Bosna i Hercegovina
AGENCIJA ZA STATISTIKU
BOSNE I HERCEGOVINE

Босна и Херцеговина
АГЕНЦИЈА ЗА СТАТИСТИКУ
БОСНЕ И ХЕРЦЕГОВИНЕ

Bosnia and Herzegovina
AGENCY FOR STATISTICS
OF BOSNIA AND HERZEGOVINA

Ref. No.: /25

Sarajevo, 2025

QUALITY STANDARDS

OF THE STATISTICAL BUSINESS PROCESS
– GUIDELINES –

March 2025

CONTENTS

Page

Introduction

The guidelines set out in this document provide important guidance in the process of monitoring and implementing the recommended quality standards of the statistical business process. The recommended standards are partly prescribed by international regulations and partly based on the application of best methods and statistical practices ("Current Best Practices").

In the interest of producing reliable, relevant and internationally comparable statistics, the competent statistical institutions in Bosnia and Herzegovina are obliged to take all necessary measures to fully adopt the guidelines and recommendations set out in this document. All stakeholders involved in the production of official statistics in the country bear significant responsibility to make maximum efforts, in the coming period, to meet the needs of users of statistical data both domestically and internationally.

The primary purpose of this document is to describe, in one place, the individual phases of the statistical survey process. In addition to the description of each phase, indicative instructions (quality guidelines) are provided for each phase as well as for the process as a whole, which producers of statistics should follow to the greatest extent possible.

The production of statistical results is a demanding process in which quality must be continuously monitored and improved. For these reasons, any document that describes this process and provides guidance on quality standards must be continuously updated and further enhanced.

DIRECTOR

Vesna Ćužić

QUALITY IN OFFICIAL STATISTICS

Definition of Quality

When referring to the concept of “quality in statistics”, it is important to recognise that this is not a one-dimensional term. The concept encompasses a wide range of aspects, including various interrelated quality components.

ISO Standard 8402 (1986) defines quality as follows:

“Quality is the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied user needs.”

In line with this definition, and in the context of statistical production practice, two key questions arise:

1. What are the essential features and characteristics of statistical products?
2. Which requirements can be explicitly defined and which are implicitly implied?

Initial considerations on how to define quality in the context of official statistics were already undertaken in the mid-1990s by Eurostat. In 1998, the Eurostat Working Group on Quality defined the concept of “quality”, which was subsequently introduced into official statistical practice for the first time in EU Member States. The Working Group (the so-called *Leadership Group on Quality – LEG-Quality*) established the currently valid version of the definition in 2003 in the document *Assessment of Quality in Statistics, Methodological Documents – Definition of Quality in Statistics*.

Accordingly, quality in statistics is assessed against the following six quality criteria/dimensions:

- Relevance
- Accuracy
- Timeliness and punctuality of release
- Accessibility and clarity
- Comparability
- Coherence

Relevance

Relevance is the degree to which a statistical product meets the current and potential needs of users. It depends on the extent to which the produced statistics, as well as the concepts used (definitions, variables, classifications, etc.), reflect user needs.

Accuracy

Accuracy is defined as the degree of closeness between the estimated values (obtained as the final outcome of statistical processing) and the true, but unknown, population values.

Timeliness and Punctuality of Release

Timeliness of statistical results refers to the time lag between the last day of the reference period to which the data relate and the date on which the results become available, i.e. the actual release date.

Punctuality refers to the time lag between the actual release date of the data and the planned release date specified in the official publication calendar. If these two dates coincide, the release is considered punctual.

Accessibility and Clarity

Accessibility of statistical results/products refers to the specific physical circumstances under which data are available to users: where the data are physically located, ordering possibilities, publication calendars, clear pricing policies, availability of microdata and macrodata, and different formats and media (e.g. paper, files, CD-ROM, internet).

Clarity refers to the statistical information environment through which users access information: whether textual information, methodological explanations and documentation are provided alongside the data; whether the data are accompanied by graphical and other visual materials; whether information on data quality is provided; and whether additional information is available to users upon request.

Comparability

The purpose of the comparability component is to measure differences arising from the use of statistical concepts and definitions for producing comparable statistics across different geographical areas or different reference periods. Comparability is assessed through two subcomponents: comparability over time and comparability between countries.

Comparability between countries is achieved by adopting fundamental principles and definitions from international standards and EU regulations.

Comparability over time is achieved through the consistent use of the same methodology, definitions and concepts when conducting surveys over a series of years.

Coherence

Coherence between two or more statistical products refers to the degree to which the statistical processes used to produce those products are based on the same concepts (classifications, definitions and target populations) and harmonised methods.

As emphasised, quality in statistics comprises multiple factors. This multidimensionality is also reflected in the definition itself.

When considering quality dimensions individually, it becomes evident that each dimension represents an important factor influencing overall quality. However, when these criteria are examined in interaction, it becomes clear that achieving certain objectives may involve competing considerations.

This "trade-off" situation (compromises between output quality components/dimensions) becomes most apparent when comparing the accuracy dimension with the timeliness dimension. The key question is whether it is preferable to release accurate data with a delay or timely data that may be inaccurate. This is only one illustrative example.

It can even be argued that, when selecting any two or more quality dimensions, goal conflicts may arise. Such situations make it challenging to achieve an optimal quality outcome for official statistics.

In any case, considering all the above criteria, the objective is to optimise quality. An important role in this context is played by quality measurement. Given the multidimensional nature of quality, it is evident that quality in statistics cannot be expressed by a single quantitative indicator.

Rather, multiple assessments of individual components must be synthesised into an overall picture. The overall assessment of a statistical product also depends on how individual quality dimensions are weighted.

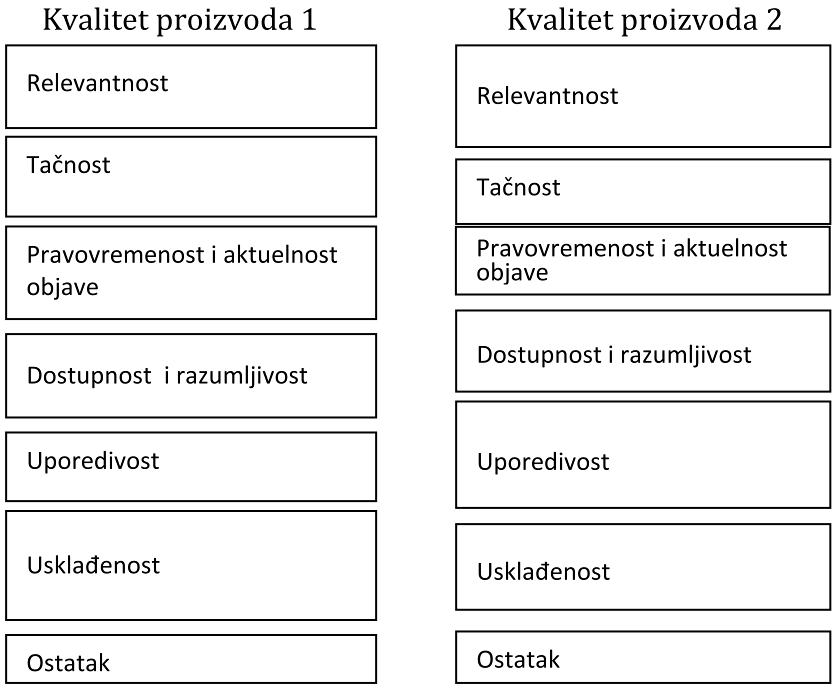

Figure 1: Weighting of individual quality dimensions may differ between two statistical products *(The width of the rectangle corresponds to the respective weight of a quality dimension.)*

Statistical Survey

Within the statistical process, the statistical survey occupies a central position. The concept of a statistical survey is very broad and encompasses many different practical implementations.

According to the "classical" understanding, a statistical survey is based on the direct collection of data from a randomly selected sample of observation units. According to the "broader" understanding (which is increasingly replacing the classical approach), a statistical survey includes various process implementations, particularly in the data collection phase.

Statistical surveys may be classified into groups depending on the mode of implementation, as follows:

By coverage:

- Census – the entire population is observed
- Sample survey – only a selected sample of the population is observed
- Derived statistics – statistical results are derived from existing statistical aggregates (primarily national accounts and balance of payments statistics)

By data collection method:

- Interviewing/surveying (by telephone or face-to-face)
- Self-completion (by post or electronic questionnaire)
- Other data collection methods (observation, monitoring, etc.)

By periodicity:

- Monthly
- Quarterly
- Semi-annual
- Annual
- Other periods (weekly, multi-annual)

By data source used:

- Statistical (primary) sources
- Administrative and other secondary sources

In practice, combinations of the above categories are frequently encountered. Many surveys use a combination of primary and secondary data sources.

QUALITY STANDARDS OF THE STATISTICAL BUSINESS PROCESS

Statistical products produced within official statistics cover a wide range of relevant socio-economic topics.

Depending on the topic, methods used to produce statistical products, applied concepts and definitions, as well as the presentation of results, may vary. However, as in any production process, there are common fundamental production steps that underpin the majority of official statistics.

The basic structure of this document is based on a general business process model, which represents a slightly adapted version of the internationally accepted *Generic Statistical Business Process Model (GSBPM)*.

The model comprises eight phases/processes, which are further divided into a number of sub-phases/sub-processes. The chapters and sub-chapters of this document follow the structure of the generic model. Each chapter provides a brief description of the overall process.

All sub-chapters consist of two sets: the first contains a general description of the sub-process, while the second provides guidelines for ensuring the quality of the process and sub-process. These guidelines serve as a checklist of quality components that should be taken into account when implementing individual steps.

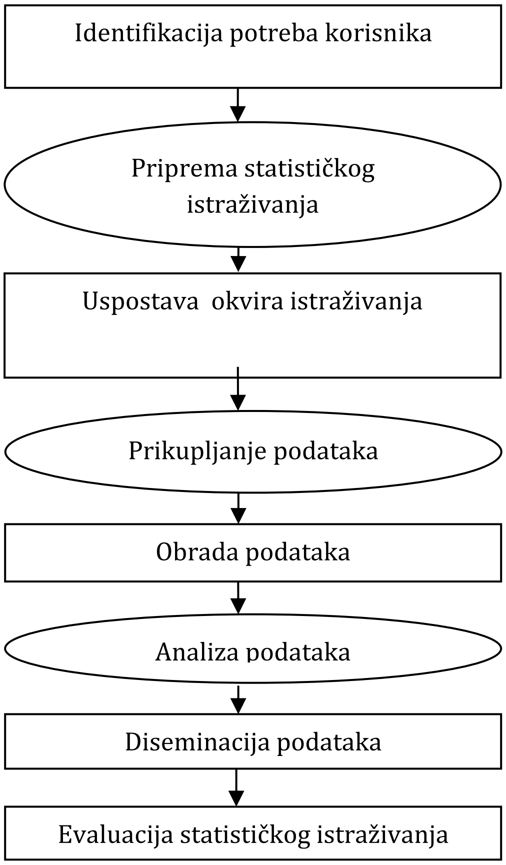
Identifikacija potreba korisnika
Priprema statističkog istraživanja
Uspostava okvira istraživanja
Prikupljanje podataka
Obrada podataka
Analiza podataka
Diseminacija podataka
Evaluacija statističkog istraživanja

Figure 3: Generic statistical business process model

1 IDENTIFICATION OF USER NEEDS

When identifying needs and analysing requirements for statistical data, attention should primarily be focused on external users of statistical data, who require such data for decision-making and analytical purposes, as well as on the requirements arising from EU regulations. Analysis of user needs and EU regulatory requirements provides the basis for planning future statistical surveys.

Expectations and needs of users of statistical results are continuously increasing and becoming more specific due to ongoing social changes, global processes and new initiatives at both international and national levels. Data needs are identified through continuous contact with users, while taking into account the capacities and operational frameworks of statistical institutions.

Currently, particular emphasis is placed on harmonising the statistics of Bosnia and Herzegovina with EU regulations in order to ensure international comparability.

1.1 Identification of Data Needs

The identification of data needs is initiated when data do not yet exist, or when existing data are insufficient to meet all user requirements.

Data needs arise from various users, including ministries and government bodies, the Central Bank of Bosnia and Herzegovina (CBBH), national and international institutions, as well as the professional community and the general public.

Through the process of identifying data needs, statistical institutions determine what users expect from the competent statistical authorities and which of these expectations can realistically be met. For this purpose, consultations with all interested users should be included in the process, using various forms of cooperation.

GUIDELINES

- The identification of data needs requires a comprehensive and systematic approach, involving all interested users as well as EU regulatory requirements.
- Various forms of cooperation with users should be organised on a regular basis, enabling statistical institutions to become familiar with user needs.
- In particular, information should be obtained on which data users require, when they require them, for what purposes, and in which form.
- When deciding on new data collection, the importance of meeting a specific data need should be assessed. The costs and benefits resulting from data collection should be carefully weighed.

1.2 Verification and Review of Sources – Data Availability

Before deciding to introduce a new statistical survey or to redesign an existing one, it is necessary to review all existing statistical surveys currently being conducted, as well as the content of administrative sources, in order to determine whether the required statistical data are already available from existing sources.

If such data are available, the next step is to assess the extent to which existing sources are aligned with the needs of the new survey, or to identify limitations that prevent the use of these sources to meet new data requirements (differences in methodology, periodicity, purpose of data collection, etc.).

The review and examination of sources is carried out primarily in order to decide whether, and to what extent, administrative sources can be used as a direct source of data.

If the reviewed and available sources do not contain the statistical data required by users, it is necessary either to redesign existing statistical surveys or to introduce a new statistical survey.

GUIDELINES

- It is necessary to assess whether the required data can be obtained by supplementing an existing survey, by using an administrative source, or whether a new survey needs to be planned.
- The quality of administrative sources should be assessed; where quality is adequate, administrative sources should have absolute priority over direct data collection.
- Data users should be specifically informed about methodological particularities, especially when statistical data are fully derived from administrative sources.
- Compliance with statistical confidentiality and the physical and information security of data from administrative sources must be ensured (access restrictions, access monitoring).
- Agreements on data transmission to the statistical institution should be concluded with the owners of administrative sources. Such agreements should define the content of transferred data, periodicity and method of data transmission.
- When establishing and regulating administrative sources, as well as when proposing changes to existing administrative sources, the participation of the competent statistical institutions in these activities should be ensured.

1.3 Preparation and Approval of the Business Case

For the successful implementation of a statistical survey, it is necessary to carefully plan the required human and material resources and to establish a timetable for carrying out all activities.

Precise resource planning and the establishment of a list of activities and deadlines are key to the efficient implementation of statistical activities. Each statistical survey should be included in the annual Work Plan (programme of statistical surveys). A newly established survey should be included in the Work Plan as a pilot survey, while a survey planned for regular implementation should be included as a regular survey.

Before the implementation of a statistical survey begins, a staffing and financial plan must be prepared and timely incorporated into the budget preparation process, as well as planning for additional material resources. The costs and benefits resulting from data collection should be carefully assessed.

Prior to the start of the survey, a pilot survey should be conducted in order to identify certain methodological aspects of the survey. The objectives of the pilot survey must be clearly defined, as the sample design needs to be adapted accordingly.

The sample design for the pilot survey depends on the purpose of the pilot. A pilot survey may be used to test the entire process or only specific parts of the process; most commonly, it is used to test questionnaire content, required sample size, sample design and data collection methods.

The sample size in a pilot survey is usually significantly smaller than in a regular survey. Following the pilot survey, results should be analysed in detail, and subsequent steps in the survey implementation should be determined on that basis.

GUIDELINES

- The objectives of the pilot survey must be clearly defined.
- Various methodological aspects should be tested in the pilot survey, including question formulation, unit and item response rates, etc.
- Results from the pilot survey must be used in the preparation of the main survey.
- When testing questionnaires, direct contact with the reporting unit is essential.
- Where necessary, procedures for statistical data processing and tabulation should also be tested in the pilot survey.
- It should be determined whether the proposed project is fundamentally feasible.
- The data collection method should be defined and clarified in advance; electronic data transmission should be given priority wherever possible. In cases where secondary sources are used, data owners should be involved in the planning process.
- For each project, a preliminary cost estimate and required budgetary resources must be calculated. In order to plan realistic cost estimates, experience from previously implemented similar or comparable projects should be taken into account.

2 PREPARATION OF A STATISTICAL SURVEY

Thorough planning is the initial phase of the production process, consisting of pre-designed future activities for the preparation and structuring of individual subprocess flows. Planning is a necessary prerequisite for obtaining a high-quality final product. This generally applicable rule also holds true for the process of producing statistics.

Planning should be directed toward a defined objective, which must, on one hand, comply with legal norms and, on the other hand, meet the requirements for adhering to quality standards. The purpose of planning is the operationalization of the subject of the statistical survey.

Activities undertaken in the planning process include:

- Specification of the basic theses regarding the subject, objectives, and results of the survey along with their essential characteristics. Using certain concepts and methods, we approach the design and definition of the project structure of the survey phases as well as the available resources;
- Once a decision on conducting a statistical survey is made, detailed planning and further differentiation of the survey phases are undertaken. All important aspects of conducting the survey must be considered and planned, such as deadlines and duration, required human and technical resources, and clarification of specific methodological issues;
- Creating an activity plan for conducting the survey (preparation of forms and instructions, defining the frame and creating a sample for a pilot survey, preparation of variables and output tables, conducting the pilot survey, analysis of pilot results, and preparation of the main survey).

Special attention in the planning process is given to the data collection process itself. Therefore, the question must be answered: where can we obtain the necessary data and how do we use it in the example of a survey?

Competent statistical institutions first and unconditionally use, as far as possible, all available data from other sources before conducting direct surveys.

Every institution maintaining a public register, as well as owners of official records and statistical data, are obliged to provide data necessary for implementing the Plans and Programs of statistical institutions in BiH. Only in cases where data provision is not possible in this way can a statistical survey be conducted via direct data collection.

When planning a survey, it is necessary to minimize the burden on reporting units. Whenever possible, surveys should be conducted on a voluntary basis and using samples, giving preference to sample surveys over full-coverage surveys.

The availability of budgetary resources is a basic prerequisite for undertaking any statistical survey activities. Preliminary cost estimation must cover all necessary costs required for conducting a given survey.

2.1 Design of Statistical Survey Products

This subprocess involves the detailed design of statistical results, products, and services to be produced, including related development tasks and the preparation of systems and tools used in the "Dissemination" phase.

Procedures that govern access to all confidential results are also developed here. Results should be designed to follow existing standards wherever possible, so inputs to this process may include metadata from similar or previous data collections, international standards, and information on practices in other statistical institutions.

This subprocess focuses on designing the layout and content of products (statistical tables, indicators, special publications, etc.) planned for dissemination. Product creation follows internally established procedures and dissemination policy.

2.2 Preparation of Methodology for Data Collection and Survey Implementation

This subprocess includes developing all necessary methodologies (methods, data collection instruments, variables, definitions, instructions, agreements and contracts with data providers, questionnaire content, dissemination plan, etc.). Preparing metadata descriptions of collected and derived variables and classifications is a key prerequisite for subsequent phases. We should use (as much as possible) harmonized concepts, variables, and classifications, but definitions must be adapted to specific needs. All definitions of concepts, variables, and classifications must be documented, as well as any deviations from recommended standards. The most appropriate methods and instruments for data collection are determined. Activities depend on the data collection mode (CAPI, PAPI, CATI, CAWI), including instrument testing. All formal agreements on data provision (e.g., technical protocols) are prepared.

In this subprocess, the target population is also determined, taking into account: the type of observation unit and characteristics defining the population; the geographic location of the observation unit; and the temporal (reference) period to which the characteristics of the population of interest relate.

2.2.1 Concepts, Definitions, and Classifications

Official statistics concepts form the foundation for numerically representing general issues related to social, economic, or other socio-political phenomena. Statistical surveys are based on appropriate, largely internationally comparable harmonized concepts.

Definitions represent operationalized descriptions within concepts, necessary for the concrete production of statistics (e.g., variable definitions). While a concept forms a general basis for many statistical projects, a definition clearly specifies how the relevant foundations (e.g., which variables are investigated) are determined for implementing a statistical project.

In this context, another aspect is important for assessing the quality of statistics: different surveys may describe a given topic using their own definitions, based on different concepts. This can lead to differing results and harmonization issues. To make the used concepts and definitions transparent to users of statistical products, sufficient metadata must be provided.

Classifications are the primary tool for producing official statistics. Competent statistical institutions apply a number of national and international classifications and nomenclatures.

GUIDELINES

- Definitions must be understandable and made available to users.
- All concepts, variables, and classifications used in the survey must be defined in detail, with special attention to potential deviations from standard concepts. Differences between phenomena measured and phenomena of interest to users should also be described (e.g., differences between target and observed populations).
- All concepts and definitions should be based on international standards.
- All relevant definitions, concepts, and classifications should be explained in the "Quality Report" for each statistical survey. This is particularly important for the purpose of statistics, statistical units, collected variables, calculated ratios and indicators, and classifications used.
- New classifications or revisions/modifications of existing classifications must be immediately entered into the classification database.

2.2.2 Statistical Business Register

The statistical business register plays a central role for all business statistics, as it ensures data coherence and is therefore considered a statistical product. The register serves as a frame for the business population, linking administrative data and transforming them for statistical purposes, linking microdata across observation units from different sources and surveys. Maintaining the statistical business register is one of the most important and fundamental tasks for official statistics. Its purpose is to collect, establish, and maintain register-relevant data. The statistical business register differs from an administrative register in that it is established and maintained according to EU Regulation 177/2008 exclusively for statistical purposes.

GUIDELINES

- The definition of register units, variables for register units, and the concept of maintenance, updating, and management of the register must be documented and aligned with EU Regulation 177/2008.
- When updating the register, available administrative data sources must be used.

- Each unit in the register must have its own identification number for internal use, which, once assigned, must not be changed.
- Data in the register are individual-level data. Adherence to principles of data protection and confidentiality must be guaranteed.
- Historical states of the register must be tracked to monitor individual units.
- The register serves as a frame for selecting observation units for statistical surveys. To reduce the burden on potential reporting units, records must be kept of which register units were used in which surveys.
- Feedback from each organizational unit using register data is essential for continuous improvement and maintenance.

2.2.3 Preparation for Using Administrative Sources

Using administrative data sources has numerous advantages for official statistics. Administrative data already exist and do not need to be collected conventionally, reducing costs and labor compared to classical surveys. This significantly eases the burden on reporting units. Data provision, if available, is generally free and mostly conducted electronically, making data immediately available for further processing.

Applications of administrative data include:

- Direct use as the primary data source;
- Use as supplementary data (complementing specific variables);
- Use as a sampling frame;
- Possibility of comparison with survey results.

Potential weaknesses of individual sources must also be identified. A major difference compared to classical surveys is that statistical institutions do not control the production of these data.

Possible issues include:

- Different concepts, definitions, and classifications;
- Coverage problems;
- Poorly maintained variables (irrelevant for administrative purposes);
- Data not available in a timely manner.

GUIDELINES

- Potential administrative data owners should be involved during planning and data collection analysis.
- Before using administrative sources, verify whether concepts and definitions correspond to the statistical purpose and sufficiently meet required standards.

- All process steps for making administrative data useful must be planned and tested in advance, and all measures documented.
- Continuous contact must be maintained with data owners, establishing protocols for regular and timely data delivery.
- Feedback on data quality from administrative owners should be provided regularly.
- Administrative data usually relate to individual records for persons or entities; maximum attention must be paid to data protection.
- Use of administrative sources must be transparently communicated to users, including implications for coverage, timeliness, and accuracy.

2.2.4 Preparation of Methodology for Selecting Observation Units

After determining the target population, a list of units with as many characteristics as possible must be prepared, called the sample frame. Variables describing unit characteristics are auxiliary variables. The selection methodology considers the data collection mode. Required data can come from administrative sources, existing statistical sources, or direct/indirect contact with observation units.

Surveys can be conducted as sample-based or full-coverage. Full-coverage surveys are lengthy and expensive but necessary to define populations without gaps over time, enabling sample selection and extrapolation frames. Most surveys are sample-based due to cost and burden considerations.

Sample Design Issues:

- **Sampling selectivity:** occurs when certain units are more likely to be included, potentially biasing results.
- **Random error:** Even without selectivity, sample results are subject to inherent variability.

GUIDELINES

- Evaluate available sources for the sample frame and how to link them.
- Include additional questions in questionnaires if frame quality is insufficient.
- For businesses, large units are included fully (census); smaller units are randomly sampled.
- Prefer simpler sample designs; consider multi-stage stratified sampling carefully.
- Consider response rates and nonresponse analysis when determining sample size.

2.3 Preparation of Data Sources for Sample Frame Construction

Understanding the difference between the target population and the sample frame is crucial. The target population includes all units whose characteristics are observed; the sample frame is the actual list of available units used for selection. Define target population characteristics, location, and reference period.

Example:
Target population: persons aged 15+ living in private households in BiH on a specific date. Define membership conditions to determine target population.

Surveys distinguish between sampling unit, observation unit, and reporting unit.

If the data source covers the reference period, population coverage is more accurate. One main source is selected, supplemented by others as needed. For business statistics, the main source is the Statistical Business Register; for surveys on individuals/households, it is the Population and Housing Census. Unique unit identifiers are essential. Coverage errors occur if sample frame and target population differ.

GUIDELINES

- Reference periods should align with the survey period. Document differences.
- Ensure consistency of variables across sources; document discrepancies.
- Sample frame should include unique identifiers, contact info, and classification variables.
- Clean sources before combining to ensure data quality.
- Document all sources and their descriptions.
- Use all available sources to minimize coverage errors; document inconsistencies.

2.4 Preparation of Statistical Data Processing Methodology

Statistical data processing encompasses all procedures after data collection or acquisition to ensure final results accurately reflect population characteristics. Procedures vary by survey type.

2.4.1 Nonresponse Handling

Address partial or complete nonresponse during planning to minimize bias. Nonresponse can also occur with administrative data.

2.4.2 Data Editing

Editing identifies and corrects errors in data. Plan to reduce time and cost, using selective or automatic editing where possible. Editing measures the quality of collection and processing, not final data quality.

2.4.3 Aggregation and Tabulation

Aggregate microdata to produce final statistical outputs. Plan which statistics to calculate and rules for calculation. Consider tabulation detail, international requirements, user needs, and data protection rules.

Additional steps may include:

- Estimation of sampling error for sample surveys;
- Deflation of financial data;
- Seasonal adjustment for time series data.

GUIDELINES

- Determine existing tools or need for custom solutions.
- Choose the most suitable software, considering functionality and standard practice.
- For nonresponse, decide on weighting or imputation, considering variables, periodicity, and external data availability.
- Use selective and automatic editing to reduce costs and burden.
- Avoid excessive detail in data publication to prevent empty or protected cells.

3 ESTABLISHING THE RESEARCH FRAMEWORK

This phase involves establishing and testing the production solution up to the point where it is ready for use in a "live" environment. The outcome of the "Preparation of Statistical Survey" phase guides the selection of processes, instruments, information, and services configured in this phase to create a fully operational environment for process execution.

3.1 Establishing Data Collection Channels and Instruments

Data collection may use one or more channels to receive data (e.g., face-to-face or telephone interviews, paper, electronic, or web questionnaires, etc.). Exchange channels may also include data extraction procedures for using data collected from existing statistical and administrative datasets. This subprocess also includes the preparation and testing of the content and functionality of the exchange channel (e.g., testing questionnaire questions). Technical specifications for creating or updating data collection instruments are prepared in this subprocess.

These procedures are carried out based on consultations with the IT department and developers. Data collection instruments are created or updated collaboratively between subject-matter statisticians and the IT department. Testing of the developed or updated statistical data collection software is performed, usually by IT staff and methodologists from the relevant department. In some cases, data entry staff or selected respondents may be involved in testing the data collection software.

3.1.1 Designing and Testing Questionnaires

Questionnaires/forms play a central role in the statistical survey and data collection process. Official statistics questionnaires usually contain only closed-ended questions, questions with predefined response options, and factual questions, while open-ended and opinion/attitude questions are included only in exceptional cases.

The design of factual questions is particularly important, as the formulation of a question with possible response options directly affects survey results. The quality of survey results largely depends on the quality of these steps undertaken during questionnaire preparation.

Questionnaires are usually divided into several sections (thematic units), guiding the reporting unit through the questionnaire via appropriate “filters.”

When designing a questionnaire, it is essential to consider whether the reporting unit fills out the questionnaire themselves or whether an interviewer records the answers provided by the reporting unit.

For questions with multiple response options (provided as assistance), it may be possible during face-to-face interviews to allow the reporting unit to read the available options. This assistance is not possible for telephone interviews.

Increasingly, electronic data reporting is gaining importance. Therefore, more work will be done in the near future on developing electronic questionnaires as a critical data collection method. Electronic questionnaires have the advantage that the software supports guiding respondents through filter questions, eliminating potential errors.

Once a questionnaire is released to the field, modifications are generally only possible in exceptional cases. Therefore, field testing and evaluation of the questionnaire are of great importance.

The choice of testing methods depends on various factors, such as the subject of the survey, available resources, or the type and significance of the survey.

For detailed questionnaires, introducing control questions can be useful to verify the credibility of the collected information.

3.1.1.1 Drafting the Questionnaire

Once variables and questions are prepared, a draft questionnaire is prepared with logical checks and skips. For electronic questionnaires, technical possibilities and standards are discussed with the programmer. Questions in the questionnaire are grouped into meaningful sets.

3.1.1.2 Internal Review of the Questionnaire

It is crucial to review the questionnaire before testing it on the target population. This internal review includes grammatical correctness, improvement of overly broad or unclear questions. Internal review means the questionnaire is checked by individuals not directly involved in the statistical survey. Reviewers may include: subject-matter experts, questionnaire design experts, interviewers, or representatives of the target population.

3.1.1.3 Technical Testing of the Questionnaire

When collecting data via electronic questionnaires (with an interviewer or web-based), technical testing is performed, including checking skips and logical operations.

3.1.1.4 Testing and Revising the Questionnaire

Before conducting a pilot survey in the field, the questionnaire should ideally undergo testing. This includes informal testing, cognitive methods, focus groups, interviewer reports, coding of interviewer or respondent behavior, and pilot testing.

Questionnaire design is an iterative process. After each major change, the questionnaire should be re-tested.

GUIDELINES

- Questionnaire questions and instructions should be clear and understandable for reporting units. Where possible, use questions proven effective in previous surveys.
- The first page of a printed questionnaire should include the survey title, statistical institution logo, and legal basis for conducting the survey. Introductory questions should generally apply to all reporting units.
- The layout of printed questionnaires should clearly separate different question blocks.
- Questionnaires should include instructions to facilitate completion, explaining how to enter allowed values for quantitative questions.
- For comprehensive questionnaires, control questions should be implemented to verify the credibility of collected data.
- Keep the number of questions to a minimum; each question should have a clear purpose.
- If different characteristics are collected from different reporting units, tailor questionnaires should be provided for each group.

3.1.2 Preparing Communication Materials for Reporting Units

In addition to the questionnaire, supplementary materials should be prepared before conducting the statistical survey, including: notification letter, brochure, reminder, thank-you letter, control letter, and follow-up letter.

The notification letter informs reporting units about the survey, its main objective, and data confidentiality.

The brochure encourages participation by presenting survey results in an appealing way.

Reminders are usually sent twice, exceptionally three times.

Thank-you letters acknowledge participation, sent when a survey is discontinued or participation is no longer required.

Control letters are intended to monitor interviewer performance and include questions related to the survey and interview process.

Follow-up letters are sent when initial contact is unsuccessful or the household did not participate for various reasons.

All materials must be proofread before distribution.

GUIDELINES

- For household surveys with interviewers, the reporting unit should be notified in advance about the interviewer's visit and survey purpose.
- Non-responding units should receive reminders requesting participation.
- For business surveys, typically two reminders are sent.
- In mixed-mode surveys (e.g., telephone and face-to-face), non-contacted units by phone should be approached on-site to reduce nonresponse bias.
- Control letters are used to monitor interviewer performance and may include key survey questions, along with interviewer and respondent identification.

3.2 Establishing Software Support

Software support and systems used in statistical business processes must be configured, from data collection to dissemination. This includes preparation of technical specifications for establishing or updating components for data processing and analysis.

Components may include control tables, databases, result tables, data transformation tools, data and metadata management tools. Establishing software support is done through consultations between IT, methodologists, and sampling staff (if needed). Software for data processing and analysis is created or updated collaboratively among methodologists, IT (or external programmers), and statisticians/sampling staff as needed.

3.3 Establishing Dissemination Components

Activities for establishing new or reusing existing components and services required for disseminating statistical products (as designed in subprocess 2.1) are described. This includes components for traditional print publications, web services, open output data, geospatial statistics, maps, or microdata access.

Technical specifications for building or updating dissemination components must be prepared, including product lists, types and functionalities of dissemination tools, visualization standards, and metadata quality links. This subprocess is carried out through consultations among methodologists, dissemination staff, and IT. Software for dissemination is created or updated.

3.4 Testing Data Collection and Processing Tools

This step involves technical testing and approval of new programs and procedures, including interaction testing between components, ensuring the production system functions as a cohesive unit. Initial tests are conducted for built or updated data processing and analysis components. This also includes collecting data in pilot surveys to test data collection instruments. After the pilot, data are processed and analyzed, and adjustments may be made as needed.

3.5 Testing and Configuring the Statistical Business Process

Configuring the production process flow covers data collection to archiving final statistical results. Activities include:

- Producing documentation of process components, including technical and user manuals;
- Introducing and training users on the production system, familiarizing them with its structure, procedures, and instructions for operating components. Training includes technical sessions and user manuals, often supported by IT;
- Moving process components into the live environment and ensuring they function as expected.

4 DATA COLLECTION

In this phase, all necessary data (data and metadata) are collected using various data collection methods (including extractions from statistical, administrative, and other non-statistical registers and databases) and loaded into the appropriate environment for further processing. While this phase may include validation of data formats, it does **not** include actual data transformations, as these occur in phase 5, "Data Processing."

4.1 Creating a Sampling Frame for Statistical Surveys and Sample Selection

4.1.1 Preparing the Sampling Frame

To prepare a high-quality sampling frame, multiple data sources are typically used. Once these sources are identified and prepared in a suitable electronic form, they must be linked using the appropriate methodology, and variables defining key characteristics of units included in the sampling frame must be prepared.

The ultimate goal is to create a unified dataset covering a list of units that closely represent the theoretically defined target population, with each unit assigned values for the variables required in the later stage of observation unit selection.

A key step in preparing a sampling frame is determining the procedure for selecting units to include in the frame. The list of units identified at the chosen point in time should be “enhanced” using supplementary administrative and statistical data sources.

The primary goal is to identify and eliminate units in the list that exist in the register but do not actually belong to the target population. These irrelevant units may exist due to errors in register updating procedures or the administrative nature of the register. In some cases, we also attempt to identify missing units—parts of the population not covered in the register—using additional data sources. This undercoverage usually stems from the predefined population of the used register not fully meeting the specific needs of the survey.

GUIDELINES

- If multiple surveys relate to the same target population and reference period, the same procedures should be used for determining the sampling frame to enhance consistency of statistical results.
- In both the data source identification and processing phases, inappropriate or duplicate units must be eliminated to improve statistical quality.
- If variables for the sampling frame are determined from multiple data sources, careful consideration of source prioritization must be made.
- The quality of sampling frame preparation procedures and the quality of data sources (based on feedback collected in the statistical process) should be regularly and systematically assessed. Significant deviations from accepted standards should trigger corrective measures.

4.1.2 Selection of Observation Units – Sample

The sample frame represents the physical realization of the units of the target population, whose characteristics would describe the statistical result.

In the next step, it is necessary to select the units (from among the units in the sample frame) that will actually be included in the survey and from which we want to collect data in later stages. For practical (technical) reasons, it is rare to include all units from the sample frame in the survey.

Usually, it is necessary (through an appropriate procedure) to reduce the list of units from the sample frame to a manageable size, which still allows sufficiently accurate estimates for the entire population.

Once the sample frame is prepared and the sample size is determined, the procedure for selecting sample units begins, which was established during the survey planning phase. This requires appropriate software environments and special algorithms.

In practice, several sampling methods are used. These include:

- **Simple Random Sample (SRS):**

In a simple random sample, the selection of units for the survey is determined in advance in cases of partial surveys. SRS is a statistical procedure in which units are chosen from the population randomly. The purpose of random sampling is to avoid sample selection bias as much as possible. A prerequisite is a complete list of all population units, which serves as the basis for selection
(Example: drawing a ball from a drum, as in a weekly lottery draw.)

- **Two-Stage Random Sampling – Cluster:**

In two-stage random sampling (cluster sampling), the population is divided into clusters, each representing a reduced version of the population (e.g., individual schools in a student survey). Elements within a cluster should reflect the range of characteristics studied, meaning that elements within a cluster should be as heterogeneous as possible. Clusters themselves should be as similar as possible to each other. Individual clusters are randomly selected, and then a simple random sample is taken within the chosen clusters, or all elements of the cluster are included (e.g., all students of selected schools are surveyed).

- **Quota Sample:**

In a quota sample, the total population is first divided into quota cells, which may, for example, be defined by age and sex for surveys of individuals. Each quota cell must include a certain number of sample units in the survey. This represents a "non-random" stratified sample, commonly used in ad hoc market research, public opinion surveys, etc. The selection of specific units to "fill" the quota cell is not strictly defined and may depend on survey costs or availability. This method is often used in income and expenditure surveys and current economic calculations.

- **Stratified Sample:**

The most commonly used sampling design, in which the population is divided into homogeneous, mutually exclusive groups called strata. In such samples, the sample allocation must first be calculated for each stratum, meaning the number of units to be selected from each stratum. A simple random sample is then drawn separately from each group. Strata are defined based on stratification characteristics (e.g., age and sex). *(Example: selecting 100 residents randomly from each region in a country.)* Units within a stratum should be as similar as possible regarding the studied characteristics, while units from different strata should be as different as possible.

Common allocation methods include:

- **Equal allocation:** select the same number of units from each stratum;
- **Proportional allocation:** select units in proportion to the stratum's size in the population;
- **Optimal allocation:** determine sample size per stratum considering both stratum size and variability of auxiliary variables, sometimes using prior survey data.

For all allocations, ensure a sufficient number of units is selected. If the sample size in a stratum is too small, a minimum fixed number (usually about ten units) is assigned, or the entire stratum is included if its size is smaller than the fixed number.

- **Purposive (Deterministic) Sampling:**

Purposive selection or selection based on deterministic rules is performed using specific algorithms. Deterministic methods differ from random selection in that fixed rules are used, without any randomness. Common deterministic methods include **census** and **threshold coverage**.

  - **Census:** All units from the sample frame are included, applicable in rare cases.
  - **Threshold coverage:** The sample frame list is narrowed based on predefined criteria. Units meeting the criteria (e.g., value above a threshold) are included, while others are excluded.

Two basic approaches exist for threshold coverage:

1. **Fixed threshold:** Units are selected if their value exceeds a fixed criterion (e.g., companies with more than 20 employees).
2. **Variable threshold:** Population is divided into subgroups, and the largest units in each subgroup are selected to exceed a predefined proportion (e.g., top 75% by size).

In practice, a combination of both approaches is often used.

The result of the sampling process is the selected sample units, with **weights** assigned later (initial weights during selection, final weights after data collection, usually adjusted for non-response).

Sample weights are values used along with collected data to calculate estimates of population parameters.

Key units are those particularly important for the study due to their expected impact on the final results. These units are treated differently during data collection and editing.

GUIDELINES:

- Efficient stratification occurs when units within a stratum are similar (homogeneous) regarding a key variable, while units across strata are as different as possible.
- For highly skewed distributions, some strata may need to be fully included in the sample ("take-all strata"), e.g., large companies in business surveys.
- Threshold coverage is used only for highly heterogeneous populations, where units below the threshold have negligible influence on results. A prior study should justify this method.
- Even when using threshold coverage, monitor units below the threshold.
- When calculating allocation, account for expected non-response. Subgroups with higher expected non-response should receive more units. Thresholds may be adjusted accordingly.
- In periodic surveys, continuously check if the accuracy of random sample results meets standards. If not, redesign the sample or consider increasing the sample size.
- Monitor whether omitted units introduce bias. If bias exceeds acceptable criteria, adjust threshold determination procedures.
- Avoid excessive numbers of key reporting units to achieve desired effects efficiently.

4.2 Organization of Data Collection

Data form the foundation for the production of statistical outputs. Data processing starts from raw data and ends with the creation of an authentic dataset. The method of obtaining data, to acquire raw data, must be defined already in the planning process.

There are several ways to organize the data collection process:
a) using already available data,
b) using administrative data, and
c) collecting data through primary statistical surveys.

The strategic goal of statistics in BiH in the coming period is to use existing administrative data sources as much as possible, and to employ traditional data collection methods (e.g., paper forms) only when data cannot be obtained otherwise. The use of administrative data alone is insufficient, primarily because official data collections in BiH are very limited or not maintained adequately.

Primary statistical data collection can be conducted through the following methods: in-person interviews by enumerators, mailing printed questionnaires, and newer or planned methods such as computer-assisted personal interviews (CAPI) and computer-assisted telephone interviews (CATI).

The table below presents the advantages and disadvantages of these data collection methods:

Data Collection Method	**Advantages (+)**	**Disadvantages (-)**
Available data	• No additional costs during collection • No burden on reporting units • Knowledge about the data is available	• Data may have been collected for other statistical purposes
Administrative sources	• No burden on reporting units • Generally achieves higher coverage	• Data often not directly statistically useful • Dependence on data owners
Personal interview (“face-to-face”)	• Personal attention to reporting units • Enumerator can correct errors	• Possible errors by enumerator • Travel time for enumerators
Printed questionnaires	• Internal proof of instrument dispatch	• High burden on reporting units • Large survey instruments needed • Intensive fieldwork • Addresses must be constantly updated

GUIDELINES

- The data collection method should be defined in the planning phase, balancing cost/benefit and reporting unit burden.
- For personal interviews, enumerators must be properly trained. A pilot survey should be conducted to simulate the survey process. Computer-assisted interviews are optimal.
- Timing of contact with reporting units should be carefully planned relative to reporting periods.
- During primary data collection, return rates and incoming material must be continuously monitored. Appropriate measures should be taken in case of low response rates.
- Reporting units should be informed about the survey purpose, potential legal obligations, and assigned a contact person for questions.
- The entire data collection process must be documented to allow future improvements.
- Individual data collected in primary or secondary surveys must be protected for confidentiality.

- Reporting units should eventually be allowed to submit data electronically, with adequate helpdesk support and explanations of benefits.
- When using secondary statistical sources, the data flow from owners to statistical institutions must be clearly defined and documented, preferably via electronic means, with accompanying metadata.

4.2.1 Organizing Data Acquisition from Administrative Sources

When collecting data through statistical surveys, efforts should be made to minimize non-response.

In studies using secondary sources (administrative registers, records), non-response is usually lower, but methodological differences from statistical concepts may cause other difficulties.

Secondary or administrative sources should be recorded centrally, allowing existing sources to be checked for additional needs. Records should be regularly updated, including the date of last data acquisition.

Key questions for data acquisition from administrative sources include:

- Which data are collected and from whom;
- How the data are obtained;
- Data structure;
- Data protection measures;
- Timing of data collection.

Assessing the current process of administrative data acquisition requires:

- Listing all administrative sources, including data type, owner, provider, and administrator (person responsible for administrative sources in the statistical institution);
- Determining how data are received;
- Documenting technical structure;
- Reviewing written documentation on procedures after receiving data;
- Recording data storage methods;
- Defining how data are further used in statistical processes.

A unified terminology is recommended:

- **Agreement:** Arrangement between administrative source and statistical institution on technical, organizational, and legal data protection. Prepared by the coordinator administrator with legal services.
- **Technical protocol:** Defines technical and timing aspects of data transfer and preparation. Prepared by the technical administrator.
- **Intranet portal for administrative sources:** Provides information on agreements and technical protocols.
- **Database:** A structured set of data received from administrative sources (e.g., VAT database).
- **Dataset:** A database consists of multiple datasets; each dataset corresponds to one table for retrieval (e.g., single register of indirect tax payers – form ZR1).
- **Time period:** Specifies the reference period of data.
- **Coordinator administrator:** Responsible for administrative sources.
- **Technical administrator:** Responsible for technical support of administrative sources.
- **Authorized data receiver:** Person authorized by the director to retrieve data.
- **Administrative source:** Data provider.

4.2.2 Concluding an Agreement with the Administrative Source

The statistical institution signs an agreement with the administrative source for data acquisition. Agreements should be standardized and define all parameters necessary for successful and transparent data transfer.

The technical part of the agreement is the **Technical Protocol**, specifying:

- Data content – precisely describing which data will be acquired;
- Appointed administrators – coordinator and technical administrator on both sides;
- Frequency of data transfer;
- Technical specifications, including:
    - Parameters for retrieving data;
    - Parameters for decrypting and processing data before loading into the database;
    - Parameters for successfully reading and writing data into the database.

Methods of data transfer:

- Portable media (CD, DVD, USB, etc.)
- FTP server
- Email
- Web services
- Direct database link (replication or selective access)

Technical specifications must include:

- Identification of the data provider;
- Unique identification of the database;
- Transfer timing;
- Data version;
- File names;
- Control data (record counts, number of defined fields).

Portable media transfer: Define media type, size, notification of new data, delivery confirmation document, authorized persons, identification of data units, and file names. Encryption is used if transfer is unprotected.

FTP transfer: Agreement on server IP/FQDN, authentication, notification of new data, directory structure, and file names. Decide whether the source or the statistical institution hosts the FTP. Use encryption if needed.

Email transfer: Define sender address, receiver address, email subject content, file names, maximum size (e.g., 10MB), and encryption method.

Web services: Internet-based applications using XML standards. Provider implements the service, recipient implements the client. Agreement on purpose, address, XML schema, response schema, and authentication method is required.

Direct database link / replication: Two options: full replication or selective data retrieval. Agreements should define data content, tables, fields, transfer timing, connection technology, and access protection policies.

Technical parameters for data processing:

- Decryption of received files
- File validation: all files received, correct record counts, other control data
- Translation of identifiers into statistical IDs

Technical parameters for database loading:

- Define file format and field definitions to allow correct reading and loading into the database.

File formats:

- **Text file:** tabular format, fields by position, standard character set (Windows 1250), end-of-line <CR><LF>, end-of-file <CR><EOF>, field types: text (left-aligned), numeric (no thousand separators, decimal comma, right-aligned), date (ddmmyyyy), time (hhmmss). Advantage: simplicity; disadvantage: no consistency checks.
- **CSV file:** same as text file, with field separator (standard: semicolon ;). Text fields must be quoted; embedded quotes doubled.
- **Microsoft Excel:** multiple datasets per file, sheets correspond to datasets, first row contains field names. Advantage: more consistent; limitation: older Excel versions allow only 65,536 rows per sheet.
- **Microsoft Access / XML:** follow dataset structure and defined data types.

GUIDELINES

- Agreement and technical protocol must be prepared and signed before data acquisition.
- Appoint coordinator and technical administrators.
- Take measures to reduce errors during data transfer.
- Update data acquisition records regularly.
- Maintain records of physical media with unique IDs.
- Store physical media in fireproof cabinets.
- Coordinator administrator is responsible for notifying users about changes, delays, or incidents.

4.3 Launching Data Collection

Data collection from reporting units (data providers) is carried out using various collection instruments. The launch of data collection includes initial contact with reporting units and all subsequent activities for monitoring timeliness and receipt of reports according to the planned schedule.

In this subprocess, a preliminary database is prepared. The responsible department and statisticians are provided access to internal IT tools to review the collected preliminary data. Validity checks of the collected data are also performed.

Basic validations of the structure and integrity of received information can occur within this subprocess, e.g., checking that files are in the requested format and contain the expected fields. In most cases, these automated validation procedures are performed by the IT department, and the results are delivered to the relevant departments for evaluation and further actions.

If technical problems are identified and additional measures are taken, reporting units are asked to resubmit their reports. This subprocess may also involve manual data entry on-site or management of fieldwork, depending on the source and collection method. The timing and method of contacting reporting units and whether they responded to queries are recorded. Service providers are managed to ensure that the relationship between the statistical institution and the data provider remains positive.

4.3.1 Communication with Reporting Units

Data collection and communication with reporting units depend on the method of questionnaire completion, the target population, the observed phenomena, and available sources.

Special attention should be paid to information loss due to human or system errors by enumerators or data processing errors. High response rates can be achieved with clear objectives, appropriate tools, and a suitably chosen data collection method.

Surveys conducted during unfavorable periods, e.g., during holidays or periods of heavy workload (such as final account preparation), often have lower response rates.

The response rate also depends on the survey topic: for example, health surveys generally have higher response rates, while surveys on crime victims typically have lower response rates.

Non-response can be categorized into two main types: **item non-response** and **unit non-response**. Item non-response occurs when some questions in a completed questionnaire remain unanswered. Reasons include lack of understanding, refusal to answer sensitive questions, enumerator errors, etc. Unit non-response occurs when data are missing for all variables.

Different completion methods have different non-response rates. Self-completion generally has the highest non-response rate. Surveys involving households or individuals usually use interviews, allowing enumerators to motivate respondents and reduce non-response.

The duration of data collection depends on the questionnaire method. Telephone interviews are usually faster than face-to-face or mail surveys. For mail surveys, collection duration is harder to estimate because it depends on reminders sent to non-responding units and their response behavior. Using multiple sources and more time generally increases the response rate.

It is generally believed that in one survey, only two of the following three goals can be fully achieved: lower cost, high response rate, and short collection time. Monitoring reasons for non-response is more frequent and detailed in interviews than in self-completion.

In self-completion, the reporting unit reads and answers the questions according to their understanding, often resulting in more errors that require additional editing.

Completed questionnaires or comments from reporting units can be received via mail, fax, email, e-reporting, or telephone. Response records are regularly monitored. Reporting units that do not respond receive a maximum of two written reminders. If necessary, key units may also be contacted by phone. Staff members responding to inquiries provide additional questionnaires or answer technical or methodological questions promptly.

In interviews, the enumerator completes the questionnaire on behalf of the reporting unit, ensuring that all questions are answered satisfactorily. The number of enumerators should be adequate for the survey. Non-response is monitored by reviewing submitted data; enumerators not meeting quantity or quality standards are warned, questionnaires checked, and repeated issues may result in termination and replacement.

For telephone surveys, enumerators are supervised by a controller who monitors quality, alerts enumerators to errors or low quality, and ensures adequate breaks. Enumerator training, supervision, and distribution of additional instructions are easier than in other methods. It is desirable that the survey manager is present in every new survey to address methodological questions from enumerators.

Follow-up (Urgency) is the procedure to minimize unit non-response. It is a key measure for ensuring timeliness and accuracy. Formal follow-ups are used only when a reporting unit fails to respond by a specified deadline. For voluntary surveys, formal follow-ups should be avoided whenever possible.

In general, success in data collection depends on whether data submission is legally required or voluntary. Voluntary surveys tend to have lower response rates. Telephone encouragement may be necessary for voluntary surveys to achieve an adequate response rate and ensure data quality.

Formal follow-up usually involves several stages:

- First reminder in the form of a written notice, followed by a second reminder referencing legal consequences of non-response.

The timing of follow-up procedures is set during planning, depending on survey frequency, processing workload, and dissemination schedule.

GUIDELINES

- Follow-up procedures (urgency) are crucial for linking timeliness and accuracy principles.

- Prior to formal follow-up, contact reporting units to request data submission.
- The type and frequency of follow-ups depend on desired response rates and project deadlines for data dissemination.
- Fieldwork must be adapted to the target population. Time schedules, notifications, brochures, instructions, methods, and collection tools should be carefully planned to reduce respondent burden.
- Communication with reporting units and enumerators must follow internal rules, resolving problems quickly and patiently.
- IT systems (phone switchboard, servers) must operate continuously during data collection.
- Procedures to reduce refusals should be applied, identifying population segments with higher non-response rates and intensifying contacts with them.
- Establish a minimal set of data that must be provided to be considered a valid response.
- Non-response can also be reduced using symbolic incentives, applied rationally.
- Systematic monitoring of the process is essential: progress of collection, non-response, and balance between quality and cost. Key quality indicators include response rate, processing errors, and reasons for non-response.
- Determine the appropriate number of contacts with each reporting unit; usually a maximum of two written reminders is sent.
- Responses from key reporting units must be obtained. If not, request in writing the reason for non-submission, recording either response date or status.
- At the start of a survey, the survey manager must participate in enumerator supervision.

4.4 Entry of Collected Data

Statistical institutions collect data in various formats and prepare them for further processing.

Data from printed questionnaires must first be converted into electronic format for subsequent processing.

Conversion into electronic format can be done in two ways:

- **Manual entry** (the result of this process is data in electronic form), or
- **Using an optical reader** (the result of this process is both images of the questionnaires and data in electronic form).

In special cases, the verification of questionnaires may be carried out before data entry. The extent of verification largely depends on the method used for data entry: scanning or manual entry.

If optical reading is used, data are entered into the corresponding fields or coded into designated fields.

Data in the questionnaire must not be crossed out or written outside the designated reading fields, as this may lead to poor quality of scanned data; these restrictions do not apply to manual data entry.

The necessary programs and procedures for manual entry and optical reading are prepared during the data collection preparation phase. The procedures and programs must also be tested, and necessary corrections made before use.

When data are collected using combined reporting methods (e.g., printed questionnaires, electronic questionnaires, CATI, CAPI) or data entry is done using combined approaches (e.g., optical reader and manual entry), or part of the data is taken from administrative sources, the collected data must be merged into a single database for further statistical processing. This allows for additional verification and editing of the data in the database, facilitating the assessment of result reliability.

GUIDELINES

- The survey manager must prepare clear instructions for data verification before entry.
- Data entry programs must be prepared on time and tested before use to avoid errors in the source data.
- Since the number of errors in digital data collection is lower, electronic questionnaires should be preferred when possible. For printed questionnaires, optical reading is more appropriate (due to lower error likelihood).
- If using optical reading or fast entry without repetition, the questionnaire should be designed to allow internal controls (e.g., a separate field for amounts).
- Data on questionnaires read by optical readers must be recorded clearly and legibly.

5 DATA PROCESSING

In this phase, we carry out the “cleaning of collected data” and prepare it for analysis. It consists of subprocesses that check, clean, and transform input data so that they can be analyzed and disseminated as statistical results. Subprocesses may, if necessary, be repeated multiple times. For statistical products that are produced regularly, this phase occurs in every iteration. Subprocesses in this phase can be applied to data from statistical and non-statistical sources (with the possible exception of subprocess 5.6 “Weighting,” which is usually specific to survey data). The "Data Processing" and "Data Analysis" phases can be iterative and comparative. Analysis may reveal a broader understanding of the data, which could clearly indicate that additional processing is required. Activities within the "Processing" and "Analysis" phases can start before the "Data Collection" phase is completed. This allows the production of preliminary results, where timeliness is a primary concern for users, and increases the time available for analysis.

5.1 Linking Different Data Sources

The statistical production process has undergone transformation in recent years. In many cases, the use of a single data source provides insufficient information to meet the analytical purpose of a project. In statistical surveys, it is justified to introduce (for reasons of reducing the reporting burden on units) characteristics from available administrative sources. Efficient and primary use of existing administrative data for statistical production is a priority task for the Agency for Statistics of BiH. For many projects, linking different data sources is an important step in the data processing process.

General conditions for the linking procedure may vary. In **direct linking**, two datasets with the same identification characteristics (“key variables”) are connected, ensuring that the linked units correspond meaningfully. As illustrated in Figure 5, such an ideal situation is not always present in practice.

Figure 5: Combining Two Data Sources

Indirect linking is more complex, i.e., linking data when the reference and secondary datasets do not share the same identifier (e.g., personal ID, company ID). Before combining datasets, a connection between different key variables must be established.

In this case, other selected variables that exist in both sources are used. Units in the reference source are divided into two parts:

1. Units that are unambiguously linked to the secondary source via intermediate connectors. For these units, the intermediate connector values match completely in both sources, and there is only one corresponding record in the secondary source.
2. Units that could not be unambiguously linked to the secondary source, meaning that each reference unit corresponds to multiple records in the secondary source. The associated records are those whose intermediate connector values are the most similar in both sources.

In some cases, there are no key variables to perform the linking. In that case, it is possible to merge data through text comparison of individual identification characteristics (e.g., addresses), using methods similar to automatic coding. If this is still not possible, one may attempt to assign the most similar duplicate to each record.

GUIDELINES

- Before linking data from different sources, accurate information about the concepts and definitions used by the data providers must be ensured and documented. This relates to quantities and units, as well as the characteristics included.
- When linking datasets at the individual level, relevant legal regulations concerning statistical confidentiality must be observed.
- If possible, merging of different data sources should be performed using a common key variable.
- When linking datasets with different key variables, connections between identification characteristics must be established. Memorizing and maintaining these connections between important data providers should be continuously carried out in the statistical register.
- Ideally, all data providers should be fully linked. If not all units can be matched via identification characteristics, text comparison should be attempted first, followed by **Statistical Matching** to achieve the most reliable connection. Any remaining unmatched units should be addressed through appropriate estimation procedures.
- All aspects of linking data sources must be described and documented, including the sources involved, the methods used, and quantitative information about the linking process.

5.2 Coding

To process the data further, they must be available in electronic form.

Coding (signing) refers to assigning an alphanumeric key to textual responses and is usually an automated process. For special cases, partial manual coding is also necessary. To enable automatic coding, a sufficiently developed coding key must exist. All records that cannot be automatically coded should, if possible, be coded by qualified personnel.

Coding using complex classifications requires specially qualified staff to achieve results of acceptable quality. Targeted training is needed to ensure correct identification and classification according to classification rules, and to achieve the required comparability.

GUIDELINES

Data Collection

- Determine the value ranges for each characteristic first. Also decide which modalities or values are used for cases where no numerical value exists (e.g., missing values, “no response”). These values/modalities must be clearly identifiable.
- In manual data collection, personnel must be careful when entering critical fields. Data entry outside permitted ranges is not allowed. Double-entry and subsequent comparison minimizes input errors.

Coding

- During survey planning, open-ended questions should be avoided whenever possible, so coding will only be necessary where absolutely required.
- Coding should be automated whenever possible.
- Parameters for automatic coding should be selected to maximize the number of records coded while minimizing coding errors.
- Records that cannot be automatically coded should be manually coded by trained staff.
- For complex classifications, staff must ensure correct identification and classification according to classification rules.

5.3 Data Acceptance/Validation – Logical Checks

To ensure statistical data can be processed adequately, the datasets must be error-free, logically consistent, and content-wise acceptable. All data delivered by reporting units must undergo acceptance checks.

We distinguish between **measurement errors** and **recognition errors**:

- **Measurement errors** occur when data is completely incorrect. Examples include:
    - Percentage distributions do not sum to 100%;
    - Data for a characteristic is completely missing (e.g., no income reported);
    - A business is classified as trade but engages fully in manufacturing.
- **Recognition errors** occur when data is suspected to be incorrect. Examples include:
    - One month’s revenue is more than double the highest value within the past 12 months, or less than half the lowest value;
    - “Revenue per employee” is outside the expected range.

GUIDELINES

- Logical checks must be consistently defined and obligatorily tested.
- In electronic data collection, major logical checks should be integrated into the collection application to prevent incorrect entry.
- Avoid the practice of applying excessive control rules that are too strict, as this may hinder proper data cleaning.
- When defining acceptance rules at the micro level, identify:
    - Values outside defined ranges,
    - Inconsistent or unacceptable value combinations,
    - Missing values.
- For quantitative variables, consider whether extreme values ("gross errors") are acceptable. Methods for detecting gross errors must comply with recognized and established standards.

5.4 Editing and Imputation

Data editing can be divided into **micro-editing** and **macro-editing**. Micro-editing deals with individual units.

- For **interview-based surveys** (field or telephone), individual records are checked for content consistency. After the survey, field or telephone data are combined with potential secondary sources into a common file; additional checks are applied, and detected errors are corrected manually or automatically. Editing rules are defined for each dataset. Outlier detection (extreme values in quantitative variables) is a key aspect of micro-editing. In computer-assisted surveys (e.g., CAPI), verification and corrections can be performed near the source.

- For **self-completion questionnaires** (printed or electronic), logical checks are performed after data entry. Full dataset checks are applied simultaneously with automatic corrections and error statistics generation. Detected errors are corrected manually or automatically.

- For surveys using **multiple data sources** (statistical and administrative), each source is cleaned individually, followed by checks on the combined dataset.

Selective editing is often used in business surveys, prioritizing important units for manual review and automatic processing for less important units.

Macro-editing verifies aggregates and distributions against previous periods or other sources. Implausible results must be further analyzed and corrected after clarification.

Imputation replaces missing, invalid, or inconsistent values to produce complete, reliable, and coherent data. Causes of missing values include refusal/non-response, technical transmission errors, interviewer mistakes, etc. Imputation addresses mainly **item non-response**; **unit non-response** is usually handled through weighting.

Imputation methods include:

- **Logical imputation** – infer value logically from available data for the unit (e.g., compute age from date of birth).
- **Mean imputation** – replace missing value with the average of similar units in the same domain.
- **Internal donor** – replace missing value with a value from another unit in the dataset.
- **External donor** – replace missing value from an external source (e.g., previous survey or administrative data).
- **Regression method** – replace missing value using a regression model.

Imputation quality should be monitored:

- **Imputed data rate for key variables** – number of units with imputed data divided by total units needing data.
- **Imputation rate for statistics** – compare statistics before and after imputation.

GUIDELINES

- Macro-editing should include reference comparisons, especially with previous periods.
- Acceptance checks should be automated where possible.
- Data editing processes must be fully documented and reproducible.
- Imputation of missing values is necessary to prevent bias and allow advanced statistical analysis.
- Mass imputation (>20% of dataset) is only applied when other options are infeasible.
- Analyze imputation results to verify effectiveness by comparing distributions before and after.

5.5 Production of Derived Variables and Units

New statistical units may be derived by aggregating or splitting data from collection units, or through estimation methods (e.g., deriving households when data collection units are persons). Derived variables and units are required to obtain aggregates, calculate indicators, indices, or other statistical results. Formulas are applied to existing variables to produce required outputs.

5.6 Weighting

Sample surveys cover only a part of the population. Sample size is significantly smaller than the full population. Sample-based surveys reduce workload and save financial and labor resources.

Two approaches exist: **probability samples** (random) and **non-probability samples**. Simple random sampling is rarely used in official statistics. Stratified random sampling is commonly used in BiH.

Random samples allow each unit in the target population to have a known, non-zero probability of selection. Results from random samples enable conclusions about the total population.

Cut-off (concentration) samples are often applied in economic statistics, selecting only units above a certain threshold, considering standardized representativeness criteria.

Weighting involves:

1. **Design weights** (initial weights), based on selection probability.
2. **Non-response weighting**, adjusting design weights for non-response.

Final weighting may also include **calibration to external sources**.

All sample-based survey estimates carry **sampling error**, which should be calculated at least for key variables.

GUIDELINES

- Sample survey results are estimates and carry sampling error.
- Estimates should consider design weights and non-response adjustments. Calibration to external sources is recommended.
- Sampling errors should be presented with appropriate measures (standard error, coefficient of variation, confidence intervals).
- The methodology for weighting and sampling error calculation must be documented.
- Users should be informed of acceptable sampling error limits, especially for regional breakdowns.
- The full sample workflow must be documented, including original gross sample size, achieved net sample, and response rates for key variables.
- All sample-related aspects (sampling plan, extrapolation methods, quantitative measures) should be presented in the Quality Report.

5.7 Calculation of Aggregates

This subprocess creates aggregate data for the total population from microdata or lower-level aggregates. It includes summing records with common characteristics, calculating mean and dispersion measures, and applying weights from subprocess 5.6. For sample-based surveys, sampling errors can be calculated for aggregates. Aggregates can be produced via a database tool or derived from initial aggregates.

All data collected for comparison purposes should be included (totals, means, medians, coefficients of variation, standard deviations). Variance estimates for validation, such as confidence intervals and sampling errors at the aggregate level, should be included for internal checks.

5.8 Creation of Final Data Files

This subprocess combines results from other subprocesses in this phase, producing a data file (usually microdata) used as input for the "Analysis" phase. Sometimes, these files may be interim rather than final, especially in business processes with strong time pressures and requirements for preliminary and final estimates. The data files should contain microdata, weights, and aggregates.

6 DATA ANALYSIS

Statistical data analysis is the process of analyzing data using various tools and techniques, with the aim of examining the situation and identifying certain patterns significant for the observed phenomenon, for the purpose of interpreting results. Typically, data is first analyzed at the macro level and, if necessary, at the micro level.

Data is analyzed to confirm its relevance, to detect possible deficiencies, and subsequently eliminate them, thus ensuring better data quality. If the analysis reveals systematic deficiencies, the results are used to improve the quality of the process or to supplement or revise the entire methodology.

6.1 Preparation of Draft Results

In this subprocess, data is transformed into statistical results. It includes the production of additional measures, such as indices, trends, or seasonally adjusted series, as well as the recording of quality characteristics. Quality indicators such as response rates, sampling errors, or other necessary indicators are also calculated in parallel with the preparation of draft results.

Quality indicators can be used for additional verification of results in subprocess 6.2 or as supplementary information in subprocess 8.2. This subprocess also involves the calculation of seasonally adjusted time series.

6.1.1 Production of Working and Analytical Tables

The purpose of producing statistics is the publication of results. Before a final decision is made on the type and scope of publication, working and analytical tables are produced to provide a basis for assessing the available data using detailed combinations of characteristics.

Working tables primarily serve internal control of results. The production of such tables is not intended for publication but is based on a well-defined and structured set of tables.

Working and analytical tables are generated from the authentic database. If analytical tables are also used to verify the database (before it becomes authentic), they also serve data cleaning purposes.

The goal of producing working and analytical tables is to answer questions relevant to publishing results, primarily including: confidentiality, accuracy of results, and appropriateness for storage in statistical databases.

Working and analytical tables are generally produced by aggregating individual data.

GUIDELINES

- Analytical tables should include as many combinations of characteristics as possible, with preference given to two-dimensional tables.
- Standard software should be used for table production, or the responsibility for producing tables should lie with IT. The production of working and analytical tables must always follow the authentic databases.
- Interpretation and control of working and analytical tables must be conducted diligently. Any issues should be discussed in a wider forum.
- Working and analytical tables are used for internal purposes. It is important to ensure access to this material is granted only to authorized persons. All tables (including electronic ones) must be archived.

6.2 Analysis of Relevance and Verification of Results

The analysis of result relevance is the process of checking the "meaningfulness" of results, their internal consistency, temporal and spatial comparability, and consistency with existing internal and external reference data sources.

Analysis of relevance, i.e., verification of results, is carried out during the macro-level result editing process and includes the following procedures:

- Checking internal consistency of results, e.g., verifying results based on known or expected relationships among results (e.g., whether production value is higher than added value).
- Checking consistency of results against results from previous reference periods (especially relevant for surveys whose primary purpose is not to measure changes over time).
- Checking consistency with related or connected results from other statistical surveys (results from surveys conducted by statistical institutions in BiH or results from surveys conducted by other institutions).
- Internal verification of results within statistical institutions.
- Occasional verification of "relevance and meaningfulness" of results with external experts.

GUIDELINES

- Before final confirmation of results, all process data (data available regarding the data processing process) should be analyzed once more, especially macro-level editing results.
- Verification procedures should be adapted to the survey periodicity, target population, data collection method, and type of data source (primary or secondary).
- The results of the analysis should be used to improve quality in the next survey implementation.

6.2.1 Time Series Analysis

A time series is a sequence of data ordered in time, in our case statistical data (e.g., industrial production, labor cost index, etc.). The main part of time series analysis is **seasonal adjustment**, while a smaller part is forecasting. During seasonal adjustment, the effects of seasonality and calendar are removed when they are specific and significant. The resulting values represent seasonally adjusted values or values with the effects of seasonality and calendar removed.

This procedure should always be performed when comparing data from different periods of the same time series or data for the same period of the same time series from different countries, as they typically vary throughout the year depending on season, number of working days, and other factors.

When comparing data for the same period across different years of the same time series, only calendar effects should be checked and removed, because the season for the same period in a year is approximately the same (e.g., comparing April 2023 to April 2022). These adjusted data are called **calendar-adjusted data** or **working-day adjusted data**.

Calendar effects include:

- Effect of the number of working days
- Effect of holidays
- Effect of leap years

However, if comparing data from the same time series but different periods, both calendar and seasonal effects must be adjusted. In this case, when seasonally adjusting the time series (Xt), we determine a model from which the seasonal component (St), irregular component (It), and trend-cycle component (TCt) are derived. In other words, the time series can be represented as:

- **Additive model:** Xt = TCt + St + It, where seasonally adjusted values are Xt - St or TCt + It
- **Multiplicative model:** Xt = TCt * St * It, where seasonally adjusted values are Xt / St or TCt * It

The trend-cycle component (TCt) describes long-term movement, the seasonal component (St) describes periodic movement (repeating each year similarly), and the irregular component (It) contains random effects, i.e., residuals of the time series after removing other components.

Models are reviewed annually (usually with the first data point of the calendar year), after which the survey owner sends all parameter model files and data to the time series methodologist for review (annual review).

Outliers are very important in time series analysis, as they represent unusual changes. If no data error exists, reasons for outliers must be investigated. There are typically three types of outliers:

- **Level shift:** break in the time series
- **Additive outlier:** sudden jump or drop, returning to usual level in the next data point
- **Temporary change:** sudden jump or drop, gradually returning to the usual level of the time series

If an outlier appears at the end of the time series, it is a **temporary outlier**, and the time series methodologist corrects it only after at least three subsequent data points are available, which allows proper classification of the outlier type. In this case, the survey owner must send the time series (every time a new data point is added) to the methodologist until the outlier is corrected.

The length of the time series for seasonal adjustment must be:

- At least **3 years** for monthly series
- At least **4 years** for quarterly series

Calendar effects with 1 or 2 regressors (1 regressor – working/non-working days (Mon-Fri/Sat-Sun); 2 regressors – working/non-working days and leap year) are considered for series with at least 5 years of data. Series with more regressors (6 or 7 regressors; 6 regressors – each weekday has its typicality; 7 regressors – each weekday and leap year) are considered only for series with at least 7 years of data.

GUIDELINES

- The survey owner must carefully review and analyze time series to improve model quality.
- The survey owner must provide the time series methodologist with all changes related to the series (sampling changes, methodology changes, etc.).
- The methodologist must consult with the survey owner regarding the significance of seasonal adjustment results (calendar effects, reasons for outliers).
- In published data, year-on-year comparisons should always use seasonally adjusted values (e.g., April 2023 vs. April 2022), adjusted for working-day effects. Trend-cycle components should be presented graphically only, noting extreme value issues.
- Very long time series (over 12 years) may sometimes be shortened (e.g., to 7 years) to improve model quality. If the season changes over such a long period, seasonal adjustment results are weaker. Data not included in the model (older than 7 years) remain available as original data.
- The survey owner must not forget the annual review and must send all models to the time series methodologist once per year.

6.3 Interpretation of Results

In the interpretation phase, data is transformed into information; statistical phenomena are explained to the user clearly and understandably.

Data must be relevant and useful, and all conclusions must be supported by the data obtained in the statistical process.

When interpreting data, the method of data collection (coverage and source: sample survey, administrative database) and other relevant information (e.g., response rate) must be taken into account. This information should be provided to the user together with appropriate metadata. Particular attention should be paid to all data limitations, such as discrepancies between the target population and the observed population. Data confidentiality principles must also be considered.

During interpretation, it is necessary to decide which phenomena will be presented to users. The focus should be on the most relevant phenomena, topicality of the subject, and output data. Data interpretation should be adapted depending on the user group (general public or experts).

- The general public is mainly interested in the most important or interesting general (popular) statistical data and information, presented clearly and understandably.
- Expert users mainly use detailed data for further analysis purposes.

GUIDELINES

- Interpretation of results must be adapted to the target population and the media (users) where the data will be published.
- Interpretation must be unbiased, objective, accurate, and understandable.
- Interpretation of short-term statistics should differ from structural statistics.
- Reference points strongly influence data interpretation. For time comparisons, use points that display the data most stably and impartially.
- For indices and other relative figures, choose a reasonable comparison period to allow users efficient interpretation of trends.
- When presenting results as indices or relative figures, careful interpretation of changes is needed when phenomena are expressed as percentages (%) or percentage points. For example, if the number of households with a PC increased from 21.3% in 2022 to 23.3% in 2023, the increase is **2 percentage points**, not 2%, which corresponds to 9.4% relative increase.

6.4 Statistical Data Confidentiality

Statistical confidentiality refers to the steps that need to be taken to mitigate the risk that a statistical unit can be directly or indirectly identified in a dataset.

Activities to protect confidential data involve two key steps:

- **De-identification of data**, i.e., removing any direct identifiers (e.g., name, address, identification/personal number) from the dataset, and
- **Assessment and management of the risk of indirect identification** that may occur in the de-identified dataset.

Removing identifying information (name, address, identification/personal number) protects the statistical unit from direct identification. However, it is still possible to indirectly identify a statistical unit in a de-identified dataset.

If a dataset contains many details, the identity of certain statistical units can be inferred from existing very rare characteristics or a combination of special characteristics.

Example 1: The identity of a person can be indirectly inferred if, in a dataset, a person aged 65 or older has an annual income exceeding 1 million KM and lives in a city with 10,000 inhabitants.
Example 2: The identity of a person with a very rare disease can be revealed even in highly aggregated data.

Statistical confidentiality includes removing or modifying information, as well as reducing (aggregating) details to ensure that no statistical unit is identifiable in the data (either directly or indirectly).

Various methods exist to protect the identity of statistical units while still allowing the publication of sufficiently detailed and useful information for statistical and research purposes.

6.4.1 Managing the Risk of Identifying Confidential Statistical Data

The greatest challenge in producing publicly available statistical data is ensuring that no statistical unit is identifiable in the data.

The process (steps) for managing the risk of identification is the same for all data published, whether **microdata** (where each record represents an observation for an individual or legal entity) or **macrodata** (where aggregated data is presented in tables).

The first step in managing the risk of identification is assessing potential identification risks. The second step is managing the identification risk using an appropriate method for data confidentiality protection.

Assessment of potential identification risks

The probability that a statistical unit can be identified depends on several factors:

- The number of details included in the data (more details → higher identification risk),

- The sensitivity of variables in the dataset (variables such as financial, health, or criminal records can increase the risk of identification), and
- The way information is presented (macrodata published in tables poses a lower risk of identification than publishing microdata).

6.4.2 How to Protect Confidential Data – Basic Principles

Identification risk in disseminated data (disclosure control) involves steps to assess and reduce the risk of revealing the identity of a statistical unit. The goal is to protect the identity of statistical units while maximizing the usability of data.

In simple cases, data can be protected manually. However, sometimes software tools are required, which demand specialized skills and knowledge to correctly protect confidential data.

Macrodata

Macrodata (usually tables) are standard products of statistical institutions and the most common way of presenting aggregated data. Tables can still contain individual information.

There are two main types of tables:

- **Frequency tables:** each cell contains the number of statistical units belonging to that cell (e.g., number of persons in age groups, or number of legal entities in an industry).
- **Value tables:** each cell contains the sum of a variable's values for the statistical units in that cell (e.g., total income or profit).

Information about a statistical unit can sometimes be inferred from these tables. For instance, in a frequency table, a cell with a very small number of cases may allow inference about a particular statistical unit using known and additional information in the table.

In value tables, cells at risk are those where dominant values (more than 85% of the cell total) correspond to one or two statistical units.

Identification risk also exists if users have access to multiple tables containing common elements. Data from one table can be used to identify a unit based on information in another table, highlighting the need to monitor all information across all published tables.

A specialized software most commonly used for assessing identification risk in macrodata records is **t-Argus**.

Microdata

Microdata represent individual-level data on statistical units and are a very valuable resource for researchers and policymakers. However, a major challenge for microdata managers is maintaining the right balance between protecting unit identities and maximizing information for statistical and research purposes.

Two key types of disclosure risk are associated with microdata:

- **Identification risk without intent** (spontaneous disclosure), and
- **Identification risk with intent** (deliberate/malicious disclosure).

Spontaneous disclosure occurs without intent, usually when statistical units have rare characteristics. For example, a dataset may include persons with unusual jobs (e.g., celebrities) or very high incomes, making them easily identifiable.

Intentional disclosure occurs when a statistical unit is identified using matching unique records in external files, combining common characteristics from both datasets.

Risk associated with access to microdata can be mitigated through:

- Direct data protection,
- Deterring attempts at identification (e.g., legal measures prohibiting attempts to identify and imposing penalties),
- Limiting access to microdata,
- Educating data users about privacy and their responsibilities (training and instructions), and
- Providing data access in a controlled environment.

It is impossible to eliminate all identification risks. The goal is to reduce the risk while maximizing the usefulness of data for statistical and research purposes.

Several software packages assist in assessing identification risk in microdata records, such as:

- **μ-ARGUS –** developed by Statistics Netherlands
- **SUDA –** developed by the University of Manchester, UK

6.4.3 Rules for Identifying Sensitive Data and Methods of Statistical Data Protection in Macrodata

Rules for identifying sensitive cells in macrodata/tables

To identify cells in tables that pose identification risks, rules for protecting confidential data must be established and applied to each cell.

If a cell does not meet these rules, further analysis is needed to minimize identification risk. Rules for identifying primarily sensitive cells include:

- **Threshold rule (minimum number of cases/statistical units):** a cell is sensitive if fewer than *t* cases contribute to it.

- **Dominance rule (n, k):** if *n* units represent more than *k%* of the cell value, the cell is sensitive. This applies where a small number of units contributes a large percentage of the cell total.

Threshold rule

In Table 1, the age group 15–19 has low income (20), which does not require protection since the result is neither unexpected nor sensitive. However, if the income were unusually high, the cell must be protected because the result is sensitive.

The minimum number of cases rule protects rare and unique combinations of characteristics that could lead to re-identification. The criterion is met if at least three cases contribute to the cell. If only 1 or 2 cases contribute, the value must be kept confidential.

Example: In Table 1, age group 50–59 has frequency 2 in the low-income cell, which must be protected. Applying a threshold of 3, any cell with fewer than 3 cases represents a disclosure risk.

Age group	Low	Medium	High	Total
15–19	20	0	0	20
20–29	14	11	8	33
30–39	8	12	7	27
40–49	6	18	24	48
50–59	2	5	14	21
60+	12	9	7	28
Total	62	55	60	177

Note: The threshold rule does not apply if only one characteristic is logically possible (e.g., all pregnant women in the dataset).

Dominance rule (n, k rule)

This rule is used for value tables. A cell is considered unprotected if the combined contribution of the *n* largest units exceeds *k%* of the total cell value. Values of *n* and *k* are determined by the data owner.

Example: Table 2 – Profit of business entities in furniture manufacturing. A cell shows total profit of 7,000,000 KM. Using the (2, 85) dominance rule, the two largest entities (A and B) must not contribute more than 85% of the total.

Entity	Profit (KM)
A	3,760,000
B	2,330,000
C	300,000
D	250,000

Entity	Profit (KM)
E	150,000
F	100,000
G	75,000
H	35,000
Total profit	7,000,000

Combined profit of A and B = 6,090,000 KM = 87% of total profit, exceeding the 85% threshold. Therefore, this cell is confidential and must be protected.

However, if entities consent in writing to release their individual data, or if some data is publicly available, protection may not be required, reducing information loss in tables.

Methods of Statistical Confidentiality Protection in Macrodata/Tables

Methods used to protect confidential data that present a risk of identification in an unprotected table cell (i.e., in macrodata) include:

1. **Primary and secondary cell suppression**
2. **Table redesign (reshaping/aggregation)**
3. **Data modification (perturbation)**

Primary and Secondary Cell Suppression

Cell suppression involves **not publishing information for unsafe cells** or deleting individual records from a microdata set.

If a table contains totals, the value of a suppressed cell can be calculated by subtracting other cell values from the total. Therefore, at least one additional cell must also be suppressed to prevent identification.

- **Primary cell suppression** refers to suppressing a cell that violates a confidentiality rule.
- **Secondary (consequential) suppression** refers to suppressing other cells to prevent indirect disclosure of the primary suppressed cell.

Example: In Table 1 (page 10), the cell showing the number of people aged 50–59 with low income is identified as an unprotected cell using the threshold rule of 3. This cell can be protected with suppression, marked as 'X' in Table 3. However, it is still possible to infer the value of the cell using remaining values. For instance, the number of people with low income aged 50–59 could be determined by subtracting medium and high income from the total (21–14–5 = 2). To protect the cell, **secondary suppression** of additional cells is required.

Table 3 – Secondary suppression

Age group	Low	Medium	High	Total
5–19	20	0	0	20
20–29	14	11	8	33
30–39	8	12	7	27
40–49	6 Y	18 Y	24	48
50–59	X	5 Y	14	21
60+	12	9	7	28
Total	62	55	60	177

- X = primary suppressed cell
- Y = secondary suppressed cells

By using this method, the X cell is suppressed, and the respective row and total column values are masked to prevent indirect identification of the unprotected cell.

Example of table protection:

Suppose the threshold rule for primary sensitive cells is 3. A table has one primary sensitive cell [11, A] with value 2.

Protection options:

- **Redesigning the table**: Combine categories to reduce the number of primary sensitive cells.
- **Suppression**: Use primary suppression for cell [11, A] and secondary suppression for [11, B], [12, A], [12, B] to prevent users from deducing the original value.

Unprotected table:

	Total	**A**	**B**
Total	40	21	19
1	18	11	7
11	6	2	4
12	12	9	4
2	22	10	12

After secondary suppression:

	Total	**A**	**B**
Total	40	21	19
1	18	11	7
11	6	Z	Z
12	12	Z	Z
2	22	10	12

After primary suppression:

	Total	**A**	**B**
Total	40	21	19
1	18	11	7
11	6	Z	4
12	12	9	4
2	22	10	12

Final protected table (after combining categories):

	Total	**A**	**B**
Total	40	21	19
1	18	11	7
2	22	10	12

Table Redesign (Aggregation/Restructuring)

Confidentiality of data providers can be protected by selecting appropriate aggregation. This method includes:

- Combining multiple response categories into one, or
- Reducing classification detail in tables.

It is important to ensure the new categories remain meaningful to users. Detailed classifications, such as country of birth, industry, or occupation, can be aggregated. For example, nurses and doctors can be grouped as "medical professions."

For quantitative or continuous data (income, age), categories can also be grouped into ranges.

Example – Reshaping table data

KD	**Size class**	**0–4**	**5–9**	**10–14**	**15–19**	**20+**
D	375	73	194	59	36	13
DA	90	29	55	1	5	-
DB	53	5	25	11	2	10
DC	30	3	4	12	11	-
DD	161	26	99	35	1	-
DE	41	10	11	-	17	3

After aggregation:

KD	Size class	0–4	5–9	10+
D	375	73	194	108
DA	90	29	55	6
DB	53	5	25	23
DC	30	3	4	23
DD	161	26	99	36
DE	41	10	11	20

Data Modification (Perturbation)

Perturbation involves slightly altering data to reduce identification risk while retaining as much structure as possible.

- The most common method is **rounding**, i.e., slightly adjusting small table values so that analyses remain reliable but exact original values are obscured.
- For example, **random rounding base 3** means all table values are rounded to the nearest multiple of 3 independently, including totals. Values already divisible by 3 remain unchanged.

Before rounding:

Age group	Low	Medium	High	Total
5–19	20	0	0	20
20–29	14	11	8	33
30–39	8	12	7	27
40–49	6	18	24	48
50–59	2	5	14	21
60+	12	9	7	28
Total	62	55	60	177

After rounding (base 3):

Age group	Low	Medium	High	Total
15–19	21	0	0	21
20–29	15	12	9	33
30–39	9	12	6	27
40–49	6	18	24	48
50–59	3	6	15	21
60+	12	9	6	27
Total	63	54	60	177

Steps in Table Protection Procedure

1. Define a rule to measure table sensitivity (determine which cells are sensitive).
2. If the number of sensitive cells is too large, **redesign the table**.
3. Choose a protection method.
4. Apply the chosen protection method to the table.
5. Evaluate protection results (balance protection vs. information loss).
6. If satisfied, **publish the table**.
7. If not satisfied, return to step 2.

Outcome: Sensitive cells cannot be partially or fully disclosed.

6.4.4 Rules for Identifying Sensitive Data and Methods for Statistical Protection of Confidential Microdata

Rules for identifying sensitive data in unprotected microdata records

To assess the risk of identification in unprotected microdata records, the following rules are used to determine sensitive data:

- **Threshold rule** – the minimum frequency of certain combinations of variables that may be published.
- **Public databases** – these are not protected since they are publicly available sources of data.

Methods for statistical protection of confidential microdata

Various methods (techniques) are used to statistically protect confidential data that present a risk of identification in unprotected microdata.

To protect microdata confidentiality, **perturbation** and **data reduction** methods are applied. These are the same basic principles used for aggregated (macrodata) protection.

Common techniques for protecting microdata confidentiality include:

- Limiting the number of variables included in a dataset;
- **Recoding/aggregation** of categories that may allow identification (e.g., reporting age in five-year ranges);
- **Top/bottom coding** of extreme values for continuous variables (e.g., income, age);
- **Suppression** of certain values or records that cannot otherwise be protected from identification risk;
- **Data swapping** – replacing values in one record with values from another record with similar characteristics to hide unique records.

When choosing methods for microdata protection, one should first decide which statistics should be modified as little as possible.

Defining variable sets

Before applying protection methods, the following variable sets should be defined:

- **Identification variables** – variables that may reveal the identity of a unit:
  - **Direct identifiers –** allow direct identification of a statistical unit (e.g., personal ID number).
  - **Indirect identifiers –** allow indirect identification, possibly in combination with other variables (e.g., name, year of birth, address, sex, occupation, region).
- **Confidential output variables –** sensitive variables carrying confidential information about the statistical unit (e.g., income, religion, political beliefs, health status).
- **Non-confidential output variables –** all other variables.

Even with all statistical protection techniques, the risk of identifying confidential data is **never zero**.

Applying statistical protection methods

1. **Remove direct identifiers** from microdata (e.g., ID numbers) and selected indirect identifiers (e.g., name, surname, address).
2. Remaining indirect identifiers (e.g., age, sex, year of birth, occupation, region) are retained as **key variables** important for statistical analysis.
3. Choose protection methods for the key variables based on the intended use of the data (e.g., public file, researcher file, student file).
4. Apply protection to reduce the risk of identification to a reasonable level.
5. Inform users of the methods used, as some values in the protected file will differ from the original file. Parameters used in protection methods (e.g., set of indirect identifiers, threshold values, dominance rules) **must not be disclosed**, as this would reduce protection.

Example of microdata statistical protection

Consider a dataset of individuals with demographic characteristics and monthly total income:

YEAR OF BIRTH	SEX	PLACE OF RESIDENCE	OCCUPATION	TOTAL MONTHLY INCOME
1985	M	ZENICA	ECONOMIST	1800
1985	M	ZENICA	ECONOMIST	1900
...	...	...	...	...

Direct identifiers (e.g., ID number) and some indirect identifiers (e.g., name, surname, address) are already removed.

Next, a **threshold of 3** is applied to remaining indirect identifiers (year of birth × sex × place × occupation). This ensures that in the protected dataset, each combination appears at least 3 times.

Combinations before protection:

Combination	Count
1985 – M – ZENICA – ECONOMIST	6
1985 – M – ZENICA – LAWYER	4
1985 – M – DOBOJ – ECONOMIST	2
1985 – M – DOBOJ – LAWYER	3
1985 – M – TUZLA – ECONOMIST	1
1985 – M – TUZLA – LAWYER	3

Two combinations appear less than the threshold (1985 – M – DOBOJ – ECONOMIST and 1985 – M – TUZLA – ECONOMIST).
Protection method: suppression – mask one of the indirect identifiers. In this example, the place of residence is replaced with a code 99 for confidentiality.

Protected microdata:

YEAR OF BIRTH	SEX	PLACE OF RESIDENCE	OCCUPATION	TOTAL MONTHLY INCOME
1985	M	99	ECONOMIST	1950
1985	M	99	ECONOMIST	1650

Recomputed combinations in the protected file:

Combination	Count
1985 – M – ZENICA – ECONOMIST	6
1985 – M – ZENICA – LAWYER	4
1985 – M – 99 – ECONOMIST	3
1985 – M – DOBOJ – LAWYER	3
1985 – M – TUZLA – LAWYER	3

All combinations now meet the threshold.

Steps in the microdata protection process

1. Define a rule to measure microdata sensitivity.
2. Select a protection method.
3. Apply protection to the microdata file.
4. Evaluate results (protection vs. information loss).
5. If satisfied, release the protected microdata.
6. If not satisfied, redesign variables and repeat step 3.

Outcome: Individual records in the microdata file cannot be recognized.

Active confidentiality protection

All statistics (except international trade in goods and manufacturing sales statistics) must apply **active confidentiality protection techniques**.

- The most important rule is the **case threshold criterion** (minimum of three observations per table cell).
- **Dominance rule** for economic variables: the two largest statistical units together must not exceed 85% of the total cell value.

Primary suppression is applied first (individual groups are considered separately, using threshold and/or dominance criteria). **Secondary suppression** is applied manually to prevent reconstruction of confidential data from published totals. This is especially important in detailed statistics with hierarchical structures (industry, firm size, region).

Passive confidentiality protection

For international trade statistics (goods/services), confidentiality is applied **only upon request** from a legal entity.

- Passive protection means table results are not treated as confidential unless requested.
- If requested, protection may be applied only for products where the entity has a dominant market position.
- If international trade statistics are classified by characteristics of legal entities, the confidentiality is similar to other business statistics, but dominance is calculated based on turnover.

Rules for user access to confidential data

External users:

- Access to anonymized microdata is granted to registered research institutions and researchers.
- Requires:
    - Completion of an access form;
    - Signing a data protection agreement;
    - Signing a data use agreement.
- Microdata are currently available on portable media or in a protected room/remote access system.

Internal users:

- Agency employees require approval from their supervisors for access, justified strictly by their work needs.

Guidelines

- Confidentiality protection is essential to maintain trust of statistical units.
- Verify confidentiality before publishing results (both aggregated tables and microdata).
- Names and addresses must be removed from microdata as early as possible.
- Inform statistical units that their data will be protected and used only for statistical purposes.

- Groups in table cells must have at least three units; a single unit must not exceed 85% of the total.

- Inform users about the confidentiality methods applied.
- De-identified data do not automatically mean full protection; both direct and indirect identifiers must be considered.
- Examine links between tables and variables to avoid disclosure.
- Limit variable detail to what is necessary to minimize information loss.
- The number of variables in a microdata file should not exceed the minimum necessary.
- Special status may be assigned to units allowing publication while reducing information loss.

7 DATA DISSEMINATION

7.1 Updating Output Results

The publication of statistical data and information is carried out according to standardized procedures across different technologies.

Standardized procedures are based on predefined structures, formats, and metadata, which are taken into account when preparing tables (at the stage of statistical data processing). The same standardized procedures apply to all content in accordance with the principle of process transparency and considering timeliness. When preparing new content (such as regular updating of output data), organized documentation, internal knowledge exchange, standard communication channels, and archiving of materials and work procedures are important.

For every form of statistical data and information release, output data should be updated in advance; these can be data in the form of databases or tables with final aggregated data, prepared for publication.

For statistical surveys where data are revised to ensure better quality, in accordance with the Data Revision Guidelines, all output data must be updated during the revision phase, and the revised data should be published.

GUIDELINES

- When updating output results, standardized procedures must be taken into account and implemented transparently (clearly and understandably). Processes that are repeated periodically (at regular time intervals) should be automated.
- Documentation must be organized, published, accessible, and regularly updated.
- If multiple people from different areas are involved in the process, internal coordination must be documented.
- Content removed due to methodological changes must be properly archived.
- Data revision rules specified in the guidelines must be considered.

7.2 Production and Presentation of Products for Publication

Elements of the release content must be prepared considering the target users and the purpose of the data. Data should be presented and published appropriately. The presentation of results should be efficient, understandable, simple, and engaging.

When presenting results, the principle of multilingualism should be considered; as a rule, all publications are published in multiple languages (Bosnian/Croatian/Serbian and English versions).

General principles for presenting results vary depending on the type of publication and medium. The Agency disseminates statistical information through:

- Website at www.bhas.gov.ba;
- Printed and electronic publications;
- Press releases;
- Press conferences;
- Media and social networks;
- Direct interaction with users based on standard requests and inquiries.

When presenting data, users' statistical literacy and varying understanding of statistical data and information should be considered. Professional users mainly use detailed data for further analyses, so such data should be published in databases in formats that allow further processing (electronic form). The general public is primarily interested in the most important or most interesting overall statistical data, presented in a comprehensible (popular) form.

The author of the data presentation must consider what the user wants to see and how to present it in a way that is understandable, engaging, and useful. Data published in the initial electronic release – press releases and popular publications – should be presented as a statistical story. This story should include commentary and data visualization. The commentary must be effective – short, simple, understandable, and interesting.

Data visualizations must be created and summarized in a clear form with simple tables and graphs. Presentation should make the results of the statistical business process accessible to the public and/or users. The following modes of presentation are distinguished:

- Tabular
- Textual
- Graphical and cartographic

Table Preparation

A table is the classic form of statistical result presentation. Publications in official statistics (paper or electronic) largely consist of tables. Tables are not produced only for printed publications. Displaying results in table form on the internet is a necessity for official statistics. A table is a tool for conveying large and complex information in a comprehensible and concise form. A table consists of a first column, table header, and table body. The distribution of characteristics in the first column and header depends on what is being compared. Tables should not be overly complicated. The table header should not have more than four levels of breakdown. The first column should ideally not exceed three levels of breakdown. Too many levels of detail reduce comprehensibility. Longer texts are often more suitable in the first column. Check if the table becomes easier to understand if the first column and header are swapped.

Text Preparation

In publishing statistical results, text serves a supporting role. Generally, the text should explain everything necessary to understand the presented tabular results. This also includes a brief description of the concepts and methods used. Detailed methodological explanations and metadata will be provided in the Quality Reports for individual statistical surveys. Press conferences are usually used for the first publication of current statistical data. They provide (in the form of a Press Release) information for the media in a very concise format (a few pages, methodological instructions, and 1–2 tables). The structure of the text follows the usual rules for this form of publication.

Graphical Presentation

Various forms of graphical presentation are used for statistical data.

- For structure and total quantities → "Column Chart (vertical)", "Column Chart (horizontal)", and "Pie Chart".
- For ranking → horizontal column chart.
- For time series or trends → vertical column chart or line chart.
- For regional comparisons → maps.
- For dispersion → scatter plot, box plot.

GUIDELINES

Table Preparation

- All tables should contain at least the following components: table title, header, pre-columns, numerical entry fields, and footer.
- The table title should include the subject (data) of the presentation, the data structure from the columns, and the time period (or date) the data refers to. Temporal, subject, and spatial delineation of the content should be clear from the title.
- The header should clearly indicate what part of the content each column's figures refer to. The measurement unit of the presented data should also be indicated.
- Numbers in cells should be legible. When choosing units of presentation, avoid showing excessive digits; round numbers where appropriate.

Additional conventions for table presentation:

- If a cell is empty → write a dash (-).
- If a value is less than half the minimum unit → write zero (0 or 0.0).
- If relevant information is not yet available but expected → write three dots (...).
- If the value will not be disclosed due to statistical confidentiality → mark the cell with "T" to conceal the true value.

Footnotes should always include the data source, e.g., "Agency for Statistics of Bosnia and Herzegovina". Footnotes may also include comments on individual numeric values within the table.

Text Preparation

- The textual complement to tabular publications should briefly and clearly cover all necessary aspects for understanding the presented figures.
- Do not repeat the table figures in the text.
- Highlight only the most important or striking results.
- Avoid terms such as “increase, growth, positive or negative rate of change,” and instead use: increased – decreased; higher – lower, etc. For example:
    - Incorrect: “The number of air passengers in 2023 compared to 2022 increased by 7.7%” → Correct: “The number of air passengers in 2023 compared to 2022 increased by 7.7%.”
    - Incorrect: “Negative rate of change of production compared to April 2022 is –5.3%” → Correct: “Decrease in production compared to April 2022 is 5.3%.”
    - Terms “growth” and “fall” can only be used for the largest increases or decreases.
- Press releases (press conferences), as the most concise form for public dissemination, are subject to special rules:
    - The main headline and first paragraph determine whether the press release attracts media attention.
    - The headline should contain the most important results.
    - The first paragraph summarizes what will be communicated further in the press conference.
    - When shortening text for media purposes, accuracy must be ensured.
- All statistical publications should follow:
    - Professional principles (precision, clarity, expertise).
    - Word choice adapted to the “typical reader/listener” (avoid jargon; define essential technical terms).
    - Provide a list for frequently used abbreviations.
- Every publication must undergo intensive editorial and content review. The director, or, with permission, the head of the publication sector, is responsible for the publication.

Graphical Preparation

- The type of graph (bar chart, line chart, pie chart, etc.) should be chosen according to the solution.
- All graphs or maps must be properly labeled (axes, data series, important data, and legend).
- The graph title must indicate what it contains.

7.2.1 Statistical Database

Publishing statistical results in print or using fixed online tables has the limitation of a restricted number of tables and estimates available. Therefore, users are provided with the opportunity to create analytical tables from existing data.

Statistical agencies store data in statistical databases to allow users to create their own output tables and results.

Statistical data stored in the database are generally characterized by two aspects:

- Certain statistical units (e.g., business entities) are stored, for which certain characteristics are then aggregated (e.g., turnover, employees).
- The aggregation process is defined according to certain characteristics (breakdown criteria). These characteristics allow temporal, spatial, and content-based differentiation of the observed values.

Producing multidimensional tables is associated with certain problems:

- Small cell counts may create confidentiality issues.
- Survey data may contain some uncertainty. Greater breakdown detail may result in cells with higher error probability, making their value problematic.

GUIDELINES

- Whenever possible, all published statistical materials should be stored in databases.
- When storing materials in databases, decide which characteristics and at what level of detail should be available. Ensure temporal and spatial delineation is possible.
- Before storing materials, ensure there are no confidentiality issues. Define how to handle cells whose values cannot be displayed due to confidentiality rules.
- New materials must be verified before release, ensuring values are identical before and after the technical conversion process.
- Stored materials must be regularly maintained. Periodically updated segments should be especially current.
- When materials are first stored, inform key users about database availability.
- Database usage must be continuously monitored. Track access numbers and ensure technical conditions allow adequate use (availability, user satisfaction, processing speed).

7.3 Management of the Publication of Dissemination Products

The mission of Statistics Bosnia and Herzegovina is to provide reliable, high-quality, understandable, timely, and internationally comparable statistical data that meet the needs of decision-makers, researchers, and other domestic and foreign users, reflecting the state and changes in the economic, demographic, and social domains, as well as the environment and natural resources.

The collection, processing, analysis, and dissemination of statistical data are carried out based on statistical standards and modern technology, while ensuring the protection of statistical confidentiality, optimal use of resources, and reasonable burden on data providers.

All published data and information are accessible to users on the websites of the Agency and the entity statistical offices.

Every publication of statistical data must be announced in accordance with the **Publication Calendar**. The Publication Calendar is prepared at the end of the year for the following year, and each publication must be confirmed no later than the Friday prior to the week of release.

The dissemination of statistical products and services of the Agency is conducted in accordance with the following principles:

- **Availability** of all users to the results of statistical surveys included in the Work Plan;
- **Equality** of all users in access to the results of statistical surveys included in the Work Plan;
- **Objectivity and impartiality** in presenting the results of statistical surveys included in the Work Plan;
- **Strict adherence** to deadlines for publishing statistical survey results according to the previously announced Publication Calendar;
- **Protection of statistical confidentiality** of individual data in accordance with the Statistics Law of BiH and the Agency's Regulation on the Protection of Statistical Data;
- **Clarity and comprehensibility** in publishing the results of statistical surveys included in the Work Plan;
- **User orientation**, including publication of main information in English.

The status of data (temporary or final) must be determined already at the planning stage of publication.

Only data that have already been published may be forwarded to international organizations or to Eurostat. If confidential data are sent to Eurostat, the confidential cells must be marked.

Procedures for data publication vary depending on the type of release. Accordingly, the following are distinguished:

- Publishing news on the website;
- Publishing printed publications;

- Publishing data in databases;
- Forwarding data to Eurostat;
- Updating data in interactive tools;
- Correcting errors in publications.

Activities related to publishing statistical data also include correcting errors. The purpose of error correction in published statistical data is to ensure accurate and high-quality statistical data and information for users. The error correction system must be structured in a way that is clear and understandable to users. Exact procedures are defined in the **Guidelines for Correction of Errors in Published Releases**.

GUIDELINES

- Data must be published exactly according to announcements in the Publication Calendar, which is available to users on the websites.
- The entire statistical survey process must be planned so that data are published in a timely manner.
- Statistical data and information, even if extremely accurate and detailed, are of limited use if not published promptly.
- Data must be equally available to all users. Users should also have as many options as possible for additional requests and data orders.
- Data publication must be as transparent as possible. Along with data, corresponding metadata must be available, such as methodological explanations, key quality indicators, questionnaires, etc. Clarity of error correction and data revision procedures must also be ensured.

7.3.1 Delivery of Data to Eurostat

A significant portion of the statistics produced by the Agency is based on recommended European standards and good practices. This primarily applies to statistics where there is a strong emphasis on harmonization of concepts at the European level.

Eurostat is the institution to which data are delivered according to agreed procedures and modes of delivery. The scope, deadlines, and content of the delivered data are determined by the relevant EU regulations. In principle, data transfers may include:

a) Transfer of table programs;
b) Transfer of individual data; and
c) Transfer of metadata in the form of a **Quality Report** (detailed methodological description).

The transfer of tables to Eurostat is generally part of the overall publication program. This primarily involves delivering selected results. Quality reports serve as evidence to Eurostat that agreed recommendations and quality standards are being followed.

GUIDELINES

- The Agency will take all necessary measures to meet Eurostat requirements. This refers to both the content and scope of the data delivery program, as well as the deadlines and mode of delivery.
- The delivery of data or metadata to Eurostat must be documented. All relevant procedures should be described in the Quality Report.
- Individual data are delivered to Eurostat only in anonymized form (statistically protected microdata). Delivery follows a pre-agreed format.
- Agency staff will work with Eurostat to ensure that the content and format of delivered metadata remain stable over time, to achieve a high degree of temporal comparability and data quality.

7.3.2 Delivery of Data in Response to International Requests

Mutual communication between national statistical offices and other national and international organizations, through which statistics are exchanged, is of utmost importance. This collaboration occurs not only in working groups, meetings, and conferences but also through requests sent directly to the Agency. The Agency strives to respond adequately to these international requests.

GUIDELINES

- Responses to international organizations' requests must be timely and as complete as possible.
- Upon receiving a request from an international organization, the **Department for International Cooperation** must be immediately informed.
- All contacts at locations where requests are processed should be forwarded to the knowledge of the Department for International Cooperation. Archiving is centralized and managed by the service organizationally responsible for it (Dissemination Sector).
- Requests must be forwarded to technically competent staff.

7.4 Promotion of Dissemination Products

While marketing can generally be considered a comprehensive process, this subprocess refers to the active promotion of statistical products produced within a specific statistical business process in order to help them reach the widest possible circle of users. This includes the use of tools for managing user relationships, as well as the use of tools such as websites and blogs to facilitate the process of communicating statistical data to users.

The best way to obtain feedback from users is through surveys measuring their satisfaction. Based on this feedback, it is possible to conduct analysis and evaluation of users' needs.

7.5 User Support

Users must be enabled to access data resulting from statistical surveys, while ensuring the protection of statistical confidentiality. User support is primarily provided via e-mail and telephone, as well as through personal contact.

User support includes mediation of statistical data and information, guidance on accessing data, advice on using data preparation tools, and assistance in searching for and preparing statistical data and information.

Statistical data and information that have been published, as well as data whose preparation does not require additional processing, are free of charge. Preparation of data according to the user's specific request is charged according to the applicable price list. Providing data within the scope of reporting to international organizations is free of charge.

A record of written requests must be maintained. This is necessary for the preparation of analyses regarding users, methods of communication, and the most frequently requested statistical areas.

Users are also allowed access to statistically protected microdata, but under special conditions and a special agreement. Statistically protected microdata may be obtained by registered research institutions and registered researchers.

Methods of user support may vary depending on the type of request. The following forms of requests are distinguished:

- Written requests for data;
- Requests via telephone;
- Personal visits by users;
- Access to statistically protected microdata;
- Access to library materials for users.

GUIDELINES

- Data provided to users must be accurate, relevant, and appropriately presented.
- Users must receive data in a timely manner. Users must be immediately informed of any potential delays.
- All data are free of charge. Only additional costs incurred for more demanding requests may be charged.
- Information on how to access statistical data must be published in multiple places (website, publications).
- A telephone number for information requests must be available during the institution's working hours.
- Basic explanations regarding access to microdata for scientific research purposes must be published on the statistical institution's website.

8 EVALUATION

Statistical surveys are generally conducted periodically, so that the entire statistical process is repeated. It is important that this process includes feedback, which enables the introduction of changes and improvements. For this purpose, each statistical survey must be fully evaluated after its completion, and this evaluation should critically assess the overall success of the survey and identify opportunities for improvement.

Collecting information on the quality of statistical data takes place throughout the entire statistical process. Systematic documentation of individual parts of the survey represents an important part of information about the survey's progress and helps identify potential systematic errors in the process. Using this information, we can assess the quality of statistical data and critically evaluate the results obtained, which are important for users, as it provides them with additional insight into the data collection process. Publishing information on data quality represents a transparent way of informing users about various aspects of statistical data.

8.1 Collection of Survey Documentation

The procedure for preparing survey documentation includes a detailed description of the statistical activity, including the description of concepts, definitions, methods used, information systems used, and working instructions. The quality of survey documentation is an important indicator of survey quality. It is also an important tool for communication between different surveys, as well as between producers and users of statistics. Survey documentation is part of the metadata.

Documentation can generally be divided into documentation for users of survey results and documentation for producers of statistical surveys.

Documentation for users of survey results describes and documents statistical results and is publicly published. The purpose of documentation for users is to inform them about what data and statistical methods are measured, to help them understand the data, to make it easier to request data, and, if necessary, to enable further processing by themselves. Examples of documentation for users include survey questionnaires, methodological instructions, and user-oriented reports on the quality of statistical surveys.

Documentation for producers of statistical surveys describes the statistical procedures and processes used in the overall statistical process. This documentation is produced during the implementation of individual steps in the survey process and is mainly intended for internal use. The purpose of documentation for producers of statistical surveys is to ensure efficient and reliable data production. The documentation for producers must, above all, describe in detail the steps used in conducting the survey (e.g., defining the target population, sample selection, questionnaire design, data editing, data publication, etc.).

GUIDELINES

- Survey documentation must be accurate, detailed, and understandable to the target group for which it is intended.
- Standard forms and templates should be used in preparing documentation (if available). Also, the content of individual processes should, as much as possible, follow a standard structure.
- Documentation for both users and producers must be accessible to all interested parties.
- Documentation should be written to ensure maximum transparency of statistical procedures and products.
- User documentation must document everything that could potentially cause confusion or misunderstanding during use.
- The content and level of detail of the documentation should be adapted to the target users.
- Process documentation should also include descriptions of the procedures to follow in case of errors, as well as the names of persons or departments to contact in case of problems or uncertainties.

8.1.1 Storage of Aggregated Statistical Data and Statistical Microdata

Statistical institutions store aggregated statistical data for further use, both in electronic and printed form.

Archiving aggregated data in electronic form is carried out as part of implementing the policy of backup and storage of electronic data. Aggregated statistical data in printed form may be published in publications issued by the Agency and entity statistical offices, or in publications obtained from other institutions. All publications issued by the Agency and entity statistical offices are stored in at least one copy.

Printed copies of publications are sent to archives. Data on publications are entered into the publication catalog. Foreign serial statistical publications are archived, each in one copy. The responsible employee of the competent statistical institution prepares an annual list of library materials for withdrawal. The material is reviewed by a commission, which determines the list of materials for removal.

Competent statistical institutions store statistical microdata for further use in electronic form for research and analytical purposes. Storing statistical microdata in electronic form is carried out as part of implementing the policy of backup and storage of electronic data.

All databases containing data from various statistical surveys and metadata, which are of permanent value to the competent statistical institutions, are subject to a unified backup system based on good professional practice.

Daily backups are made only for data that was modified on that day. Weekly backups are performed for the entire server. These backups are stored in a secure storage location for media. The backup process of database updates is tested at least once a month. Statistical microdata in electronic form are stored on a shared file server, where both data and metadata files are saved.

GUIDELINES

- It is necessary to regularly monitor professional requirements regarding backup and archiving.
- Copies must be regularly sent to the archive, and publications must be recorded in the PUBLICATION CATALOG.
- Backup and storage must be carried out in accordance with instructions for creating and storing backup copies.

8.2 Conducting Evaluation – Evaluation of Results

To make the application of methods used in data production understandable to users of statistical products, sufficient metadata must be made available.

In statistical databases, metadata refers to data that is not directly presented in statistical content but is essential for understanding the process of producing a statistical product (e.g., industry or occupation names, a list of municipalities, etc.).

A statistical metadata item also includes descriptions of questionnaire data fields, variable definitions, explanations, and instructions for completing forms. Actual statistical data are referred to as (in contrast to metadata) microdata and macrodata. Without sufficient metadata, users may not have all necessary information, which can lead to incorrect interpretations and decisions.

For these reasons, preparing a Quality Report for individual statistical surveys is of particular importance. These reports are not only for internal use but also contain metadata necessary for the complete use of the statistical product.

Documenting the statistical process and its results, providing access to metadata, and reporting on quality through appropriate indicators are directly related. In this sense, the Agency and entity statistical offices have a standardized system for reporting the quality of statistical products.

The Quality Report for statistical surveys is structured according to an established fixed schema. The structure of the Quality Report for statistical surveys has the following format.

1. STATISTICAL PROCESS AND STATISTICAL PRODUCT

---

2. RELEVANCE
2.1 Users of Statistical Survey Data
2.1.1 Key Users of Statistical Survey Data
2.1.2 Assessment of User Needs
2.1.3 Measuring User Perception and Satisfaction
2.2 Data Completeness
2.2.1 Quality and Performance Indicator – Data Completeness Rate (R1)

---

3. ACCURACY AND PRECISION
3.1 Sampling Error
3.1.1 Quality and Performance Indicator – Sampling Error (A1)
3.1.2 Activities to Reduce Sampling Errors
3.2 Non-sampling Errors
3.2.1 Non-sampling Errors – Coverage Errors
3.2.1.1 Quality and Performance Indicator – Overcoverage Rate (A2)
3.2.1.2 Quality and Performance Indicator – Proportion of Duplicated Units (A3)
3.2.1.3 Undercoverage Errors
3.2.1.4 Measures to Reduce Coverage Errors
3.2.2 Non-sampling Errors – Measurement Errors
3.2.2.1 Causes of Measurement Errors
3.2.2.2 Measures to Reduce Measurement Errors
3.2.3 Non-sampling Errors – Nonresponse Errors
3.2.3.1 Quality and Performance Indicator – Unit Nonresponse Rate (A4)
3.2.3.2 Quality and Performance Indicator – Item Nonresponse Rate (A5)
3.2.3.3 Procedures in Case of Nonresponse
3.2.3.4 Procedures to Reduce the Nonresponse Rate
3.3 Revisions
3.3.1 Quality and Performance Indicator – Average Size of Data Revisions (A6)
3.4 Imputation
3.4.1 Quality and Performance Indicator – Rate of Imputed Data (A7)

---

4. TIMELINESS AND ACCURACY OF RELEASE
4.1 Timeliness of Release
4.1.1 Quality and Performance Indicator – Timeliness of First Results (TP1)
4.1.2 Quality and Performance Indicator – Timeliness of Final Results (TP2)
4.2 Accuracy of Release
4.2.1 Quality and Performance Indicator – Accuracy of Release (TP3)
4.3 Reasons for Major Delays and Measures to Improve Timeliness and Accuracy

---

5. CONSISTENCY AND COMPARABILITY
5.1 Consistency
5.1.1 Quality and Performance Indicator – Consistency with Reference Source Results (CH1)

5.1.2 Reasons for Major Deviations
5.2 Comparability
5.2.1 Quality and Performance Indicator – Discrepancy of Comparable Statistics (CC1)
5.2.2 Quality and Performance Indicator – Length of Comparable Time Series (CC2)
5.2.3 Breaks in Time Series
5.3 Geographical Comparability
5.3.1 Comparability with Other Members of the European Statistical System

6. ACCESSIBILITY AND CLARITY, DISSEMINATION FORMAT
6.1 Press Releases for Data Publication
6.2 Publications for Data Release
6.3 Online Database
6.4 Access to Microdata
6.5 Availability of Methodological Documentation
6.6 Measures to Improve the Understandability of Disseminated Results
6.7 Quality and Performance Indicator – Consultation of Data Sets (AC1)
6.8 Quality and Performance Indicator – Consultation of Metadata (AC2)
6.9 Quality and Performance Indicator – Completeness Rate of Metadata (AC3)

7. SURVEY COSTS AND RESPONDENT BURDEN
7.1 Costs of Conducting the Statistical Survey
7.2 Respondent Burden
7.3 Measures to Reduce Costs and Burden

8. CONFIDENTIALITY
8.1 Confidentiality – Policy
8.2 Confidentiality – Data Handling

9. STATISTICAL PROCESSING
9.1 Data Source
9.2 Frequency of Data Collection
9.3 Data Collection
9.4 Data Validation
9.5 Data Compilation
9.6 Adjustments
9.6.1 Seasonal Adjustment

8.3 Action Plan for Improvement

This sub-phase brings together all necessary decision-making capacities to form and agree on an action plan based on the evaluation report. It includes steps or actions to monitor the implementation of recommendations in the future through the proposed mechanism at the institutional level.

LIST OF REFERENCES / LITERATURE

- **Eurostat, 2005**, *Handbook on Data Quality Assessment Methods and Tools* http://www.cmi.edu/research/rmi_government/EC/DatQAM_handbook_final.pdf

- **Eurostat, 2009**, *ESS Handbook for Quality Reports* http://epp.eurostat.ec.europa.eu/portal/page/portal/ver/quality/documents/EHQR_FINAL.pdf

- **Making Data Meaningful (2009), Part I – A Manual for Writing About Numbers**, UNECE, United Nations Economic Commission for Europe https://unece.org/DAM/stats/documents/writing/MDM3_Croatian_version.pdf

- **Making Data Meaningful (2009), Part II – A Manual for Presenting Statistics**, UNECE, United Nations Economic Commission for Europe https://unece.org/DAM/stats/documents/writing/MDM_Part2_Croatian.pdf

- **ABS, 2010**, *Information Paper: Quality Management of Statistical Processes Using Quality Gates*, Dec 2010, cat. no. 1540.0, ABS, Canberra

- **Statistik Austria, 2010**, *Quality Guidelines, Version 1.1 – as of 14.12.2010* http://www.stat.at/web_de/ueber_uns/aufgaben_und_grundsaetze/qualitaet/index.html

- **Statistical Office of the Republic of Slovenia, 2012**, *Guidelines for Ensuring Quality* http://www.stat.si/doc/pub/Smernice.pdf

- **UN, 2014**, *Fundamental Principles of Official Statistics of the United Nations* https://unstats.un.org/unsd/dnss/hb/E-fundamental%20principles_A4-WEB.pdf

- **Eurostat, 2017**, *European Statistics Code of Practice* https://op.europa.eu/en/publication-detail/-/publication/661dd8ef-7439-11e8-9483-01aa75ed71a1/language-hr/format-PDF

- **Statistics Canada, 2018**, *Statistics Canada's Quality Assurance Framework* https://www150.statcan.gc.ca/n1/pub/12-586-x/12-586-x2017001-eng.htm

- **Statistics New Zealand, 2019**, *A Guide to Good Survey Design, Fifth Edition* https://www.stats.govt.nz/assets/Uploads/Methods/A-guide-to-good-survey-design-fifth-edition/a-guide-to-good-survey-design-fifth-edition.pdf

- **European Statistical System (ESS), 2020**, *Handbook for Quality and Metadata Reports – 2020 Edition*