors must be kept to an acceptable level Metrics: $a = \frac{\sum \lvert \text{final weights} - \text{design weights} \rvert}{\sum \text{design weights}}$ $b = \max\left[\frac{\lvert \text{final weights} - \text{design weights} \rvert}{\text{design weights}}\right]$ Design weight is the product of the inverse of probability of selection both at EA and Household level (W1*W2) & final weight is the product of weights calculated at EA level (W1), household level (W2) & result code (response) W3 (i.e. Final Weight is (W=W1*W2* W3)

Non-sampling errors occur at all points of the data collection and processing procedures and decrease the accuracy of the data. Thus it is important to estimate their relative weight in the total error and allocate appropriate resources for their control. Non-sampling errors should be controlled and reduced to a level at which their presence does not obliterate the usefulness of the final sample results. Unlike in the control of sampling errors, this error may increase with increases in sample size. If not properly managed, non-sampling errors can be more damaging than sampling errors for large-scale household surveys. When possible, estimate the effects of potential non-sampling errors which include, amongst others, coverage error, measurement errors due to interviewers, respondents, instruments, and mode; non-response error; processing error, and model assumption error.

Frame Coverage Errors

Coverage error is an error that arises from not being able to sample from the whole or complete target population, i.e. businesses/establishments/households. In instances where the sample was drawn from an incomplete frame, there must be a statement to this effect and its causes, supporting the estimates in the metadata.

This section ensures that necessary steps are taken to develop and maintain - sampling frames, and that coverage of sampling frames is evaluated and documented. There is need to assess the coverage of the survey in comparison to a target population, for the population as a whole and for significant subpopulations. This may mean assessing the coverage of a sampling frame (e.g. a business register by industry), the coverage of a census that seeks to create a list of a population (e.g. the coverage of a census of population by district or by age and sex), or the coverage of a sampled survey in comparison to independent estimates of the target population (e.g. the difference between sample-based population estimates and official population estimates). Frames are constructed and maintained through the development of both primary and secondary sampling units registers.

The statistical units register is a list or equivalent procedure of identifying population units that are part of the survey population. The statistical units register is then maintained through updates from the source administrative register (introduction of new units into the survey population and removal of dead units from the survey population based on updates of the administrative register). Coverage is defined in terms of accuracy and completeness of the list of study units. Thus divergences between the frame and the target population are the sources of coverage errors. There are two types of coverage errors: under- coverage and over-coverage errors.

Under Coverage Errors

Standards	**6.2.2**	Delays between newly-registered administrative units and the corresponding statistical unit births must be known. Update procedures are enforced to correct for under-coverage
	6.2.3	The measures of under-coverage fall within acceptable standards

Under-coverage occurs when some units which are supposed to be included in the frame are excluded (i.e. when the sampling frame is incomplete, SBR in our case), thus giving them no chance of being selected, and results in biased estimates. Under-coverage may happen due to delays in update procedures, lost registration applications for businesses or residential units, units that are alive (active units) and are marked improperly as dead on the main frame, and thus not appearing on the main sampling frame. The characteristics of the excluded units introduce a bias in the estimates; for example, if the frame is based on all units which have an email address, it is biased towards units which are technologically enabled in this way. This is different from non-response. The effect of under-coverage may also underestimate the variance. This is so, if the excluded units differ from the rest of the sampled population.

$$x = \frac{\text{(No.of units not in frame)}}{\text{(Total no.in the frame+units not in frame)}}$$

Over Coverage Errors

Standards	**6.2.4**	Delays between de-registering of administrative units and the corresponding statistical unit deaths must be known. Update procedures are enforced to correct for over- coverage
	6.2.5	The duplication rate must be at an acceptable level

Over-coverage occurs when units are present in the frame and in fact do not belong to the target population or to units not existing in practice. This may be due to time lags in taking death of a business or person into account, or units which are dead and then marked as live (active units) improperly on the main frame, and thus incorrectly appearing on the main sampling frame. The effect of over-coverage will underestimate the variance.

Duplication in the Frame/Register

Duplication in the frame or register may contain multiple (duplicates) listings. That is, target population units may be present more than once in the frame. This may arise when the frame is created from different or multiple sources of data providers. Duplicated units have a high probability or chance of being selected more than once. Since a unit which appears more than once on the frame has a greater chance of being selected, the sample is biased towards these duplicated units. If the variable which is used as a sampling unit contains the same information, the effect will be to underestimate the sampling variance.

The most effective way to reduce coverage error is to improve the frame by excluding erroneous units, duplicates and updating the frame regularly, through fieldwork (use of SBR Review Forms or administrative data sources.

Estimating the rate of duplicate records on the sampling frame gives an indication of the extent of the problem. The duplication rate may be estimated using the following formula:

$$x = \frac{\text{No.of units duplicated}}{\text{Total no.of units in the frame}}$$

Frame Coverage Errors Continued

Quality Indicator	**6.2.6**	The proportion of units which are out-of-scope must be at an acceptable level
Standards	**6.2.7**	The proportion of units which are misclassified must be at an acceptable level
	6.2.8	Systematic errors must be identified and reported

Out of Scope Units

The out-of-scope units (i.e. ineligible units) are units that do not belong to the target population or do not form part of the sampling frame but have been included in the frame. These units have a

tendency of causing over coverage of units. Estimating the rate of ineligible units gives an indication of the reduction in accuracy. This may be estimated as follows;

$$x = \frac{\text{No.of units out of scope}}{\text{Total no.of units in the frame}}$$

Misclassification Errors

Misclassification errors occur when an entity is put into the wrong category – usually as a result of faulty measurements or wrong classification, e.g. classifying a business as manufacturing whereas it falls within creative arts industry.

Estimating the rate of misclassified units on the sampling frame gives an indication of the loss of accuracy. This may be calculated as follows;

$$x = \frac{\text{No.of units misclassified}}{\text{Total no.of units in the frame+not in the frame}}$$

Systematic Errors

Systematic errors are referred to as 'bias' as opposed to random errors. Systematic errors are errors that are committed consistently over time either by interviewers or respondents. They determine the extent of bias introduced for both administrative records and surveys. These errors cannot be detected during editing, for example, consistent misunderstanding of a question in a questionnaire during data collection can cause this kind of problem.

Measurement Errors

Measurement error is the difference between a measured value of a quantity and its true value. It occurs from failing to collect the true value from the respondents. It usually occurs during the data collection phase. Commonly known sources of measurement errors are the survey instrument used, mode of data collection, the interviewer and respondent, e.g. ambiguous questions. Measurement errors are not usually calculated. However, they are usually identified during editing processes by comparing responses to different questions of the same value, for example age and date of birth. In any case, the results should have a statement defining the error and a description of the main source of the error. Describing processes to reduce measurement errors indicate to users the accuracy and reliability of the measures.

The data quality report should ideally reflect on the bias due to these measurement errors for the main variables. However, since this is often difficult to achieve, providing evidence of measurement errors and a fair idea of their magnitude is sufficient.

In Statistics Botswana there are measures put in place to reduce the extent of measurement errors. These include the following but not limited to;

a. Engagement of stakeholders, the formation of Task Teams, Technical Working Groups and Reference Groups to qualify instruments and related equipment
b. Extensive training of trainers on the questionnaire and all other processes in relation to data collection
c. Extensive training of supervisors and enumerators
d. Pretesting and/or piloting of all instruments and other related processes of data collection
e. Quality control checks during data collection

Quality Control Processes

Standards Cont	**6.2.9**	Every kth statistical unit is independently double collected. The two outputs must be compared and corrective action must be taken. Records must be kept
	6.2.10	Data collection error rates calculated from fieldwork records must be at an acceptable level
	6.2.11	The effects of data collection instruments must be determined, reported and measures are taken
	6.2.12	The effects of the data collection mode must be determined, reported and corrective measures are taken

Quality Control Checks

The data-collecting agency should develop protocols to monitor data collection activities, with strategies to correct identified problems. This will include good practice to minimize cheating by data collectors, such as protocols for monitoring interviewers and re-interviewing respondents. One method usually practiced by data-collecting agencies is to appoint data collection supervisors who independently verify data collected by enumerators. This is done by re-administering the questionnaire/form to the same statistical units.

Questionnaire Effects

The data collection error rates detected by the fieldwork records must be very low. The standard requires every kth statistical unit to be checked independently.
Questionnaire effects (or collection instruments) may be directly due to the layout of the questionnaire that is being filled in or due to the type of questions that have been asked. Examples may include some of the following;

a. Leading or presumptuous questions
b. Absence of a correct scale/unit of measurement
c. Sensitive questions can lead to false responses or non-responses.
d. Close-ended questions can limit the respondent.
e. Language of questionnaire also can lead to problems, especially misunderstanding of questions.
f. Long questionnaires can cause respondent fatigue and thus lead to incorrect information being provided.

Data Collection Mode Effects

The choice of data collection method can lead to errors. For example, data can be collected via post, face-to-face interviews, telephone interviews, self-enumeration, administrative record system, direct observation or via the Internet. Data collection mode effects need to be investigated, where data collected using one or more methods are compared.

Some of the different challenges/effects based on different modes of data collection include the following;

a. Telephonic interviews exclude people without access to a phone.
b. Face-to-face interviews may not occur due to natural disasters in some locations.
c. Problems can occur through failure to understand handwriting, and/or the type of writing tool used.
d. Postal questionnaires are often not mailed back and also may not be completed by the intended participant.
e. Refusal and non-response can occur in all methods.

It is emphasized that the effect of the data collection mode should be determined and reported.

Quality Control Processes Continued

Standards Cont	**6.2.13**	The effects of the interviewers must be determined, reported and corrective measures are taken
	6.2.14	Respondent effects must be determined, reported and corrective measures are taken
	6.2.15	Proxy responses must be separately categorized (flagged) and must be at an acceptable level

Interviewer Effects

Interviewer effects occur when interviewers–

- lead respondents in the way they phrase or ask the question
- guess responses instead of ask the respondent;
- deliberately cheat by completing the forms "sitting under a tree";
- do not handle questionnaires with proper care;
- and respondents encounter a language barrier;
- are not allowed to enter gated communities.

It is helpful to provide information around the minimum qualification required for an interviewer, the intensity of training given to interviewers, the number of training days and the rate of success on the skills test for interviewers.

Respondent Effects

The extent of sensitivity of questions may lead to item non-response. Respondents may not be available for the interview, or even may refuse to respond for security or other reasons. Encouraging respondents to participate maximizes response rates and improve data quality. The following data collection strategies can also be used to achieve high response rates:

a. Ensure that the data collection period is of adequate and reasonable length;
b. Send questionnaires with covering letters describing the survey to respondents in advance;
c. Plan an adequate number of contact attempts (callbacks).

Train interviewers and other staff who may have contact with respondents in techniques for obtaining respondent cooperation and building rapport with respondents.

Proxy Response

Proxy response is a response made on behalf of the sampled unit by someone other than the unit. This is the rate of complete interviews by proxy. It is an indicator of accuracy as information given by a proxy may be less accurate than information given by the desired respondent. Most survey programs do not identify proxy responses. It is however good practice to separately categorize proxy responses while releasing data. This may be calculated as follows for both fully or partially completed responses;

$$\text{Complete proxy} = \frac{\text{Number of units with complete proxy response}}{\text{Total number of eligible units}}$$

and

$$\text{Unit response} = \frac{\text{Total number of responses (Total number of eligible units) Less Number of units with complete proxy response/}}{\text{Total number of eligible units}}$$

Data Processing Errors

Processing errors are errors that occur during data processing (entry, coding, editing and imputation), and cause the recorded values to be different than the true ones. After the data have been collected, they go through a range of processes before the final estimates are produced. These include data coding, data capturing, editing, weighting and tabulating.

Standards Cont	**6.2.16**	Data entry error must average an acceptable accuracy rate
	6.2.17	Coding error must average an acceptable accuracy rate
	6.2.18	Editing rate must average an acceptable level
	6.2.19	Editing failure rate must average an acceptable level

Data Entry Errors

Data entry error occurs when a character or item is erroneously entered into the computer, enters items at the wrong location, or omits some of the items recorded in the questionnaire. The experience and honesty of the data entry personnel has an effect on the period of data entry and quality of data captured.

Training and regular checking of survey personnel can minimize data entry errors. A questionnaire which is not designed for scanning may give faulty results if it is scanned. That is, for effective scanning results, questionnaires should be designed for such. Otherwise a number of questionnaires will be rejected if the scanning software misinterprets a character. Normally the rejected characters are collected and keyed in manually.

Data Coding Errors

This is the error that occurs due to the wrong assignments of codes to data. They need to be assessed depending on the coding methodology use. For example manual coding schemes, classification

codes are used where coders are trained on their use; or an electronic program is developed to check for the correctness of codes (consistency checks).

The dataset should be coded to indicate any actions taken during editing, and/or retaining the unedited data along with the edited data. When setting up a manual coding process to convert text to codes, create a quality check process that verifies at least a sample of the coding to determine if a specific level of accuracy coding is being maintained.
Editing Failure Rates

Editing errors arise as an effect of checking for inconsistencies and outliers. Editing failure rates express the extent of distortion occurring between the raw and the edited data.
Prior to imputation the data must be edited. Data editing is an iterative and interactive process that includes procedures for detecting and correcting errors in the data. As appropriate, check data for the following and edit if errors are detected:

Responses that fall outside a pre-specified range;

a. Consistency, such as the sum of categories matches the reported total, or responses to different questions are logical;
b. Contradictory responses and incorrect flow through prescribed skip patterns;
c. Missing data that can be directly filled from other portions of the same record;
d. The omission and duplication of records; and
e. Inconsistency between estimates and outside sources.

Data editing must be repeated after the data are imputed, and again after the data are altered during disclosure risk analysis. At each stage, the data must be checked for credibility based on range checks to determine if all responses fall within a pre-specified reasonable range. Consistency checks must also be conducted across variables within individual records for non-contradictory responses and for correct flow through prescribed skip patterns. Completeness checks must be made based on the amount of non-response and must involve efforts to fill in missing data directly from other portions of an individual's record.

Data Processing Errors Continued

Standards Cont	**6.2.20**	The imputation rate for item non-response must average an acceptable level
	6.2.21	The imputation rate for unit non-response must average an acceptable level
	6.2.22	The model assumption must be stated. All models used in the estimation of statistics must be described

Imputation Rate

Imputation is a way of compensating for missing data by assigning a value. The imputation rate is the percentage of values which have been imputed. Imputation can be applied to statistics units, e.g. households/businesses/individuals that did not respond to certain questions or certain items that were not answered in the questionnaire. When imputation is done at the level of the statistics unit, it is referred to as imputation for unit non-response; otherwise it is referred to as item non-response. There are many techniques or models available to impute for missing values. Depending on the problem, good estimates can be obtained through multiple imputations which adopt a simulation approach, or weighted estimation numerical algorithms like the Expectation Maximization (EM) algorithm. This is further discussed below under the sections covering model assumption errors and non- response errors.

Model Assumption Errors

The estimation process of parameters in a survey process will often incorporate various models. These include generalized regression estimators, seasonal adjustment, sampling design, treatment of missing data etc. The use of a model is associated with various assumptions of the feasibility of the model. Using the wrong models might produce results that are counter-intuitive or wrong, and thus the assumptions made need to be revisited, resulting in the model being changed or amended.

Model assumption errors will most likely lead to bias in the final statistics. The variability of their parameter estimators will lead to increased variance of the statistics output. Prior to producing estimates, establish criteria for determining when the error (both sampling and non-sampling) associated with a direct survey estimate, model-based estimate, or projection is too large to publicly release the estimate/projection;

a. Develop model-based estimates according to accepted theory and practices (e.g. assumptions, mathematical specifications).

b. Develop projections in accordance with accepted theory and practices (e.g. assumptions, mathematical specifications).

Document methods and models used to generate estimates and projections to help ensure objectivity, utility, transparency, and reproducibility of the estimates and projections.

Non-response errors

Non-response is the failure of a survey to collect the data on all survey variables, from all the units designated for data collection in a sample and/or a complete enumeration. There are two types of non-response:

a. unit non-response, which occurs when no data are collected about a designated population unit (e.g. business or household or person); and

b. item non-response which occurs when data on only some part but not all the survey variables are collected about a designated population unit.

Data Processing Errors Continued

Standards Cont	**6.2.23**	Item non-response rate must be within acceptable levels
	6.2.24	Unit non-response rate must be within acceptable levels

Item Non-Response Rate

Item non-response occurs when respondents refuse to respond to some of the questions or if the interviewer forgets to capture some of the questions. It is the ratio of the number of units which have provided data at least on some variables over the total number of units designated for data collection.

Unit Non-Response Rate

Unit non-response refers to a particular survey variable and occurs when sampling units (e.g. businesses, individuals, households) refuse or fail to participate in the survey. Unit non- response results in under-estimation of characteristics of interest and it also introduces bias in the estimates. In panel or longitudinal data, this is referred to as attrition. Weights need to be adjusted to account for unit non-response.

Source Data Consistency

Quality Indicator	**6.3**	The extent to which the primary data is appropriate for the statistical product produced
Standards	**6.3.1**	Source data must be consistent with the scope, definitions, and classifications of the statistical product produced

It is important that all key variables in the source data used to produce the statistics have definitions and classifications that are consistent. For example, Statistics Botswana uses data collected by immigration officers from the Department of Labour and Home Affairs on travelers who pass monthly through Botswana entry/exit to produce its Tourism and Migration data using the consistent definitions and classifications as the ones used in Statistics Botswana.

Quality Assessment of Data from Primary Source

Quality Indicator	**6.4**	Data from the primary source have been quality assessed
Standards	**6.4.1**	Source data must be accompanied by a quality assessment report

The primary source of data is data that have not been collected by the user. These data may be sample survey, census or administrative record data. Ideally, the user and producer of these data should produce a quality declaration for the data used in producing the statistics. The user should then expect that data received are accompanied by a quality declaration from the primary producer. In the absence of the foregoing, it is incumbent upon the user to produce a quality declaration of its own.

Statistical requirements of the output should be outlined and the extent to which the administrative source meets these requirements stated. Gaps between the administrative data and statistical requirements can have an effect on the relevance to the user. Any gaps and reasons for the lack of completeness should be described, for example if certain areas of the target population are missed or if certain variables that would be useful are not collected, any methods used to fill the gaps should be stated in the declaration.

Frame Maintenance Procedures

Quality Indicator	**6.5**	Register/frame maintenance procedures are adequate • updates • quality assurance • data audit
Standards	**6.5.1**	Maintenance procedures of register/frame must be documented. Registers/ frames must be updated on a regular basis in line with what has been documented
	6.5.2	The impact of frame maintenance must be measured, monitored, analyzed and reported on

To improve and/or maintain the level of quality of the register or frame, maintenance procedures should be incorporated to eliminate duplications and to update for births, deaths, dormant, out-of-scope units and changes in characteristics. These procedures must ideally be continuous or be implemented as close as possible to the survey period.

There are various methods that can be used for this purpose, e.g. the use of alternative sources of data to complete the existing list and/or feedback from the collection procedure, some of which are as follows;

a. Maintenance based on the most reliable stable source of sufficient coverage must be identified.

b. Maintenance must consist of both automatic updates, and investigation cases.

c. Profiling and delineation of large and complex businesses in the case of business statistics.

d. Maintenance takes place to an acceptable level of improvement.

e. Maintenance performance is measured by a quality management framework.

f. Register improvement is based on Statistical Business Register (SBR) Feedback Forms undertaken to correct particular problem areas.

g. Quarterly Business Surveys also provide feedback regularly and is incorporated into update procedures.

In general, Statistics Botswana follows the Guidelines for building Statistical Business Registers in Africa for development and maintenance of the SBR.

Data Collection Systems Flexibility

Quality Indicator	**6.6**	Data collection systems are sufficiently open and flexible to cater for new developments
Standards	**6.6.1**	The system must be flexible enough to cater for new developments

From time to time, changes in legislation, definitions, classifications, etc. can cause a change in some main variables. Therefore the data collection system needs to be made sufficiently open and flexible enough to remain easy to use in the case of possible future developments that may or may not be foreseen at the development stage of the survey. Any current system needs to be periodically reviewed to determine whether it satisfies the current survey requirements, and not necessarily the requirements that were expected at the development stage. The practice of adapting core business processes of statistical collection to suit the demands of a data collection system must be avoided. The number of ad hoc methods of using the system must be kept to a minimum. If this number gets too large or results in the risk of poor quality statistics, the system will require a complete overhaul.

7 Timeliness and Punctuality

Timeliness of statistical information refers to the time lag between the reference point to which the information pertains and the date on which the information becomes available. Timeliness also addresses aspects of periodicity and punctuality of production activities within the statistical value chain. Punctuality of statistical product is the time difference between the date the data are released and the target date on which they were scheduled for release, as announced in an official release calendar and laid down by regulations or previously be agreed with users.

a. Key components

- Statistics production time
- Periodicity and Punctuality of statistical release.

b. Indicators and standards

Standard for Preliminary Results Release

Quality Indicator	**7.1**	Average time between the end of the reference period and the date of the preliminary results (the ratings are informed by the GDDS and the SDDS document)
Standards	**7.1.1**	The preliminary results must be released according to the prescribed standard

The reference period is the date or time to which the survey or census refers and upon which a decision is made whether to include or exclude the unit sampled or enumerated, e.g. for the 2011 Population and Housing Census, the reference period was referred to as 'census night' which was the previous night . Babies who were born 1 minute before 6am were counted, while those born after 6am were not counted as part of the household members.

Reliable data, fit for the purpose for which they were compiled, should be ready for dissemination as soon as is feasible after the period to which they relate. For some programmes, the release of preliminary statistical information followed by revised and final figures is used as a strategy for the timely release of statistical information. In such cases, there is a trade-off between timeliness and accuracy. The earlier the data are released, the less complete and accurate it could be. Nevertheless, preliminary data can be used to inform interim decision-making. The tracking of the size and direction of revisions can serve to assess the appropriateness of the chosen timeliness-accuracy trade-off. Where it is applicable to publish preliminary results, these should be communicated to users well in advance of the release date. It is desirable that the release of preliminary results be communicated when the survey manager makes public his/her schedule of key deadlines.

All results are to be released in accordance with prescribed standards. The General Data Dissemination System (GDDS) provides current best practice standards on periodicity and timeliness of macroeconomic and financial data as well as socio-economic data. The Special Data Dissemination Standard (SDDS) provides standards to guide countries that had, or might seek, access to international capital markets. Otherwise every release should be as per release calendar. For example, Statistics Botswana Service Charter provides that preliminary results should be released after four (4) months of data collection. In this case, the Service Charter should indicate the same timelines as per the release calendar.

Where no standard(s) exists, the decision is informed by user requirements balanced with the feasibility of such releases. This also applies to situations where only one set of results is applicable. Any deviations should be explained through an accessible medium and should clearly state that the results are preliminary findings.

Standard for Final Results Release

Quality Indicator	**7.2**	Average time between the end of the reference period and the date of the final results
Standards	**7.2.1**	The final results must be released according to the prescribed standard

Final results should be released before the next round of data collection or administrative record processing. In accordance with international best practice, the release of final results should follow the prescribed timeframes set by the GDDS and SDDS for specific releases including GDP, Trade and CPI. These should also conform to the Statistics Botswana Service Charter provision as well as the release calendar developed in house.

Implementation of Project Plan of Activities

Quality Indicator	**7.3**	Production activities within the statistical value chain are within planned timelines, viz.: • Data collection • Data processing • Data analysis • Dissemination • Archiving
Standards	**7.3.1**	Project plan/schedule of key deadlines related to the statistical value chain must be compiled
	7.3.2	Updates to registers must occur within clearly specified timeframes
	7.3.3	A protocol for the timely delivery of administrative data must exist and must be adhered to
	7.3.4	Data collection must follow the project plan/schedule
	7.3.5	Data processing must follow the project plan/schedule
	7.3.6	Data analysis must follow the project plan/schedule
	7.3.7	Dissemination must follow the project plan/schedule
	7.3.8	Archiving must follow the project plan/schedule

Preparation of a Project Plan

Before proceeding with statistical production activities, it is desirable to compile a project plan for the statistical programme. This plan should say what process follows which and how long each process will take. The project plan should be of sufficient detail to compile a schedule of key deadlines within the SVC. Compliance with this project plan and schedule should be closely monitored.

An overrun of one process could adversely affect the quality of output from another process, thereby affecting the quality of the overall statistical product. Anticipated delays across the SVC should be noted and planned for with lag time built into the project timeframe. This implies that overruns for each of the phases in the statistical value chain are anticipated and planned for. There must be a considerable time lag between surveys to ensure smooth running of different survey phases, thus improving on the quality of data. A well thought survey schedule is necessary. For example Statistics Botswana has a ten year inter-censal household survey program and this needs to be followed closely to avoid overruns.

Timeframes to Update Registers

Ideally, the time frames for updates of registers should be planned not only according to the requirements of the data collection program but also by taking into consideration the needs of the users of the registers. Those who have the responsibility for maintaining the registers should make sure that the time frame has been adhered to and made known to all users.

Where administrative records were used in compiling statistics and to ensure their timely receipt, it is beneficial to have a formal arrangement between the statistics-producing agency and the data collection agency. Formalized arrangements through SLAs or MOUs will protect both parties since it will unambiguously indicate what each agency is responsible for. For example, the MOU signed between Statistics Botswana and Botswana Unified Revenue Services for exchange of data in compiling trade statistics, GDP and updating SBR. This ensures that statistical activities are not negatively affected due to a delay in receiving data.

Project Plan Implementation

It is important to monitor the timely implementation of the project plan, and not only concentrate on the dissemination date, as this will ensure that justice is done to all production processes viz data collection, processing as well as analysis; since they contribute towards quality statistical outputs.

Survey managers should allocate sufficient time for all processes within SVC and therefore expected to be meticulous with regard to updating project plans and schedules of key activities. The documentation should be part of the operation and system metadata.

Archiving

Archiving involves the long term storage of the organization's records, documents and products that have enduring value and have therefore been designated as "permanent". The purpose of this is for preservation of the publications and future use without wear and tear. This includes micro data files that have been disseminated and also the master files.

All statistical publications and data sets which are eligible for archiving shall be digitally archived and they must follow the project schedule.

For example, in Statistics Botswana all iterations of each dataset is retained and all archival copies are securely preserved and migrated as technology changes, to ensure they are always accessible.

Periodicity of Releases

Quality Indicator	**7.4**	Periodicity of releases
Standards	**7.4.1**	The periodicity (e.g. monthly, quarterly, and annual) of releases must conform to a data dissemination standard

Data Dissemination Standard

The periodicity (annual, biannual, quarterly, and monthly, etc.) of all releases must conform to a dissemination standard and should be adhered to. Details of any time lag between the scheduled time and actual release dates for specific products (GDP, CPI, Merchandize Trade data, employment, etc.) must be given, and reasons for delays should be documented along with their effects on the statistical product in the metadata. When disseminating the results, the periodicity of the release

should be clearly identified and reported as part of the metadata with the periodicity displayed on the front cover of the publication; and in an advance release calendar, when the product is advertised. It is important to follow the organizational standard in this regard, but users also need to be consulted when considering a review of the data dissemination standard.

8 Accessibility

The accessibility of statistical information and metadata refers to the ease with which it can be obtained from the agency. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed. The cost of the information may also be an aspect of accessibility for some users.

a. Key components

- Catalogue systems are available in the parastatals/government or statistical agency
- Delivery systems to access information
- Measure of release calendar and delivery systems performance

b. Indicators and standards

Statistical Products Dissemination

Quality Indicator	**8.1**	Statistical products (e.g. data, metadata) are available to the public
Standards	**8.1.1**	Statistical products must be disseminated to the public

Statistical products and data must be disseminated to the public and made available to users through various media which include the following: statistical releases, CD-ROMs, posters, exhibition stands at conferences, conference papers/presentations/posters, computer printouts, Internet press releases, and billboards.

To ensure that the contents of the product reflect the need of intended users, authors should consider user needs early in the publication development process. A data dissemination plan should be available as early as in the planning of the product. Once the product has been approved for release the data producer should organize a meeting to review proposed dissemination strategies including press releases, targeted mailings, libraries, the number of copies to be printed, web release, the use of print on demand, and the use of both print and electronic announcements. For instance, in addition to the above data and statistical products, Statistics Botswana makes these available on portals which are linked to the website.

Administrative Records Policy

Quality Indicator	**8.2**	Rules governing the restricted availability of administrative records are well described and documented
Standards	**8.2.1**	A policy document having clear rules governing the restricted availability of administrative records must exist

Administrative records are usually collected for purely administrative purposes and help the collecting agency conduct it's day-to-day business. Personal/business details are often captured and this makes the data very sensitive, which in turn renders the use of the data for statistical purposes difficult because the confidentiality of the data may limit the accessibility of the data unless a legal framework is in place.

Procedures to request access to administrative records should be clearly described. This legal agreement is between the administrative record data collector and the data users. The arrangement should be via formal agreements such as MoUs, SLAs or letters citing confidentiality clauses. Recipients of administrative data must adopt a strict policy on the use and dissemination of the data under their custody. The use must be confined to statistical purposes only and must never be disseminated at a level where respondents would be identifiable

Data Sharing Media Channels

Quality Indicator	**8.3**	Types of media and/or channels used for sharing data amongst stakeholders are adequate and preserve confidentiality
Standards	**8.3.1**	Data must be accessible through various channels with mechanisms that ensure confidentiality

One of the key components necessary for maximum accessibility is the number of ways in which data are available to the users. The more the options for the user obtaining the data, the more accessible it is. The following media channels are commonly used to share data with users and producers; website, social media, portals, print outs, CDROMs, etc.

Data Accessibility Formats

Quality Indicator	**8.4**	Data are accessible in a user friendly format
Standards	**8.4.1**	The data must be available in different file formats

Data-producing agencies should make sure that data are available in formats that satisfy the requirements of users, for example, publishing data with SPSS, and Excel or ASCII format. A variety of dissemination techniques should be used to accommodate the majority of users. Innovative ways to disseminate data/products should also be explored.

Statistical Products Release Calendar

Quality Indicator	**8.5**	Statistics are released according to Release calendar
Standards	**8.5.1**	Statistics must be released according to the release calendar

The data-producing agency should have an advance release calendar that is published on a monthly basis or yearly on the Internet. It will contain information on which statistics are being released at a particular time. These schedules also contain the dates and time of release of such products. For instance, Statistics Botswana has an annual release calendar for its products, and it is published on the website. Some of the key releases include CPI, Merchandise Trade, GDP and employment products from administrative records, surveys and censuses

Statistical Products Release

Quality Indicator	**8.6**	Statistical releases are made available to all users at the same time
Standards	**8.6.1**	Statistical releases must be made available to all users at the same time.

All statistical releases or products should have an embargo date and time i.e. all users get the information at the same time.

Data Requests Policy

Quality Indicator	**8.7**	Statistics/administrative records not routinely disseminated are made available upon request
Standards	**8.7.1**	Statistics/administrative records not routinely disseminated must be made available, and the terms and conditions on which they are made available must be publicized.
	8.7.2	Special requests are considered and be met.

Statistics or information that is not disseminated routinely is referred to as special requests. Data-producing agencies are encouraged to have a policy in place on handling special requests. These requests should be answered promptly. Procedures to request access to confidential micro data should be clearly described in a format easily understood by users, and made readily available.

The availability of unpublished statistics and data, and the terms and conditions on which they are made available should be publicized. This may include the publication of users' details and uses of data.

User Support Services

Quality Indicator	**8.8**	User support services exist and are publicized
Standards	**8.8.1**	User support services must exist and widely publicized.
	8.8.2	User support services are effective.

User Support Services (USS) provides a single point of access to an organization's information. This service will promote the increased effective use of the organization's data products and services. In response to contacts, the staff will do the following:

a. help define the information requirements of the client;

b. provide data and information on the data producer's products and services;

c. develop customized, cost-efficient data solutions, facilitate and serve as the direct link to the rest of the organization's researchers, analysts, consultants and other technical experts; and

d. advise users to make the enquiries by telephone, fax, email, post and through the website Any statistical release should have people assigned to key user response roles; e.g. analysis, queries, etc. The contact details for these people must be communicated to the central USS for every product.

Data Dissemination Policy

Quality Indicator	**8.9**	A data dissemination policy exists and it is accessible
Standards	**8.9.1**	A data dissemination policy must exist and be accessible

A data dissemination policy is a document that guides the organization in dealing with data dissemination issues such as:

a. The nature of the data that are released (e.g. full data versus sample or aggregated versus disaggregated data);
b. confidential information;
c. cost of data;
d. periodicity of release;
e. metadata availability and limitations on published documents; and
f. choice of media.

The organization responsible for producing and disseminating the data has the right to review this document from time to time without prior notice. The policy should be available to users.

Pricing Policy

Quality Indicator	**8.9**	A pricing policy exists and it is accessible.
Standards	**8.10.1**	A pricing policy must exist and be accessible

Data collection, compilation, processing and dissemination are costly. A data producer may decide that users of the data must share the cost of the data. The amount of money charged may differ by type of users, e.g. student and academic, private and public sector organizations. In some instances, data are supplied free of charge to certain users such as students, government departments or government-financed research institutes, international organizations such as the World Bank and UN agencies. Where data are supplied free of charge, the user may be asked to provide the media (e.g. disk or CD) or pay a nominal amount that covers the cost of the media and shipment. The pricing policy should be made available to users.

Publications Catalogue

Quality Indicator	**8.11**	Catalogues of publications and other services are available to users of statistics
Standards	**8.11.1**	Catalogues of publications and other services must be freely accessible to users of statistics.

A major component of ensuring accessibility is providing efficient search mechanisms to help users find what they need. The Internet site should offer an array of search and navigation tools and features that permit users to discover information such as;

a. data and product browsers and search, by theme and subject;
b. catalogue search;
c. key word search (supported by terminology thesaurus);
d. search of the organizational library; and/or
e. guides to data, to search tools and methods

Given the current rate of technology change, the nature of both catalogue and delivery systems is evolving fast. The traditional printed catalogue are outdated, thus giving way to on-line catalogues of statistical products.

Users will determine the effectiveness in the accessibility of the catalogue. The feedback may be derived from;

a. automated usage statistics for the various components of these systems,
b. surveys for user satisfaction with particular products, services, or delivery systems; and
c. Voluntary user feedback in the form of comments, suggestions, complaints, or appreciations.

Metadata Accessibility

Quality Indicator	**8.12**	Metadata are readily accessible to users.
Standards	**8.12.1**	Minimum metadata required for interpreting the product must be accessible

To ensure the usefulness and usability of data files created, all data files or any statistical product should be accompanied by a readily accessible document containing metadata that clearly describes and explains the data. This is the minimum metadata required by users to know, locate, access or use data/statistics appropriately. This type of metadata can be found in the **Interpretability dimension.**

9 Interpretability

Interpretability of statistical information refers to the ease with which users understand statistical information through the provision of supplementary information (metadata and relevant supporting documents).

a. Key Components

- Metadata on concepts and definitions, classifications and methodology used within the statistical value chain
- Key findings giving the summary of the results;
- Presentation of statistics in a meaningful way.

b. Indicators and Standards

Metadata Documentation Standard

Quality Indicator	**9.1**	Documented metadata (definitional, operational, methodological, system and dataset) are sufficient to understand data
Standards	**9.1.1**	Metadata must be documented according to the accepted standards, guidelines or good practices

To make sure that information is interpretable, data producers are required to give descriptions of the underlying concepts, variables and classifications that have been used, the methods of collection, processing and estimation used in production of information and its own assessment of the quality of the information. Statistics released should be accompanied with complete metadata information, which is documented and complies with a metadata standard template. Completeness of metadata describes the extent to which metadata are available for the users and the extent to which it covers the topic, i.e. the metadata should be sufficient enough for users to replicate the output.

In the case of public-use micro data files, information regarding the record layout and the coding/ classification system used to code the data on the file is an essential tool to allow users to understand and use the data files. Different types of metadata as indicated above should be documented according to standard(s), or sourced from the standard, for example:

a. Concepts and definitions should be sourced from the relevant standard manual (compendium of statistical concepts and definitions) and made available to users;

b. Classifications should be sourced from the relevant standard coding system or document and made available to users;

c. Variables should follow the standard variable naming convention,(Household Classification Scheme);

d. There should be a standard template for documenting a record layout which provides proper descriptions for the dataset (flat file);

e. A standard metadata capturing template should be developed that provides an overview of minimum metadata required to explain the data.

Where a deviation from the standard is required, such deviation should be documented, including reasons for deviating and approval thereof. Metadata is a description of the data for the users to understand the data in detail. This includes, among others, description of the source, compilation, and methodology, time of dissemination, institution and persons responsible for the compilation.

Statistics Presentation Standard

Quality Indicator	**9.2**	Statistics are presented in a clear and understandable manner
Standards	**9.2.1**	The presentation of the statistics must be according to a standard

Statistics should be presented in a way that facilitates proper interpretation and meaningful comparisons (layout and clarity of text, tables, and charts). There is a need to have a standard for producing a report or statistical release, including a data tabulation standard (tabulation plan). Data should be published in a clear manner; charts and tables are disseminated with the data to facilitate the analysis. It should offer adequate details and time series. Analysis of current period estimates should be available. Depending on the intended audience and purposes, data of different degree of aggregation, sub-components and additional data should also be made available. For instance, Statistics Botswana has editorial guidelines which standardize the presentation of statistical products.

Key Findings Summary

Quality Indicator	**9.3**	Statistical releases contain a summary of the key findings as defined in the major objectives
Standards	**9.3.1**	Statistical releases must contain a summary of key findings

The statistical releases should contain a primary message that clarifies the interpretation of the data here referred to executive summary, which is directed at the media. Such commentary increases the chance that at least the first level of interpretation to the public will be clear and correct.

Private Bag 0024,
Gaborone
Tel: 3671300
Fax: 3952201
Toll Free: 0800 600 200

Private Bag 47,
Maun
Tel: 371 5716
Fax: 686 4327

Private Bag F193,
City of Francistown
Tel. 241 5848,
Fax. 241 7540

Private Bag 32 ,
Ghanzi
Tel: 371 5723
Fax: 659 7506

E-mail: info@statsbots.org.bw
Website: http://www.statsbots.org.

STATISTICS BOTSWANA