Official Statistics for SDGs- Lesson 4

4.1 Lesson4

Notes:

Lesson 4 outlines statistical methodologies whose standards, guidelines or manuals are compiled internationally.

4.2 Lesson Outline

Lesson Outline

4-1 Statistical Standards

4-2 Classifications

4-3 Survey Methodologies

4-4 Statistical Registers

4-5 Dissemination Methodologies

4-6 Rough Sketch of Official Statistics by Scope

Notes:

This lesson focuses on the statistical methodologies for official statistics.

Statistical methodologies for official statistics are discussed and reached consensus as the statistical standards. The main statistical standards are classifications, survey methodologies, statistical registers and dissemination methodologies that are applied to various official statistics across scopes. Some of them are often deliberated, reviewed and authorized as international standards. There are a lot of international standards for each subject of official statistics. This lesson only gives an overview of them by scope.

4.3 4-1 Statistical Standards

Notes:

4.4 International Statistical Standards
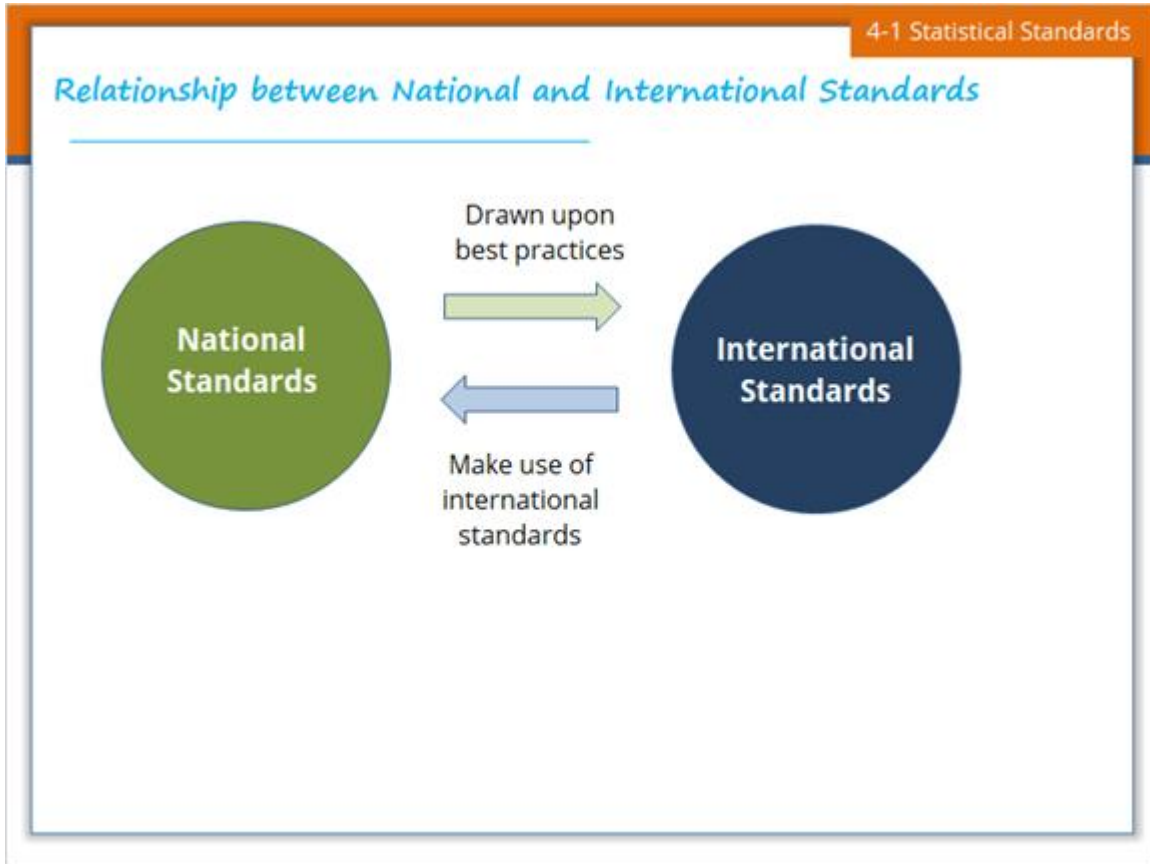

Notes:

Standards exist for all phases of statistical production and for nearly all statistical subjects.

Standards may be handbooks, recommendations manuals, classifications, guidelines, etc.

UNSD hosts a list of existing international statistical standards, which can be viewed by type, subject matter area or maintenance organization.

International statistical standards covered methodological guidelines, handbooks and manuals in the 20th century, but extended to establishing goals such as MDGs and SDGs paraphrased by statistical indicators in the 21st century.

4.5 Relationship between National and International Standards

Notes:

International and national standards inform each other: when international standards are created or updated, they drawn upon best practices from national standards, and offices that are in the process of creating or updating their national standards typically make use of existing international standards.

4.6 Process of Creating or Updating an International Statistical Standard
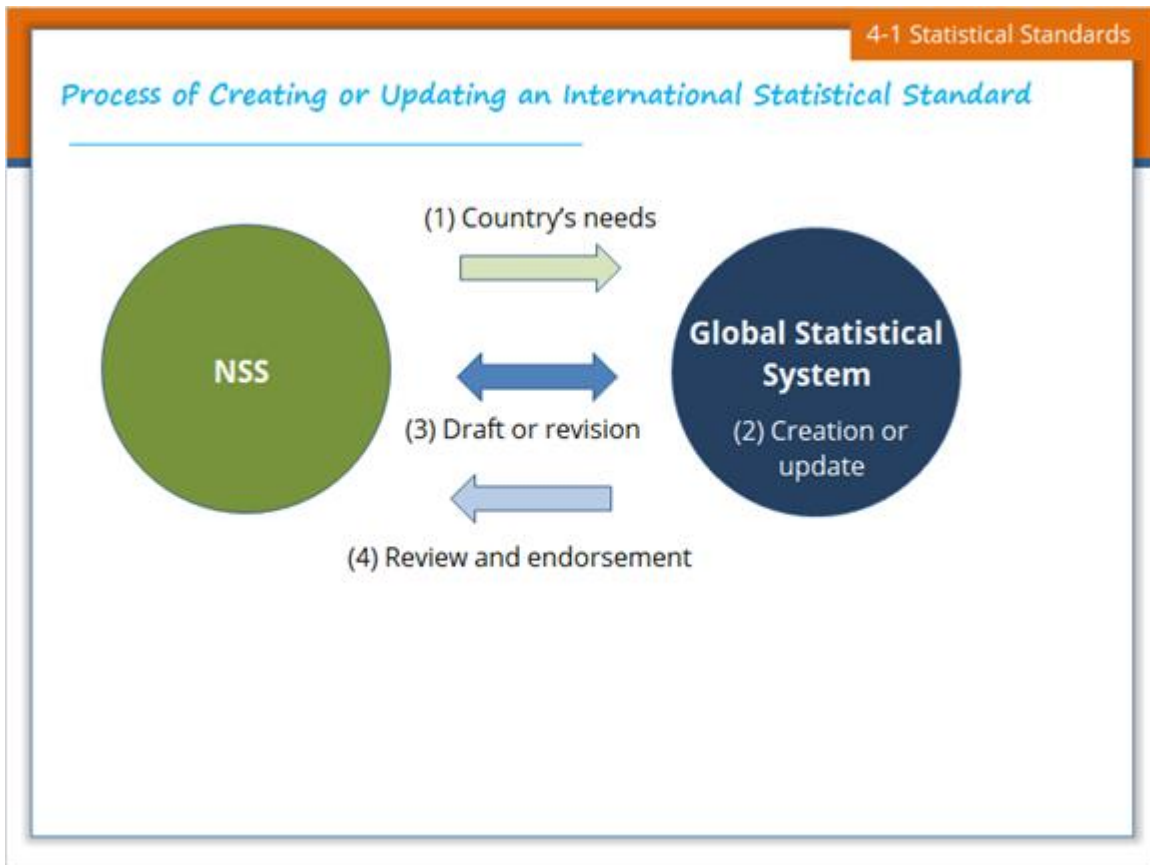
Notes:

The process of creating or updating an international statistical standard is usually started as a response by the Global Statistical System to countries' needs.

Typically the international organization responsible for creating or updating the standard will need to obtain a mandate for the work from its governing body.

The process of drafting or revising the standard will involve close communication with subject-matter experts from the NSSs and other agencies in the Global Statistical System.

Once the standard is completed, it will be sent back to the governing body of the agency for review and endorsement.

4.7 4-2 Classifications

4-2 Classifications

Notes:

4.8 Classifications

Classifications

- Have a set of discrete categories, which may be assigned to a specific variable registered in a statistical survey or in an administrative file, and used in the production and presentation of statistics
- May be flat or hierarchical
- Typically involves dividing a variable into categories that are **mutually exclusive** and **collectively exhaustive**

Notes:

In order to compare statistics, they need to follow common classifications. The classification depicts not only types or segments but also directions for data collection, compilation and dissemination of each statistics. Moreover, classification has the role of promoting connection between statistics.

Mutual exclusivity means that there should be no overlap between categories.

Collective exhaustiveness means that all categories together should cover every possibility.

4.9 Examples of Classifications Led by UNSD

Examples of Classifications Led by UNSD

Field	Abbrev.	Name
Area		Standard Country or Area Codes for Statistical Use
Industry	ISIC	International Standard Industrial Classification of All Economic Activities
Product	CPC	Central Product Classification
	SIEC	Standard International Energy Product Classification
Trade	SITC	Standard International Trade Classification
	BEC	Classification by Broad Economic Categories
Expenditure		Classification of Expenditure According to Purpose
	COICOP	Classification of Individual Consumption According to Purpose
	COPP	Classification of the Outlays of Producers According to Purpose
	COFOG	Classification of Functions of Government
	COPNI	Classification of the Purposes of Non-Profit Institutions Serving Households
Time	ICATUS	International Classification of Activities for Time-Use Statistics

Notes:

UNSD acts as the leader of various international classifications.

Standard Country or Area Codes for Statistical Use was established in 1969 and revised in 1998.

In terms of industry, ISIC was approved in the Economic and Social Council in 1948, revised for the first time in the Statistical Commission in 1958, and revised at the fourth Statistical Commission in 2008.

As for product, CPC was approved in the Statistical Commission in 1989 and revised in 2015, within which a part of Energy was revised as SIEC in 2014.

Regarding trade, general SITC was approved in the Statistical Commission in 1950 and revised in 2006. To provide a set of broad product categories for the analysis of trade statistics BEC was approved in 1971 and revised for the use in the System of National Accounts (SNA), Supply and Use table (SUT) in 2003, etc.

UNITED NATIONS
SIAP
Statistical Institute for
Asia and the Pacific

With regard to expenditure, Classification of Expenditure According to Purpose was established in 1999, COICOP for individuals, COPP for corporations, COFOG for governments, COPIN for non-profit organizations. They follow the concepts and definitions in 1993SNA.

For time use, ICATUS was drafted in 1999, tested in 2005, formalized in 2016 and approved in the Statistical Commission in 2017.

4.10 Examples of Classifications Led by WHO

Examples of Classifications Led by WHO

Field	Abbrev.	Name
Medicine	ICD	International Statistical Classification of Diseases
	ICF	International Classification of Functioning, Disability, and Health
	ICHI	International Classification of Health Intervention
	ICD-O	International Statistical Classification of Diseases for Oncology
		The ICD-10 Classification of Mental and Behavioural Disorders
	ICD-DA	Application of ICD to Dentistry and Stomatology
	ICD-10-NA	Application of ICD to Neurology
	ICPC	International Classification of Primary Care
	ICECI	International Classification of External Causes of Injury
		The Anatomical Therapeutic Chemical Classification System with Defined Daily Doses

Notes:

With regard to medicine and health, WHO established the family of health statistics classification: ICD for diseases, ICF for disabilities, and ICHI for health intervention. As derivative classifications, ICD-O was established for oncology, ICD-10 Classification of Mental and Behavioral Disorders for mental and behavioral disorders, ICD-DA for dentistry and stomatology, ICD-10-NA for neurology and ICF for children and the

younger generation. Furthermore, there are ICPC for primary care, ICECI for external causes of injury, and the Anatomical Therapeutic Chemical Classification System with Defined Daily Doses for anatomy, therapeutics and chemistry as related classifications.

4.11 Examples of Classifications Led by Other International Agencies

Examples of Classifications Led by Other International Agencies

Field	Abbrev.	Name
Labour		International Classification of Occupations for Migration and Employment Placement
	ISCO	International Standard Classification of Occupations
	ICSE	International Classification of Status Employment
Trade	HS	Harmonized Commodity Description and Coding System
	NST	Classification Systems Used in Transport Statistics
		International Classification of Non-Tariff Measures
Agriculture		Commodity Classification
Education	ISCED	International Standard Classification of Education
Crime	ICCS	International Classification of Crime for Statistical Purposes
Activity		Classification of Statistical Activities

Notes:

ILO created classifications related to occupations including industries. Classification of industries and occupations was resolved based on the System of Classification of Industries and Occupations at the first International Conference of Labour Statisticians in 1923. International Classification of Occupations for Migration and Employment Placement was established in 1952. ISCO was established based on the resolution of ILO in 1958, revised in 1968, 1988 and 2008. In terms of employment status such as employees and self-employed, ICSE was established in 1958 and revised in 1993.

UNITED NATIONS
ЛIap
Statistical Institute for
Asia and the Pacific

As for trade, HS, comprising about five thousand products identified with six-digit numbers, was established in 1988 and have been revised once in three to four years by WCO. NST was established by ECE in 2000 and revised in 2004 and 2006. For non-tariff measures the International Classification of Non-Tariff Measures was established in 2012 and classified technical methodology (sanitation, environment protection, etc.) and other methodology (allotment, price control, export regulation, etc.) which had been used as the methodology for commerce policy by UNCTAD.

For agriculture Commodity Classification is established and maintained in the Website by FAO.

For education ISCED was established 1970 and revised in 1997 and 2011 by UNESCO.
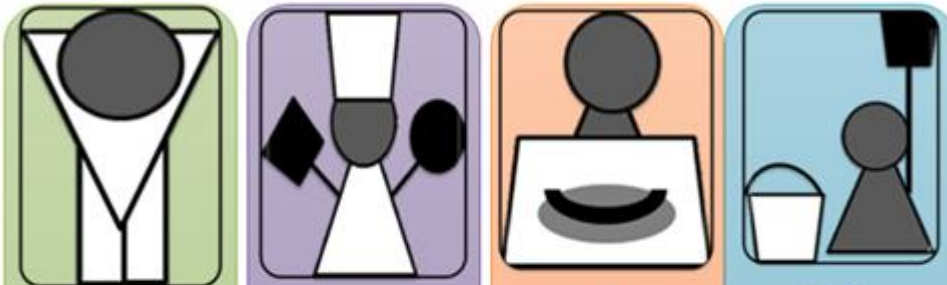
For the purpose of crime statistics ICCS was established in 2015 by UNODC.

For the facilitation of statistical activities by international institutes Classification of Statistical Activities was established by ECE in 2014.

4.12 Coding on ISCO Example: Workers in a Burger Shop

Notes:

In order to make your picture of classifications clearer, an example is shown on the slide.
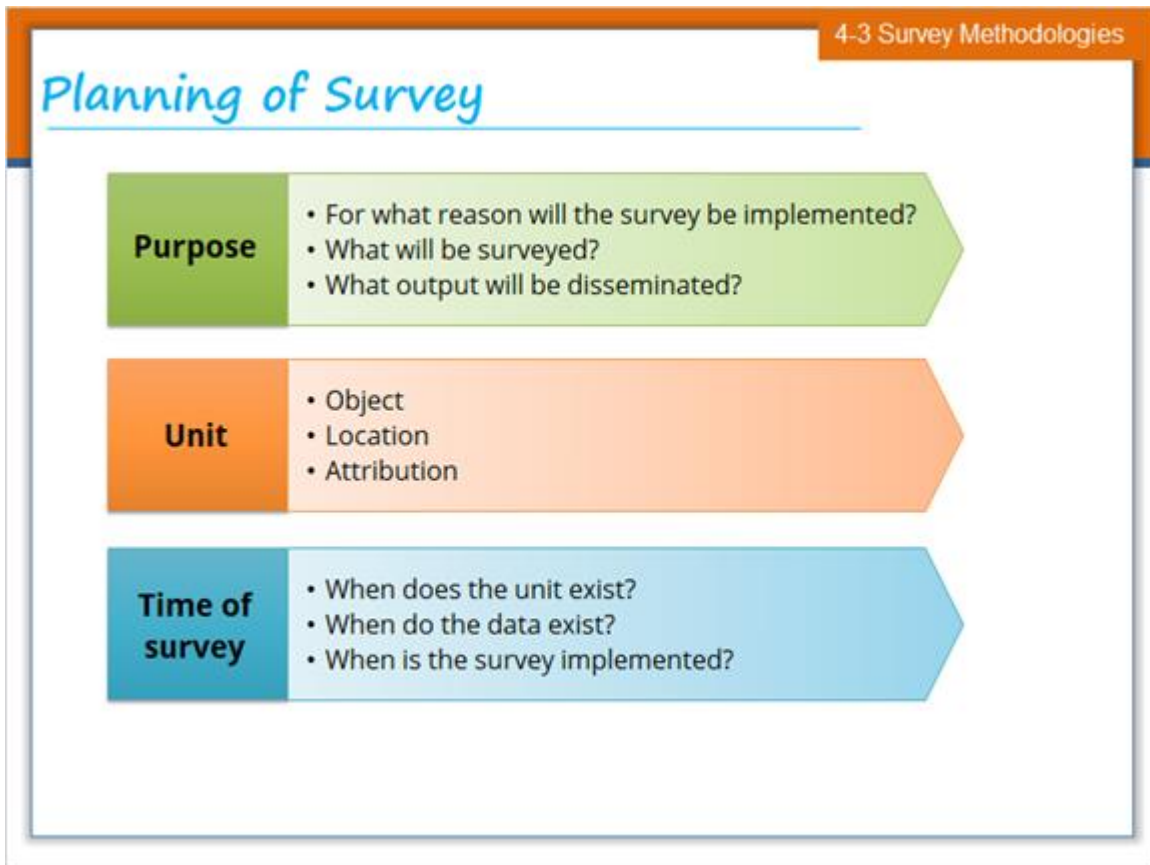
These four people are working in the same hamburger shop, and they are classified in different occupations based on ISCO.

4.13 4-3 Survey Methodologies
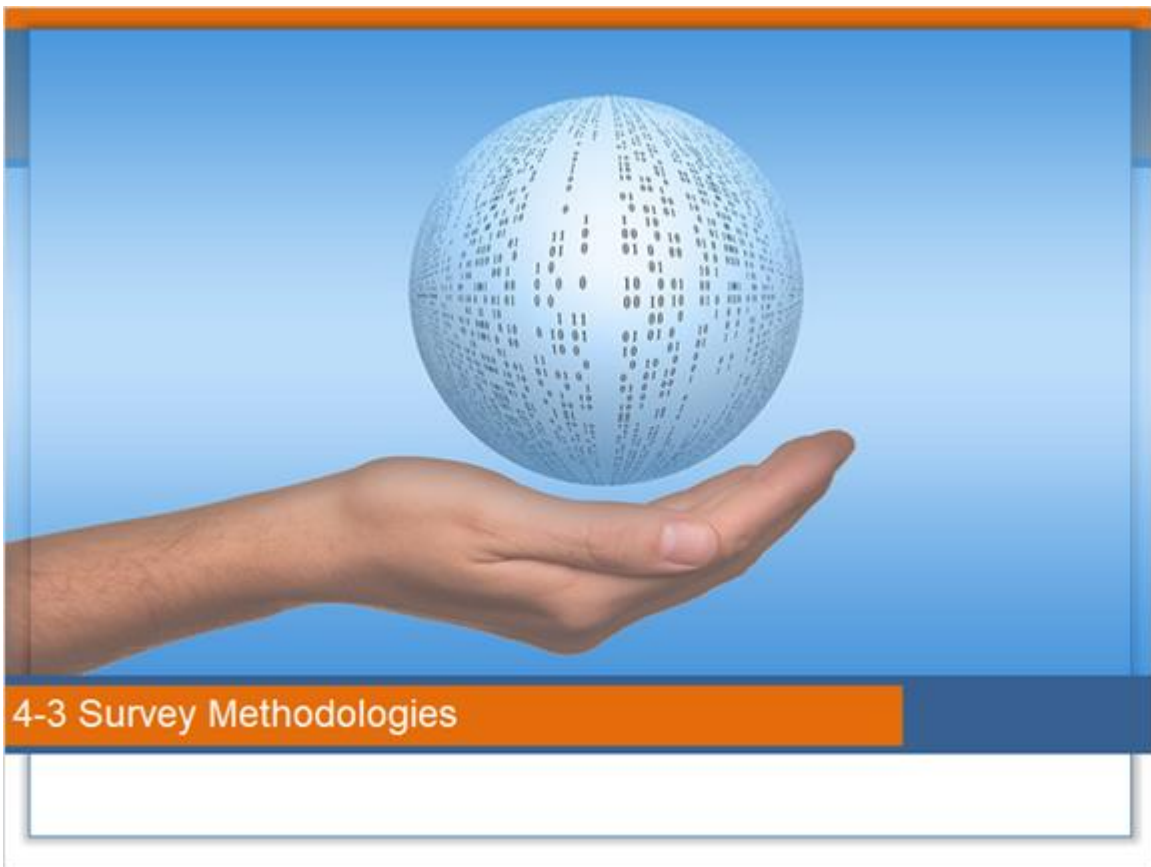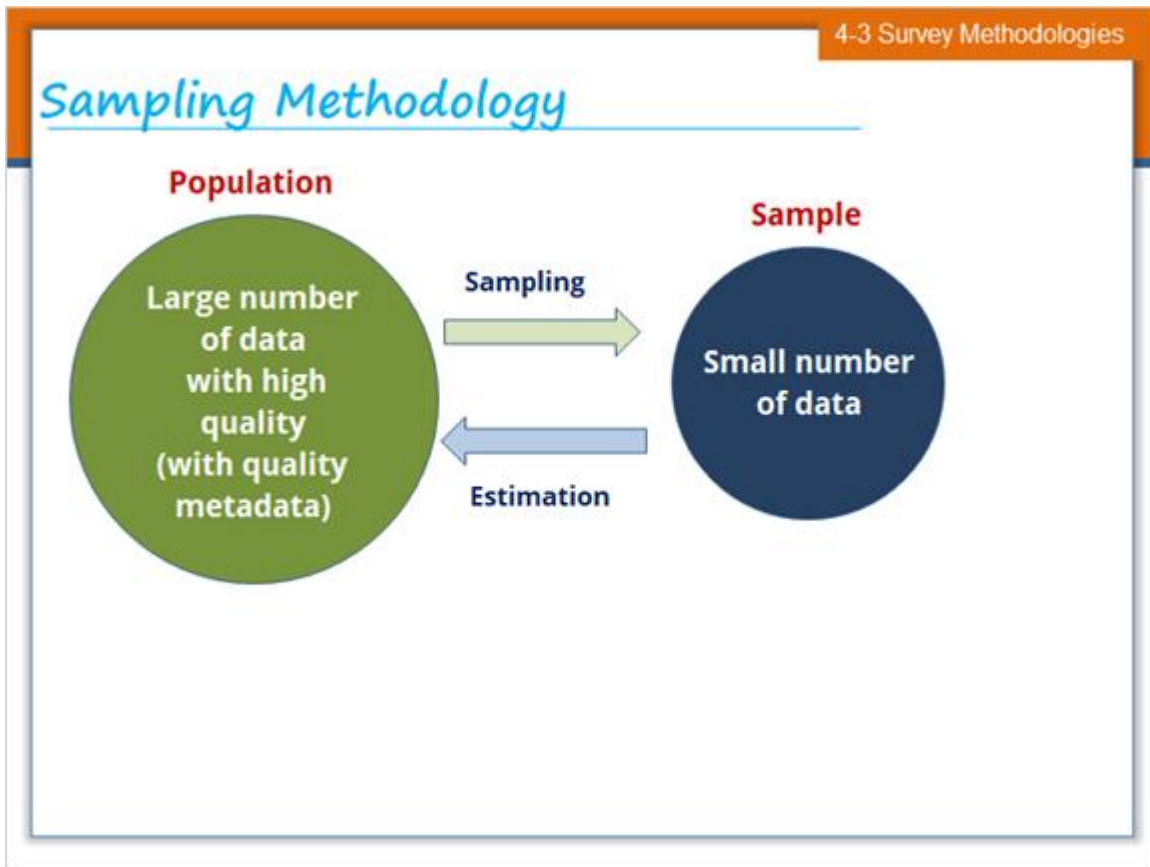

Notes:

4.14 Planning of Survey

Notes:

Core factors of survey methodologies are introduced on this slide.

For the planning of a survey, three factors are involved: purpose, unit and time should be considered regardless of census or sample survey. There are three kinds of time. For instance, in the Population Census, the target population would be the people staying in the country for several months. The information in the questionnaire will be reported as of a designated day and time. The reporting period will be designated, too.
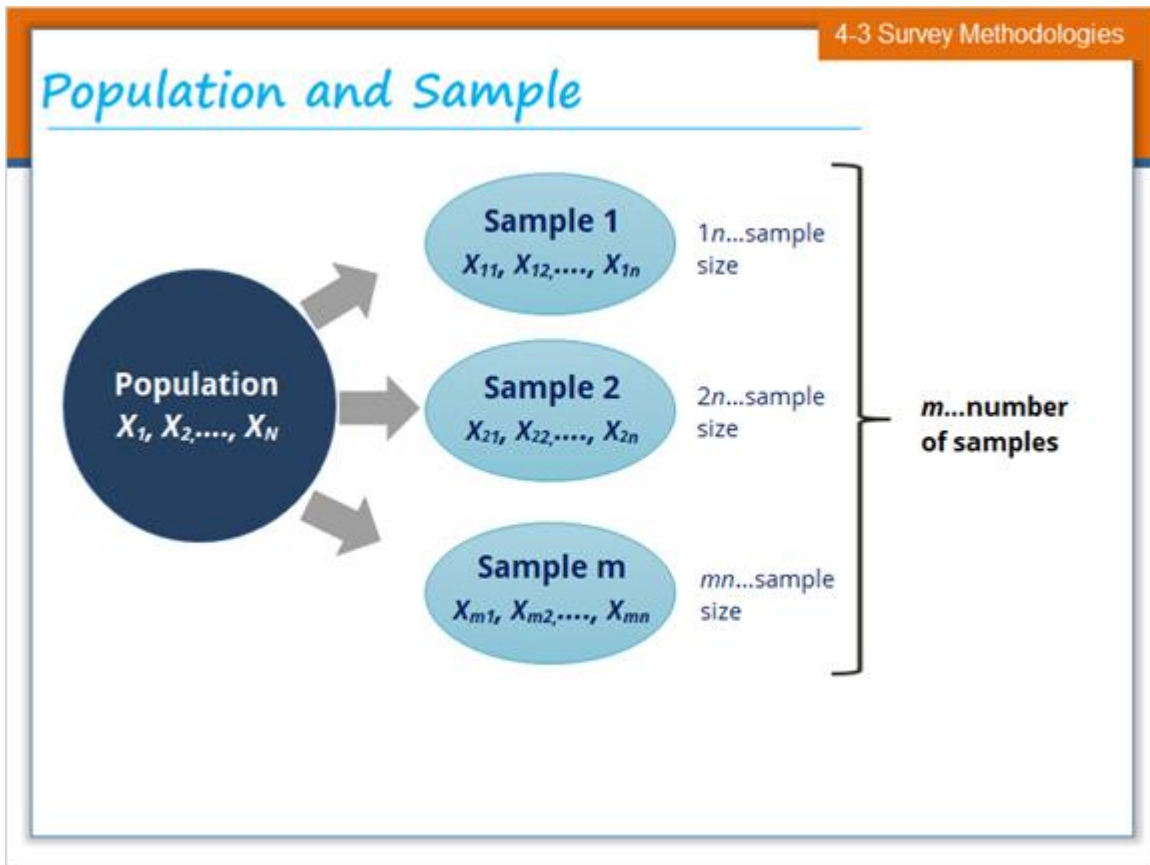
4.15 Sampling Methodology

Notes:

Although the purpose of capturing information is important, it may take a huge amount of time and costs to survey all. In order to carry out the survey with efficiency, a small number of units is often selected for the survey. The process of selection is called sampling. The group of units is called a sample while all original units are called the population. The information of the population is estimated from the sample.

4.16 Population and Sample

Notes:

The methodology to select units are multiple. The **number of sampling** (selection) is the **number of samples**. On the other hand, the **number of units** of the sample is called the **sample size**. In practice, sampling is carried out once. Nevertheless, the number of units in the sample is not called the number of samples but the sample size.

4.17 Sampling in Practice
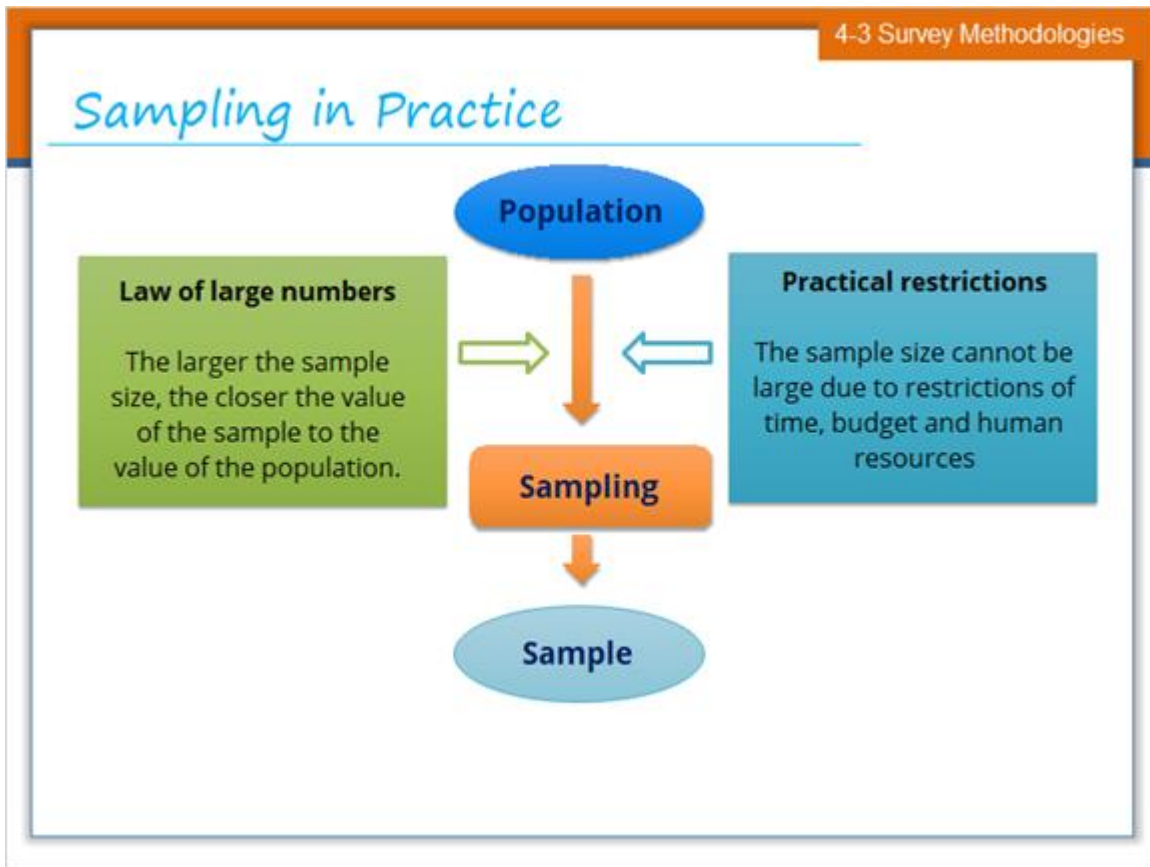
Notes:

Based on the law of large numbers, the larger the sample size, the closer the value of the sample to the value of the population. On the other hand, the sample size cannot be large due to restrictions of time, budget and human resources. With such a trade-off, statisticians need to devise a methodology of sampling to achieve the utmost accuracy under a small sample size.

4.18 How to Create a Population
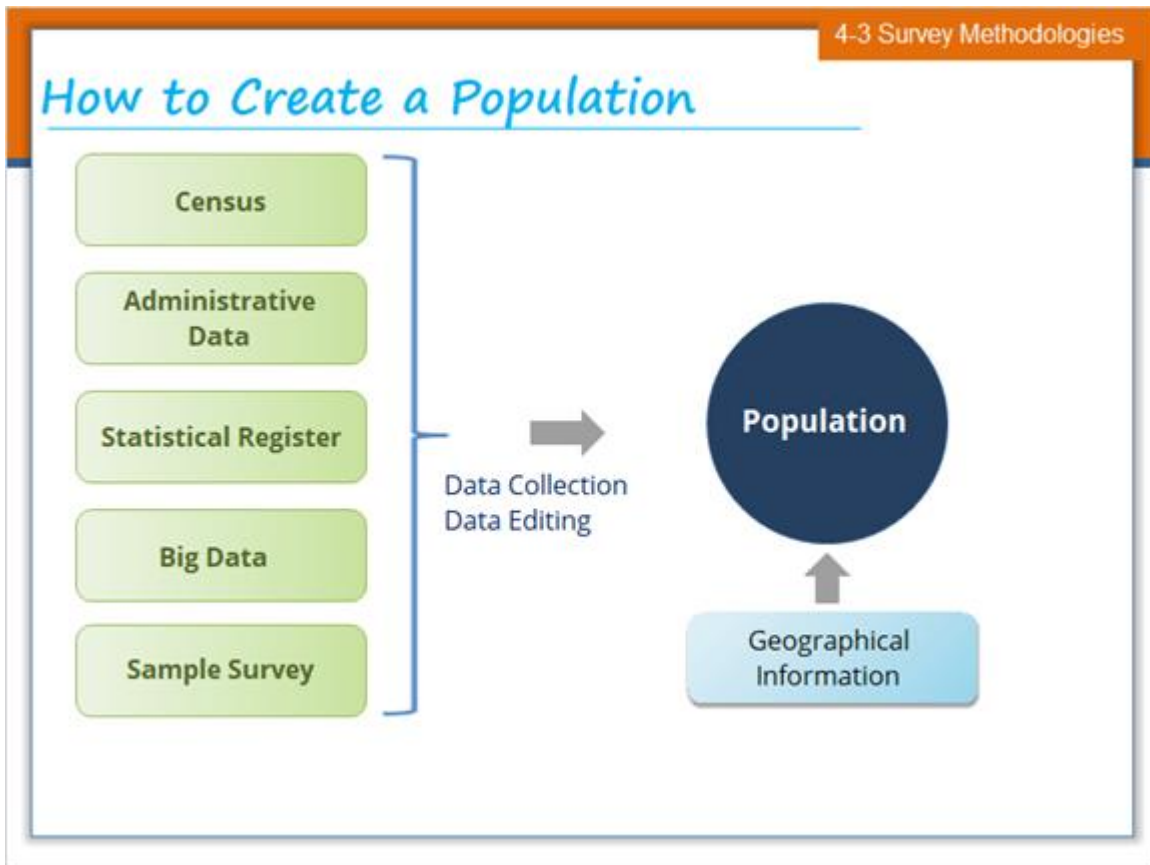
Notes:

Data of the population are collected from survey data, administrative data or Big Data. They are connected such as in a statistical register. As shown in Lesson 3, these data should be of high quality. Some data are so crude that they must be edited before comprising the population. These data are often matched with geographical information such as maps, longitude and latitude.

Data collection and editing will be explained later in this lesson.

4.19 Using Information of Census
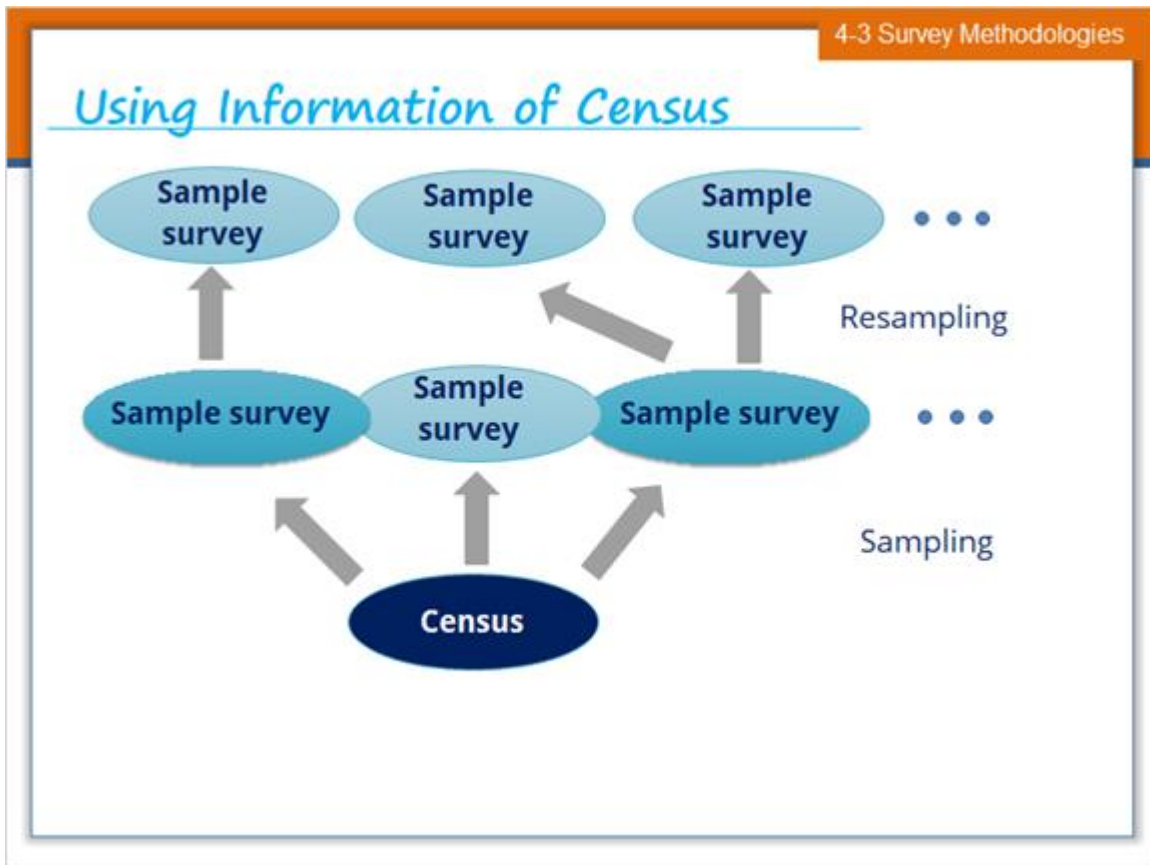
Notes:

Censuses provide the information necessary for strict decision-making, and the values that serve as the standards for other surveys is called a "Benchmark", which serves as the foundation for the design of sampling surveys and information such as the list of respondents and enumeration districts for sampling surveys. Some sample surveys with large sample sizes are sometimes used as the population for sample surveys with smaller sample sizes.

4.20 Purposive Sampling and Random Sampling

Notes:

Sampling is like tasting a cup of soup for the purpose of evaluating the whole soup. The cook can evaluate the whole soup if he/she has it but wants to keep it. One methodology is to pick and eat the part he/she favors. If the part is favored by customers, the methodology would be effective. Nevertheless, customers are different and would evaluate the soup from various aspects in most cases. Mixing the soup is also a good alternative tactic before tasting a cup of soup if he/she uses it as the miniature of the whole soup. The first methodology (picking up the favorite part) is purposive sampling while the latter methodology (mixing the soup) is random sampling. Although purposive sampling has some advantages, random sampling can satisfy the essential quality of statistics based on science. Therefore, for official statistics, random sampling is applied in sample surveys.

4.21 How to Select Individuals
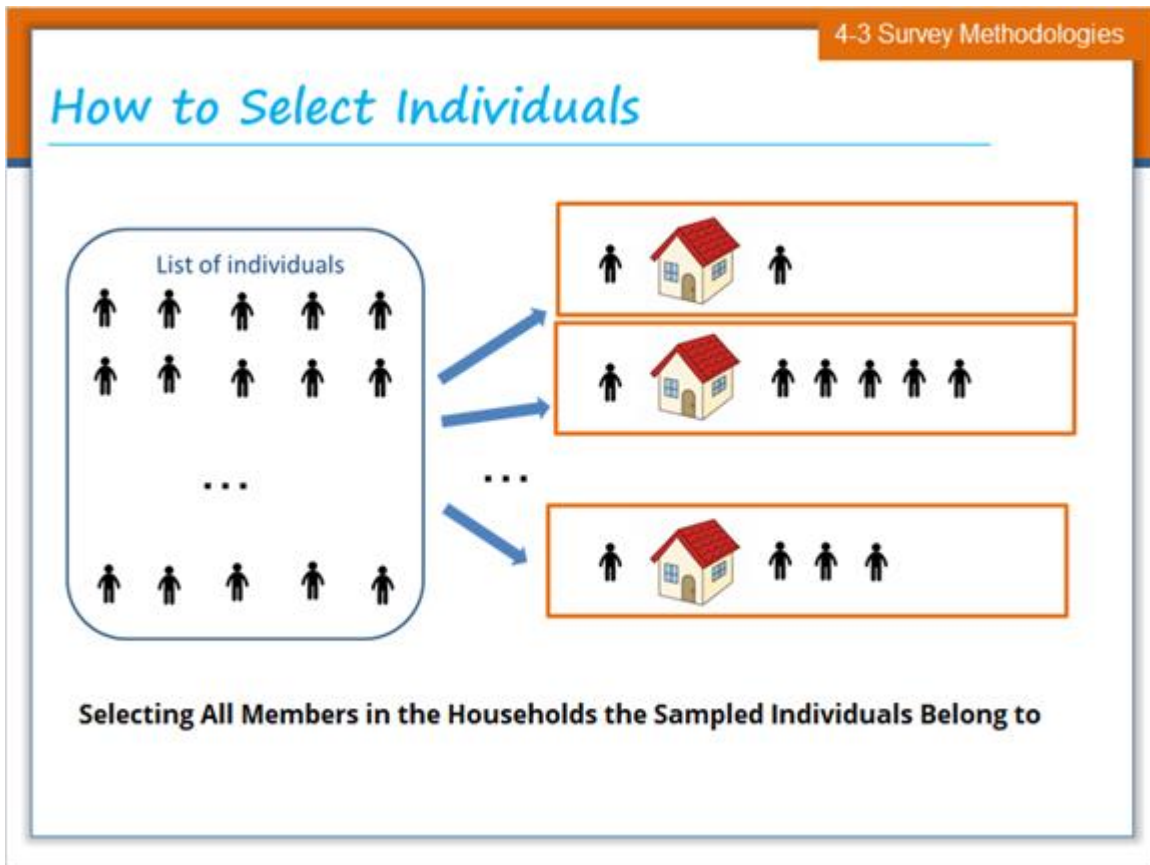
Notes:

Although random sampling is selected following a theorem, selecting individuals at random needs devices in practice. For example, a survey is designed to gain information on employment, time use or expenditure from individuals. Even if the lists of individuals exist, visiting or contacting individuals separately would need time and costs. Households can be units for the surveys instead of individuals. If a household is selected, all members of the household are selected. Given the lists of individuals exist, firstly individuals are selected randomly, then households to which these selected individuals belong are selected, and all members of the selected households are selected for the survey. In case lists of individuals do not exist, households themselves would be selected randomly and all members of those households would be selected. In this case, all households are selected with the same probability regardless of household size (number of household members). This means that single household members are apt to be selected more than multiple household members.

4.22 How to Select Trucks
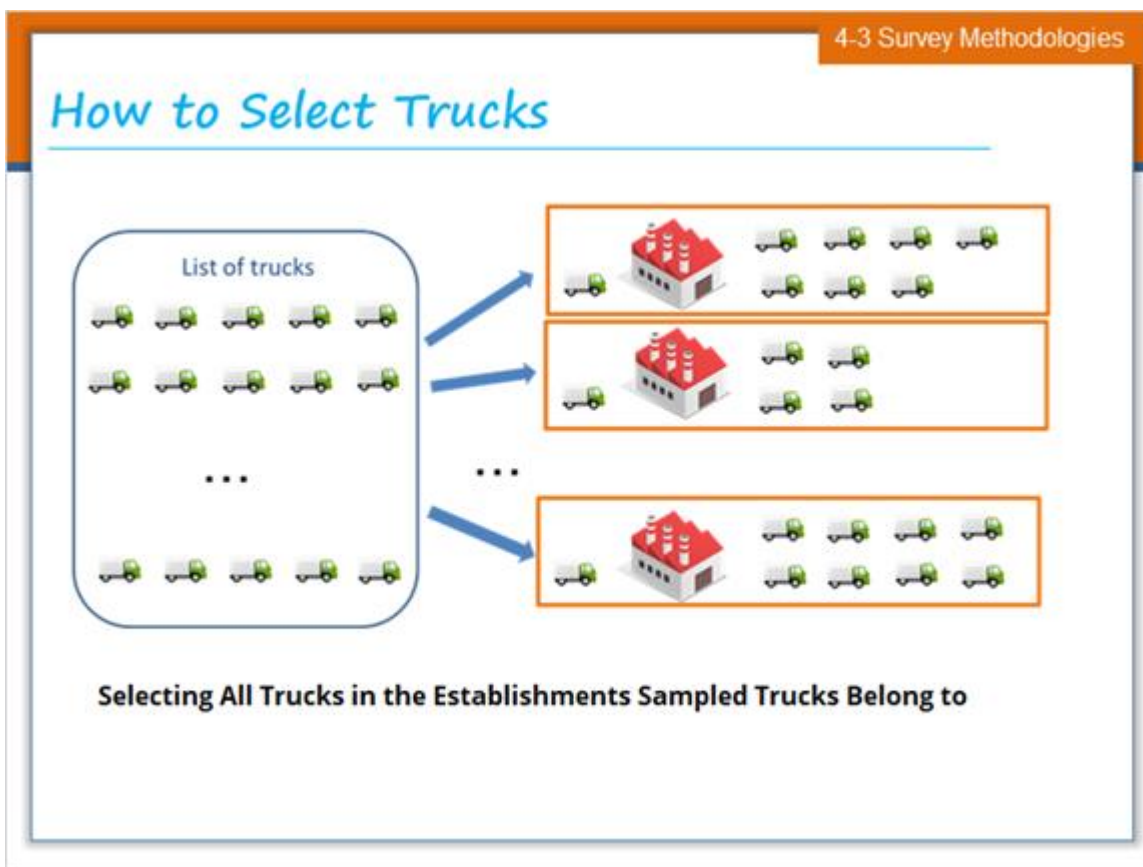
Notes:

Similar methodologies can be applied to establishments. For example, a survey is designed to gain information on miles or tons of loads of trucks. Firstly trucks are selected randomly, then establishments to which the selected trucks belong are selected, and all trucks of the selected establishments are selected for the survey. Lack of the lists of trucks would cause more serious issues than selecting individuals from households because the number of trucks in establishments are more diverse than the number of household members. Therefore, random samplings of establishments should be avoided. Additionally, establishments with many trucks should be selected with higher probability.

Similarly, if the unit of surveys is employee, it is important to gain lists of employees. Instead, establishments with many employees should be selected with higher probability as well.

4.23 Estimation of Ratio
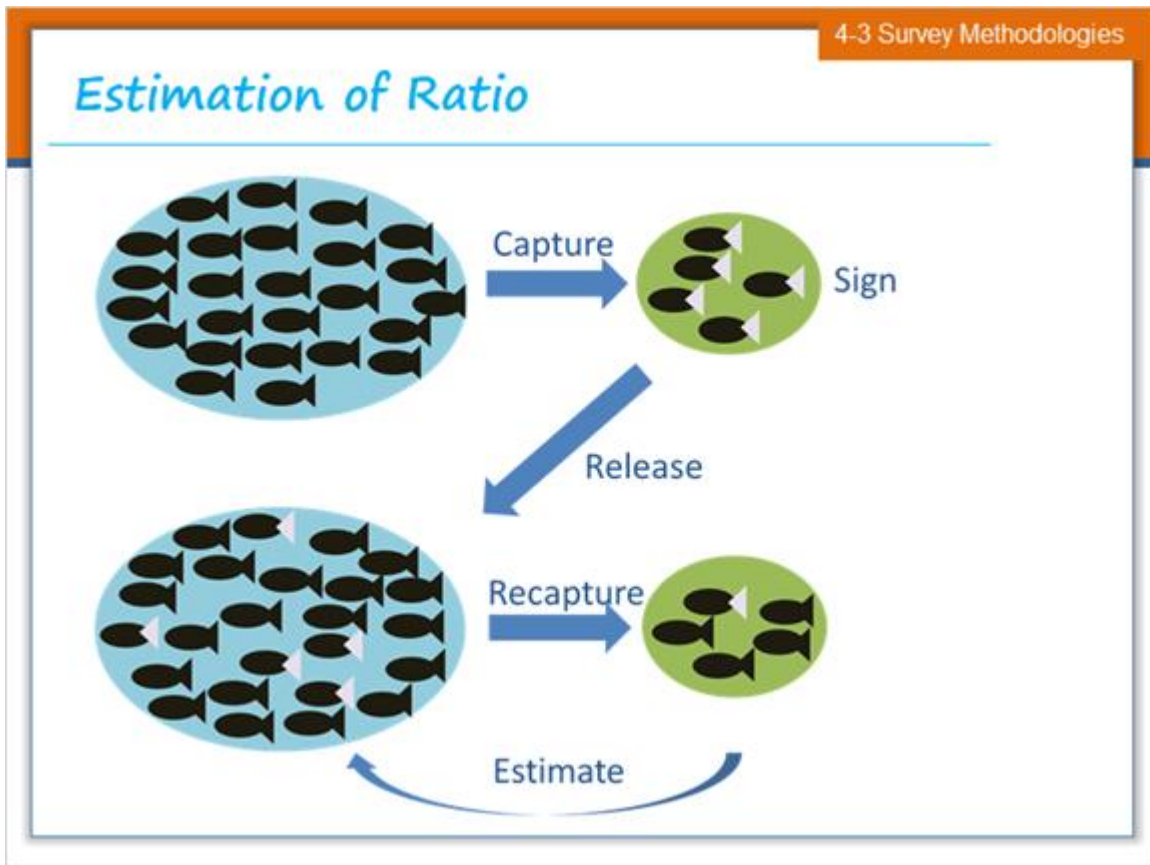
Notes:

Estimation is based on the methodology to multiply the inverse of the rate of selection to result from the sample.

The slide shows a simple case: the capture-recapture methodology. When the number of fish in a pond is unknown, firstly catch five fish from the pond. Those five fish are signed and released back into the pond. Next five fish are captured again. If the signed fish is only one, the number of fish in the pond is estimated as five times five = 25.

Different from this case, for surveys in society and economy, the population is much clearer so that only the second (recapture) stage of the capture-recapture methodology reflects on the sampling design. For example, the unemployment population is estimated by multiplying the inverse of the rate of selection to the unemployment in the sample.

4.24 Process of Design of Sampling
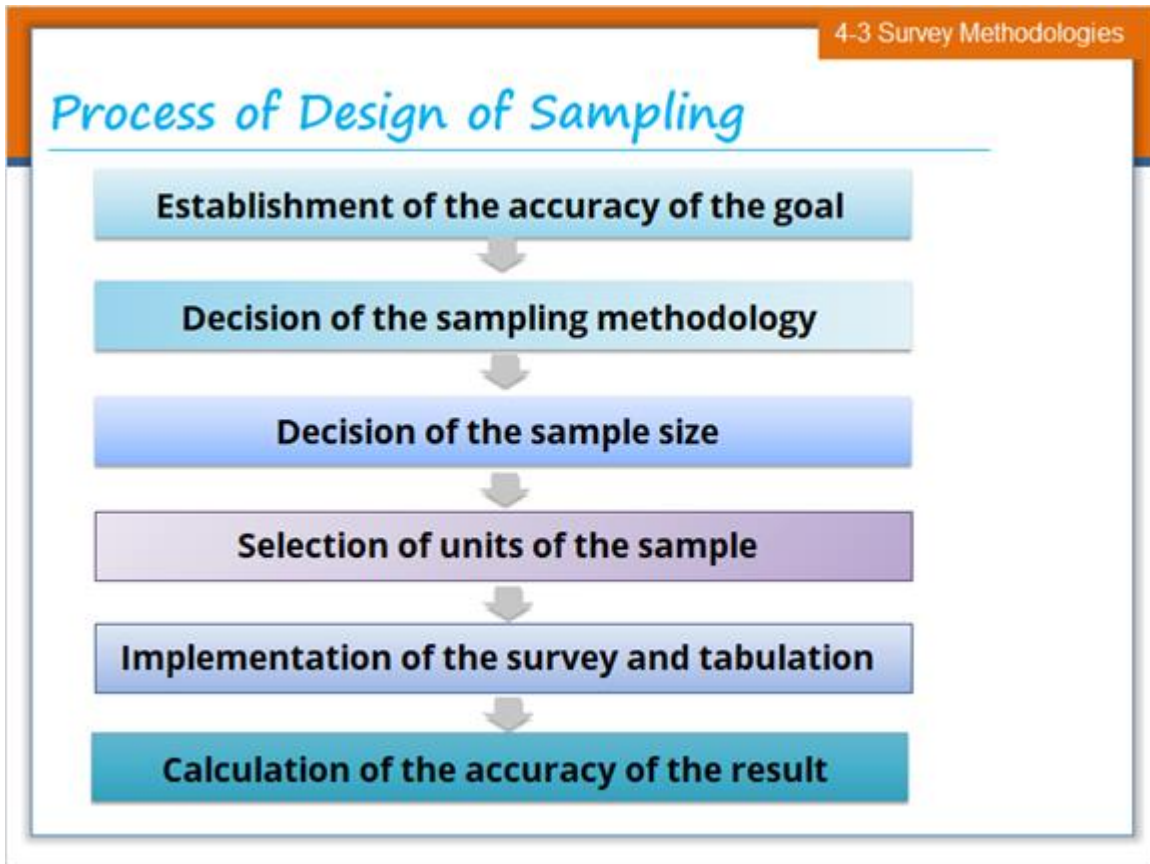
Notes:

For sampling surveys, the process of design is on the slide. Firstly, the accuracy of the survey is decided as a goal. If a sampling methodology is decided, the sample size can be calculated. The sample size means the number of units. During the survey, some units do not respond to the request to fill out the questionnaire. Thus the accuracy will be changed based on the number of units after the implementation of the survey.

Eurostat published Survey Sampling Reference Guidelines which include survey planning, survey quality, sampling and estimation, and nonresponse in 2008.

4.25 Doubtful Surveys
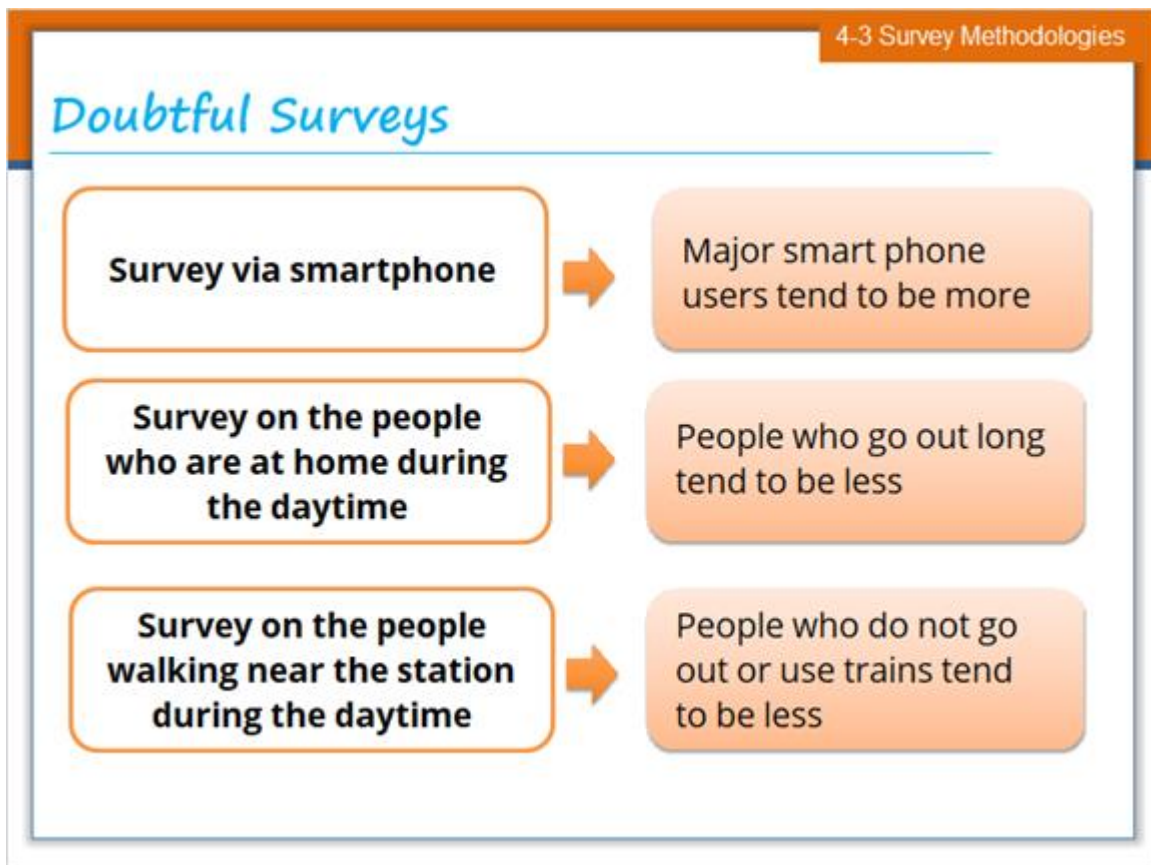
Notes:

Biases can happen even in the census. Typical mistakes in the design of the survey are in the slide. It is important to design surveys impartially so as to avoid biases.

4.26 4-4 Statistical Registers

4-4 Statistical Registers

Notes:

4.27 Why is Counting Births and Deaths Important?

Notes:

As mentioned in Lesson 1, statistical register is one of the significant data sources of official statistics.

This lesson elaborates on more specific matters quoting practices related to statistical register.

Civil registration is the way by which countries keep a continuous and complete record of births, deaths and the marital status of their people.

Countries need to know how many people are born and die each year - and the main causes of their deaths - in order to have well-functioning health systems. The only way to count everyone and to track all births and deaths is through civil registration. Civil registration provides the basis for individual legal identity but also allows countries to identify their most pressing health issues.

WHO receives cause-of-death statistics regularly from about 100 Member States.

However, globally, two-thirds (38 million) of 56 million annual deaths are still not registered and every year, almost half of the world's children go unregistered.

When deaths go uncounted and the causes of death are not documented, governments cannot design effective public health policies or measure their impact. Civil registration is something that all developed countries have, and that developing countries need. Information on births and deaths by age, sex and cause is the cornerstone of public health planning.

Civil registration brings multiple benefits. An individual's right to be counted at both extremes of life is fundamental to social inclusion. In the absence of insurance or inheritance, death registration and certification are often required prerequisites for burial, remarriage, or the resolution of criminal cases. There are risks associated with civil registration. The information it provides can be used to discriminate against certain groups, however there are ways to design systems to reduce these risks.

## 4.28 Example of Civil Registration in Australia
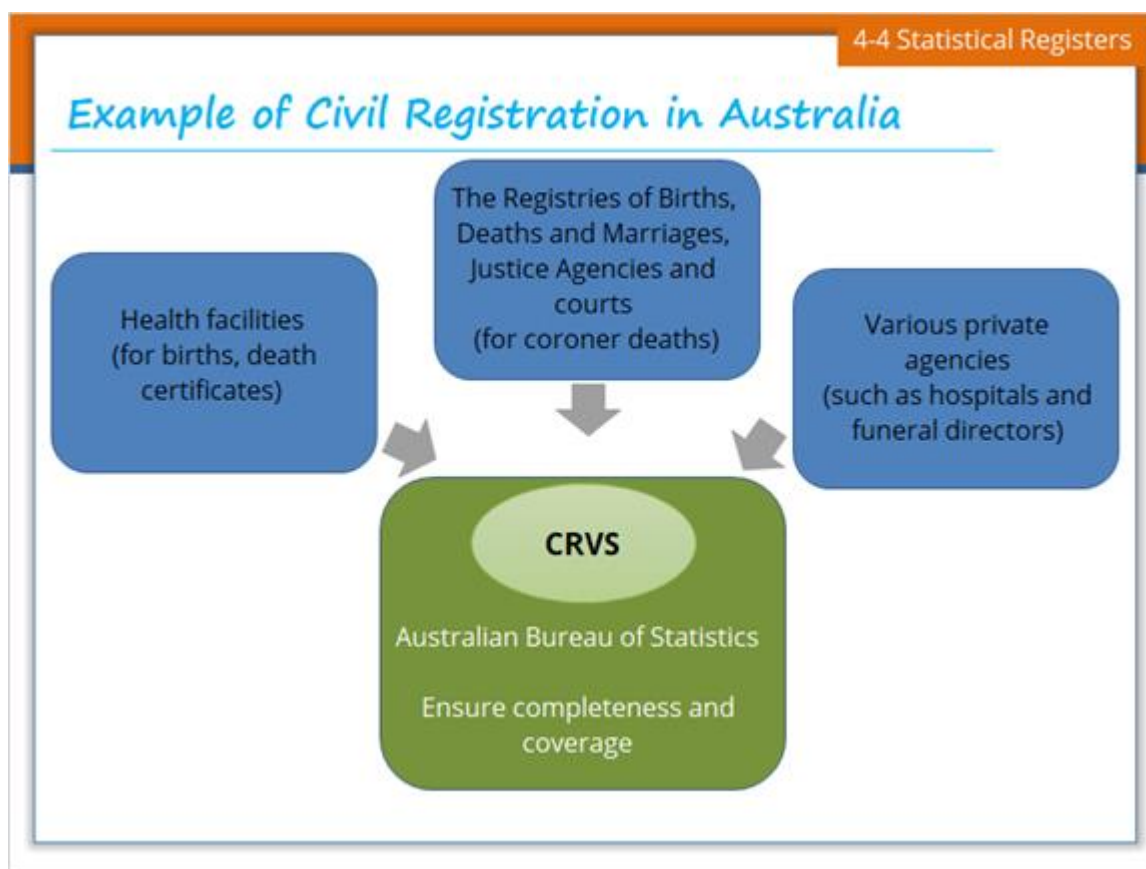
Notes:

In Australia, there is strong coordination between the different actors involved in CRVS, based on interagency cooperation between people and agencies to achieve good quality data on births, deaths and cause of death (COD). These include health facilities (for births, death certificates), the Registries of Births, Deaths and Marriages, Justice Agencies and courts (for coroner deaths), and various private agencies (such as hospitals and funeral directors). The Australian Bureau of Statistics serves as the point of consolidation of data from these different sources, and to ensure completeness and coverage. In Australia, health information is treated as an essential component of the civil registry (CR) process, and birth notifications and medical certification of death must be received before a birth or death can be registered. In particular, this creates a very strong incentive for deaths to be properly certified.

The registration of deaths is the responsibility of the eight states and territory Registrars of Births, Deaths and Marriages, who receive the confirmation of a birth or death from

the health system. Information on COD is supplied by the medical practitioner or by a coroner to the registrar, who then forwards it to the Bureau of Statistics for coding, compilation and supplementation with information from the National Coroners Information System. Data quality assurance is provided at every point in the production of statistics, including receipt from registries, editing, coding and publication. The key experiences from Australia can be summarized as follows:

• Health data are treated as an essential component of CR.

• The Bureau of Statistics plays a key role to coordinate information collection, coding and publication and ensure coverage and quality.

• There exists a widespread involvement of various actors of the CRVS systems, and their systems of coordination.

• There are strong processes in place at different stages of the information flow to strengthen data quality.

• The system is especially strong in bringing in COD information, which remains historically weak in most national CRVS systems.

4.29 How to Create a Statistical Business Register
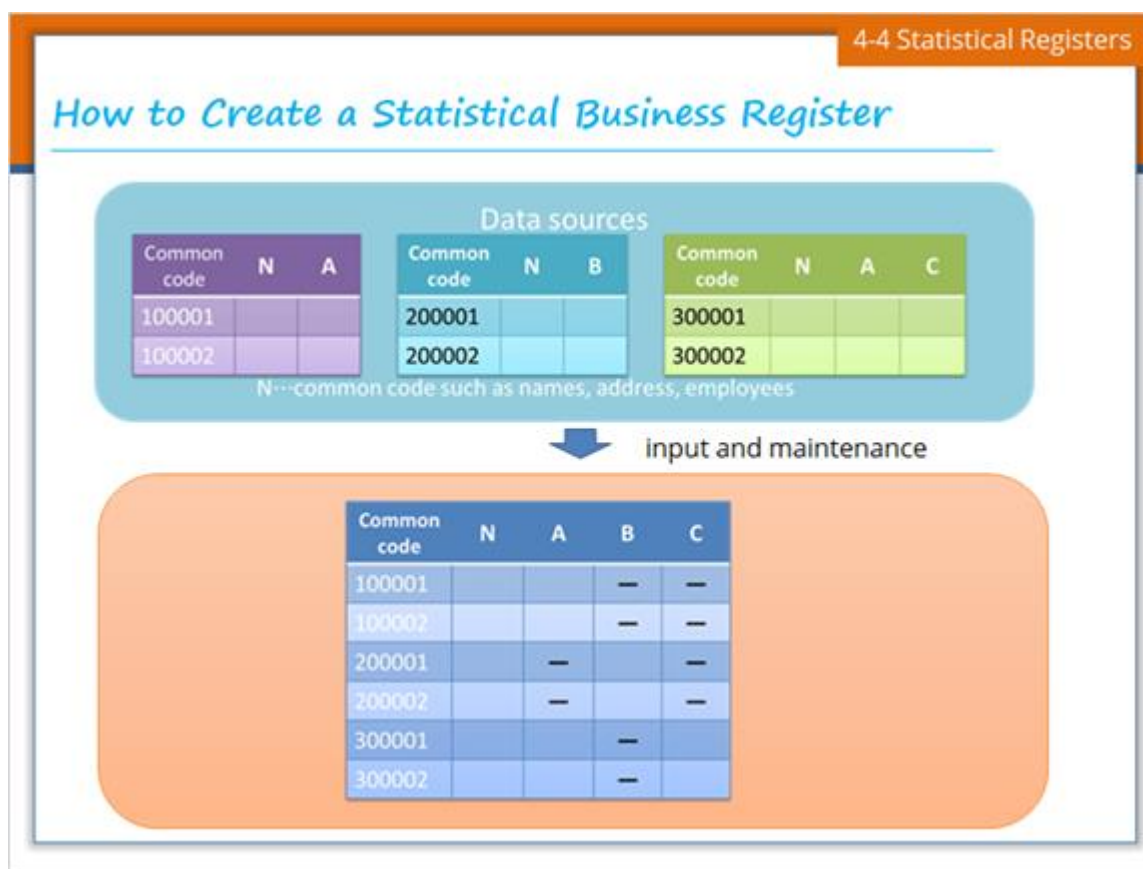
Notes:

The demand for economic statistics is growing in all countries as well as at an international level. To a large extent this is the result of the fact that economies have become more complex, with increasing division of labour both at national and at global levels. To remain relevant and to meet user needs, economic statistics should, therefore, cover new areas and should provide more detailed information about activities and variables. There is also a general need to improve timeliness, coherence and reliability at both national and international levels.

While business statistics should achieve these new and enlarged goals, it should do so in an efficient way and with a response burden as low as possible. This means efforts must be made to modernize the production processes, to use administrative data to reduce survey burden, to redesign survey systems, to harmonize surveys and variables, to comprehensively link administrative data and survey sources, and to cooperate more closely with administrative authorities on a continuous basis to improve the quality of

the Statistical Business Register. Better consistency between the various statistical areas at national and international levels has also become a much more important goal than it was in the past, when economic development was less effected by international dependencies, and the globalization of production and markets was at a lower level.

In many countries, especially in the EU, Statistical Business Registers have been developed for statistics on corporations and establishments.

A Statistical Business Register is a regularly updated, structured database of economic units in a territorial area, maintained by an NSO, and used for statistical purposes. In this definition the terms have the following meanings.

This is a complex property. A Statistical Business Register is a structured database, where defined characteristics are stored for each economic unit. In circumstances where there is more than one kind of economic unit the relationships between these units are also included. For example, for each local unit there is a link to the enterprise to which this unit belongs; for each legal unit there is a link to which enterprise it is associated. Economic units are assigned unique numerical identifiers so that they are easily identified and to make sure that no units are included twice.

Economic units comprise legal/administrative units and statistical units. Legal/administrative units are units registered in administrative registers, such as taxation registers, social security register, company register, register of the chamber of commerce, etc., including also non-market units such as government departments and non-profit institutions. Legal/administrative units serve as the basis for delineation of statistical units, by which they are represented for statistical purposes. Examples of statistical units are the enterprise, the local unit and the establishment.
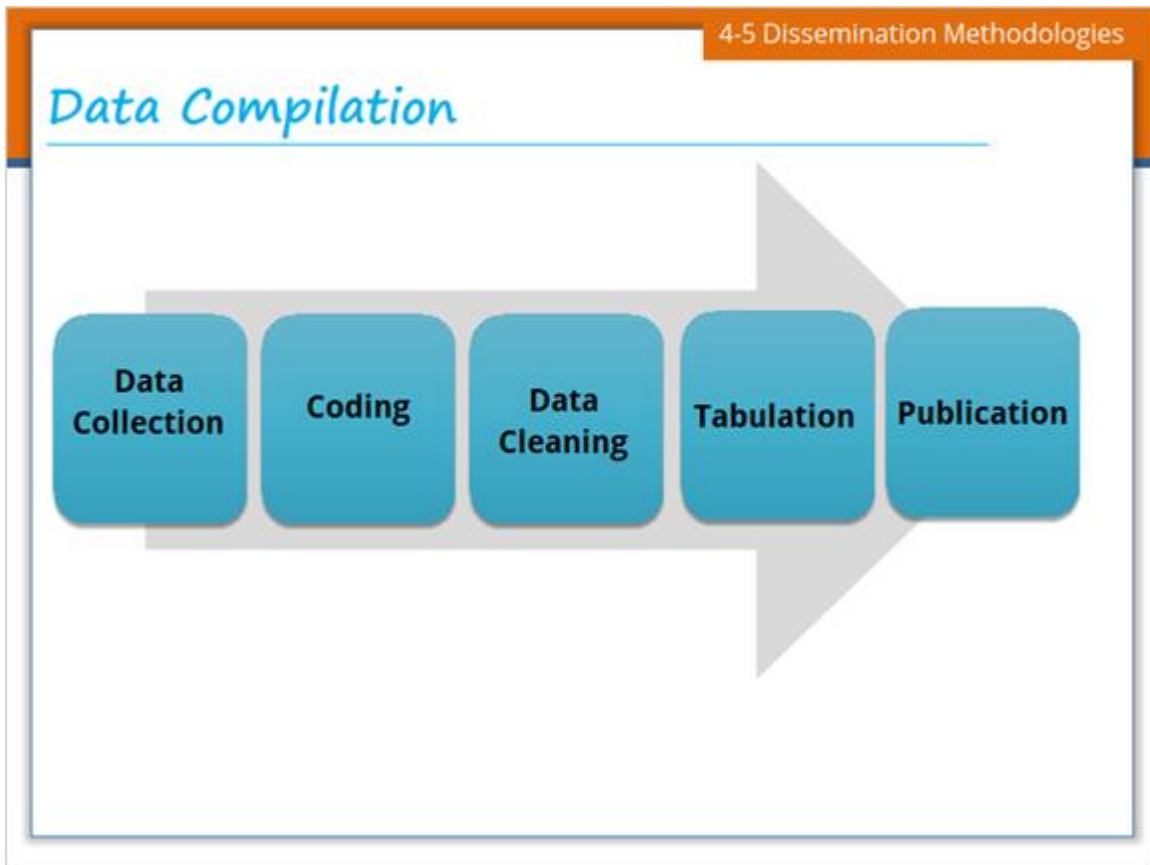
Data are recorded by unit with its registration number from various sources like in the slide.

4.30 4-5 Dissemination Methodologies

Notes:

4.31 Data Compilation

Notes:

This is the typical flowchart for the dissemination of statistics after data are collected through surveys, administrative procedures and from statistical registers.

Data need to be coded and cleaned for the purpose of dissemination of statistics. Code of data follows classifications, favorably the international classifications to secure international comparison. Data cleaning includes data editing which is the activity aimed at detecting and correcting errors (logical inconsistencies) in data. These processes are elaborated in Phase 5: Process, Lesson 2-2 Generic Statistical Business Process Model. The data are tabulated and published based on the classifications.

4.32 Functions of Generic Statistical Data Editing Models (GSDEM)

Functions of Generic Statistical Data Editing Models (GSDEM)

- Review of data validity (by checking combinations of values)
- Review of data plausibility (by analysis)
- Review of units
- Selection of units
- Selection of variables
- Variable amendment
- Unit amendment

Notes:

For data editing, generic statistical data editing models exist.

The descriptions of the function categories are as follows:

Review of data validity (by checking combinations of values): Functions that check the validity of the combination of data values against a specified range or a set of values and also the validity of specified combinations of values. Each check leads to a binary value (TRUE, FALSE).

Review of data plausibility (by analysis): Functions that calculate measures for the plausibility of data values in a data set. It results in quantitative measures that can be used to evaluate the plausibility of data values, which may include aggregates. This also includes less formally specified "functions" such as analysis by inspection of graphical displays.

Review of units: Functions that calculate scores that provide quality measures for

making a selection of a record. A score function can be whatever measure which describes a unit. The outcome of a score function is often needed for further use in the next phase of the process step in which the score function is.

Selection of units: Functions that select units from a data set for separate processing. Automatic selection appears e.g. when values of score functions are compared with a predefined threshold value. Correspondingly, manual selection is usually based on macro-editing, e.g. with aggregates and graphics.

Selection of variables: Functions that point out variables in units for a different treatment than the remaining variable, usually referring to their observed errors. As for units, this operation can be done either manually or automatically.

Variable amendment: Functions that alter observed values or fill in missing values in order to improve data quality. Usually the amendment functions are dedicated to correcting different error types (e.g. systematic errors, errors in unit properties). The functions may lead to method solutions that are conducted automatically or manually.

Unit amendment: Functions that alter the structure of the unit by combining and reconciling the different units residing in multiple input sources. The aim is to derive and to edit the target statistical units that are not given in advance.

4.33 Methodology of Expression of Results by ECE

Methodology of Expression of Results by ECE

Guidelines for Statistical Metadata on the Internet (2000)

- Standards of statistical metadata on the Website

Communicating with the Media: A guide for Statistical Organizations (2004)

- Guide for effective communication with the media and the people to countries

Making Data Meaningful (2009-2012)

- A guide to writing stories about numbers
- A guide to presenting statistics
- A guide to communicating with the media
- A guide to improving statistical literacy

Notes:

Metadata should assist the user in searching for statistical information, interpreting its content and, after its potential downloading from the Internet, they should help in post-processing statistical applications.

The main objectives of the **Guidelines for Statistical Metadata on the Internet** are the following:

I. To support the world-wide dissemination of statistical data on the Internet;

II. To promote a consistent interpretation of statistics from different sources by considering data quality and international comparability as strategic issues;

III. To promote proper usage and processing of public statistics.

Communicating with the Media is intended as a practical guide to assist countries that are setting up a NSO to communicate effectively with the media, and in turn, with the general public.

Making data Meaningful is intended as a practical tool to help managers, statisticians and media relations officers in statistical organizations, particularly those organizations that are in the process of developing their communication strategies. The guides provide advice on the use of text, tables, charts, maps and other devices to bring statistics to life for non-statisticians.

4.34 Guidelines for Statistical Metadata on the Internet

4-5 Dissemination Methodologies

Guidelines for Statistical Metadata on the Internet

Some guidelines for minimum standards for statistical metadata on the Internet

Minimum set of metadata required for the correct interpretation of statistics:
- Title/content description
- Labels for rows/columns in tables and elements of graph
- Definitions of labels
- Measurement unit
- Time reference/period
- Regional units
- Comparability over time
- Footnotes highlighting specific precautions
- Source of the data
- Explanation of standard symbols in tables
- Any information on copyright, restrictions of usage
- Contact points for additional information

Notes:

The requirement to supply information for interpretation, to some extent, depends on the subject area, the type of information provided and the target user groups for any specific service. It is obvious that the needs may vary substantially between general users, media, researchers, etc. However, following the general objectives, and taking into account the needs of users of all skill levels, as much information as necessary to

make correct interpretations and avoid misuse should be made available. It may be important to supply information on the level of comparability with data from different sources. A further challenge is to link data/metadata in such a way that users can choose the option they really need to use and interpret statistical information correctly. It is therefore recommended that for any form of statistical data presented on the Internet (e.g., statistical indicator as a separate figure, graph or table) the meta-information on the slide should be available.

4.35 Guidelines for the Modeling of Statistical Data and Metadata of ECE

Common Metadata Framework

Notes:

The Guidelines for the Modeling of Statistical Data and Metadata of ECE describes a common metadata framework. Parts of the framework have been updated separately.

Metadata provide information on data and about processes of producing and using data.

The aim of this methodological material is to guide NSOs on how to design meta-information systems for statistical survey processing. It recommends ways of metadata and data modelling for individual phases of the processing of statistical surveys.

As for Part C - Metadata and the Statistical Business Process, the Generic Statistical Information Model (GSIM) is a reference framework for statistical information, designed to play an important part in modernizing and streamlining official statistics at both national and international levels. It enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. It provides a set of standardized, consistently described information objects, which are the inputs and outputs in the design and production of statistics. The GSIM helps to explain significant relationships among the entities involved in statistical production, and can be used to guide the development and use of consistent implementation standards or specifications.

4.36 Data and Metadata Reporting and Presentation Handbook by OECD

4-5 Dissemination Methodologies

Data and Metadata Reporting and Presentation Handbook by OECD

1. Introduction and overview
2. Standards for the reporting and presentation of statistics
3. Importance of the use of common terminology
4. Guidelines for the reporting of different types of data
5. Guidelines for the reporting of different forms of time series data
6. Guidelines for the reporting and dissemination of metadata
7. Guidelines on key reporting practices

Notes:

OECD also provides a Data and Metadata Reporting and Presentation Handbook. The handbook includes guidelines for the reporting of different types of data, the reporting of different forms of time series data and the reporting and dissemination of metadata.

The Metadata Common Vocabulary of the handbook specifically aimed at identifying commonly used terms to describe the different types of metadata, intended to be used by international organisations and national statistical agencies, contains a core set of metadata items (for both structural and reference metadata) and their related definitions and is designed to improve the standardization of metadata content for the purposes of data exchange and to promote the use of common nomenclatures that can foster international comparability of international data. The current version contains several fields - term title, definition, source, hyperlink to definition source where available, related terms and context.

## *4.37 Statistical Data and Metadata eXchange (SDMX)*
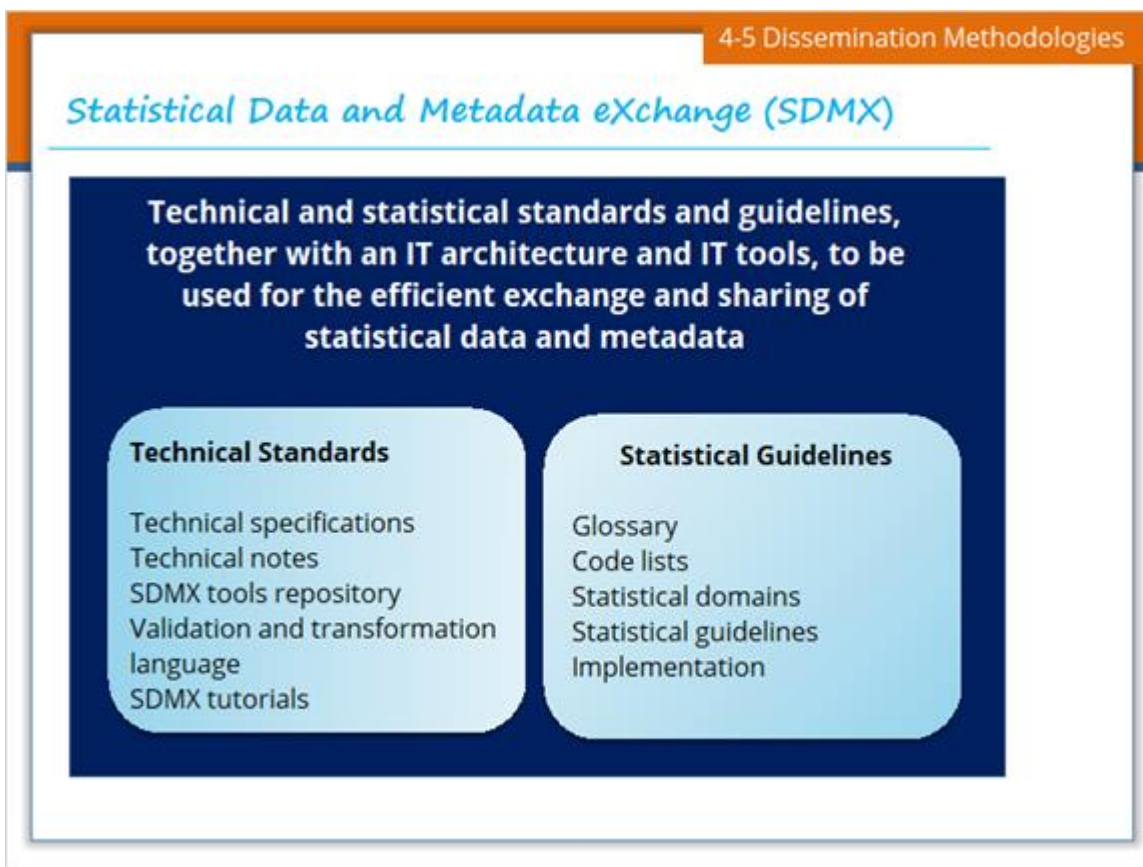
Notes:

Statistical Data and Metadata eXchange (SDMX) consists of technical and statistical standards and guidelines, together with an IT architecture and IT tools, to be used for efficient exchange and sharing of statistical data and metadata. Standardized file formats for data and metadata, and standardized contents of these files are the pre-condition for the automated production, processing and exchange of SDMX data and metadata files between national and international statistical organizations.

Statisticians tend to be concerned with the standards and the model, whilst IT staff tend to be interested in the technical stuff. Where the two worlds collide in SDMX, the aim is to speak the same language so that each understands the needs of the other.

4.38 SDMX Glossary
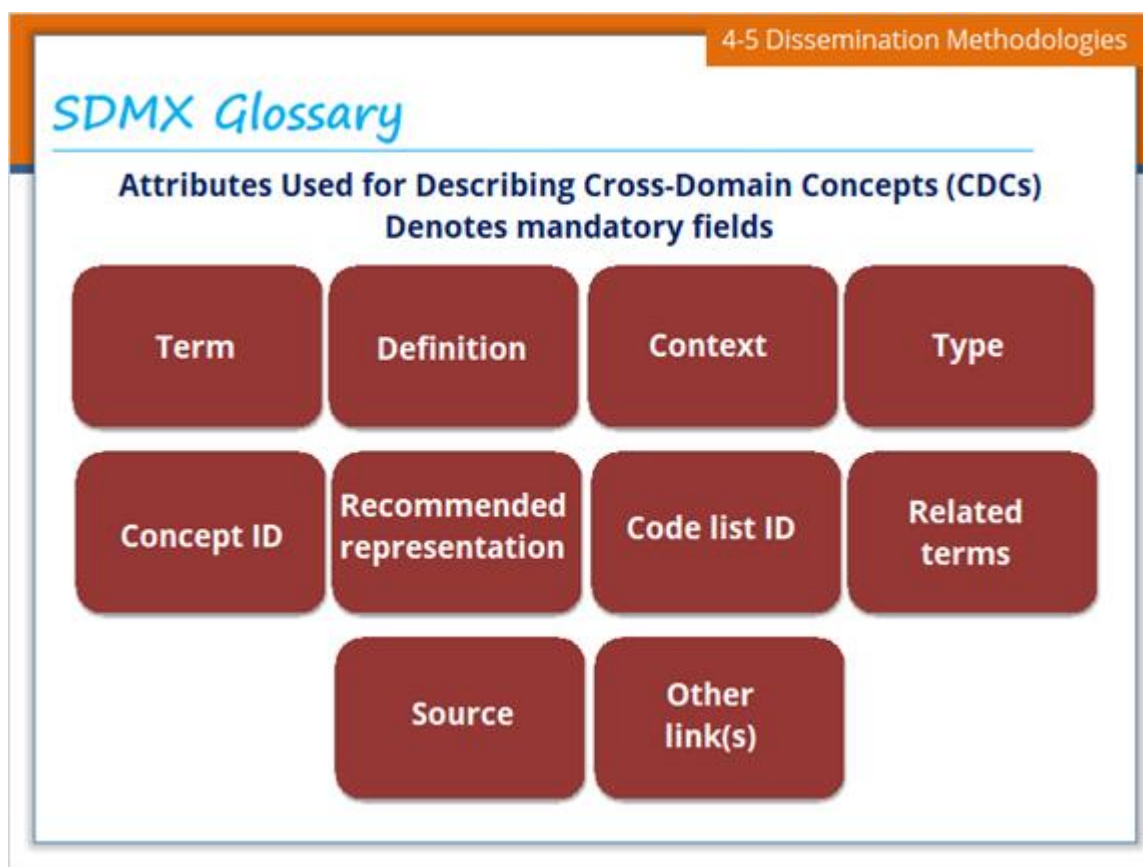
Notes:

The SDMX Glossary is an SDMX guideline containing concepts and related definitions that are useful for building and understanding data and metadata exchange arrangements based on SDMX. The Glossary provides definition of terms found in the SDMX Information Model, Data Structure Definitions (DSDs), and Metadata Structure Definitions (MSDs) at the time of the present release. It is recommended as a single entry point to a common SDMX terminology to be used in order to facilitate communication and understanding of the standard.

In short, the overall message of the glossary is the following: if a term is used, then its precise meaning should correspond to the SDMX Glossary definition, and any reference to a particular phenomenon described in the SDMX Glossary should use the appropriate term.

The glossary is not intended to cover the whole range of statistical terminology, as this area is already covered by other general or domain-specific glossaries. The focus of the

glossary is largely those terms that are normally used for building and understanding metadata systems and SDMX data exchange arrangements. Capitalized concepts are concepts which are part of the SDMX Information Model.

Term is the name of the concept. The term should preferably be entered in the singular form and upper cases should be avoided to the largest extent possible.

Definition is a short statement explaining the meaning of the concept. This textual description of the concept should answer the question "What is it?" rather than "How is it done?" or "Why do we have it?, etc. It is recommended to keep definitions short and add any explanatory text under the field "Context".

Context is the complementary information on the background, history, use, status, etc. of the concept. This field is used to add information on how and where the term may be used. It describes SDMX use cases for the term and may contain examples of its use. This field is optional, though strongly recommended.

Type is used to explicitly denote concepts that are cross-domain.

Concept ID is a unique identifier for the concept that allows it to be unambiguously used for machine-to-machine exchange.

Recommended representation is a recommended type of value for the concept term. Examples are "primitive" types such as string (i.e. free text), or complex types such as Codelist, that is used for those terms that have an associated Codelist in Codelist ID. There may be more than one recommended type; in this case, the first type is recommended over the others. For time types, it could be possible to use a more precise representation of time than the recommended type (e.g. Reporting Time Period instead of Observational Time Period).

Codelist ID is a unique identifier for the Codelist associated with the Concept. Most often it is the term's Concept ID prefixed by "CL_". For example, the "Observation Status" term has the Concept ID of OBS_STATUS, and the Codelist ID of CL_OBS_STATUS. This attribute is used only if the concept's "Recommended representation" includes "Codelist".

Related terms are entries in the SDMX Glossary that are closely associated with the concept term. It is possible here to create relationships between concepts, e.g. between "Reference metadata" and "Structural metadata". No hierarchy is created between the concepts linked, i.e. if a link is established between "Reference metadata" and "Metadata", a similar link will be established between "Metadata" and "Reference metadata".

Source is Source information from which the definition was extracted. The reference must be as complete as possible. When available, the source is followed by a hyperlink, i.e. a link to the source material for the term.

Other link(s) is/are Link(s) to material that is related**,** closely or loosely, to, but not directly associated with the concept source of the term, e.g. link to a general methodological document.

4.39 Statistical Products for Use outside the NSO
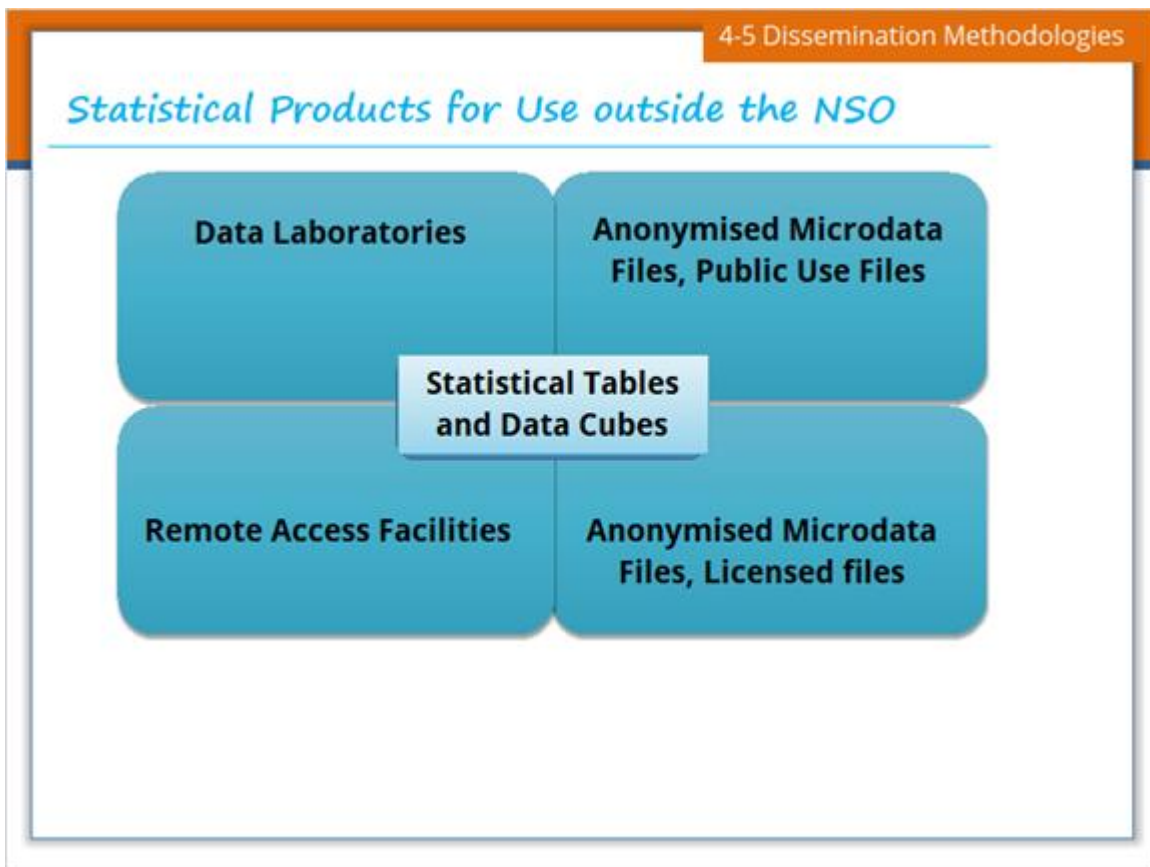
Notes:

Microdata are original individual data by unit in surveys: data on the characteristics of units of a population, such as individuals, households, or establishments, collected by a census, survey, or experiment. Microdata are not only data sources for the original surveys but also those for other statistics. Microdata need to be documented.

Any principles for microdata access must be consistent with the Principle 6 of the Fundamental Principles: "Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes". The following principles should be used for managing the confidentiality of microdata.

Principle 1: It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected.

Principle 2: Microdata should only be made available for statistical purposes.

Principle 3: Provision of microdata should be consistent with legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected.

Principle 4: The procedures for researcher access to microdata, as well as the uses and users of microdata, should be transparent and publicly available.

There are various ways a NSO can support research work. There is one important point which is not always understood. Microdata files can be anonymized by removing names and addresses and taking other steps (e.g. collapsing geographical detail) to ensure that the identification of individuals is highly unlikely when these files are looked at in isolation. This could be referred to as eliminating spontaneous identification. But other microdata files exist in the public and private sector, sometimes with individuals identified. Studies have shown that by statistically matching the NSO microdata files with existing files, unique records can be identified. The number can be quite significant depending on the amount of detail available in the NSO microdata file. Also in relative terms, the number of unique cases will be greater for smaller countries. These risks are not always well understood by NSOs. Of course, they will be reduced if techniques such as data perturbation or data swapping are used in the NSO microdata file.

Statistical products for use outside the NSO are the following:

Statistical Tables and Data Cubes

This can include both standard tables and special tables (or special analyses for that matter) generated at the request of the researcher. Some offices now release very detailed matrices, known as data cubes, which researchers can manipulate to support their own needs. However, if these are very detailed, the level of confidentiality risk can be similar to microdata.

Anonymised Microdata Files, Public Use Files

These are microdata files that are disseminated for general public use outside the NSO.

They have been anonymised and are often released on a medium such as CD-ROM, sometimes through a data archive. The level of confidentiality protection in Public Use Files should be such that identification is not possible even when matched with other data files. Public Use Files are a way of providing access to researchers in some countries.

Anonymised Microdata Files, Licensed files

Licensed files are also anonymised but are distinct from Public Use Files in that their use is restricted to approved researchers and an undertaking or contract is signed before files are provided to the researchers. Even if advertised as generally available to the public, they are not released before an undertaking or contract is provided by the researcher. Even though anonymised and other steps are taken to ensure that identification of individuals is highly unlikely when used in isolation, they may contain potentially identifiable data if linked with other data files; this is one reason why a preventive undertaking or contract is required. There may be other conditions of use that the NSO may impose on researchers.

Remote Access Facilities

Arrangements are now being made in many countries that allow researchers to produce statistical outputs from microdata files through computer networks, without the researchers actually 'seeing' the microdata. Because of the additional controls that are available through Remote Access Facilities, and the fact that microdata do not actually leave the NSO, access to more detailed microdata can be provided this way.

Data Laboratories

On-site access to more identifiable microdata, typically with stringent audit trails and NSO supervision. The access to more detailed data creates some inconvenience to the researcher, because of the requirement of working at the NSO, or at an NSO enclave.

4.40 Data Documentation Initiative (DDI)
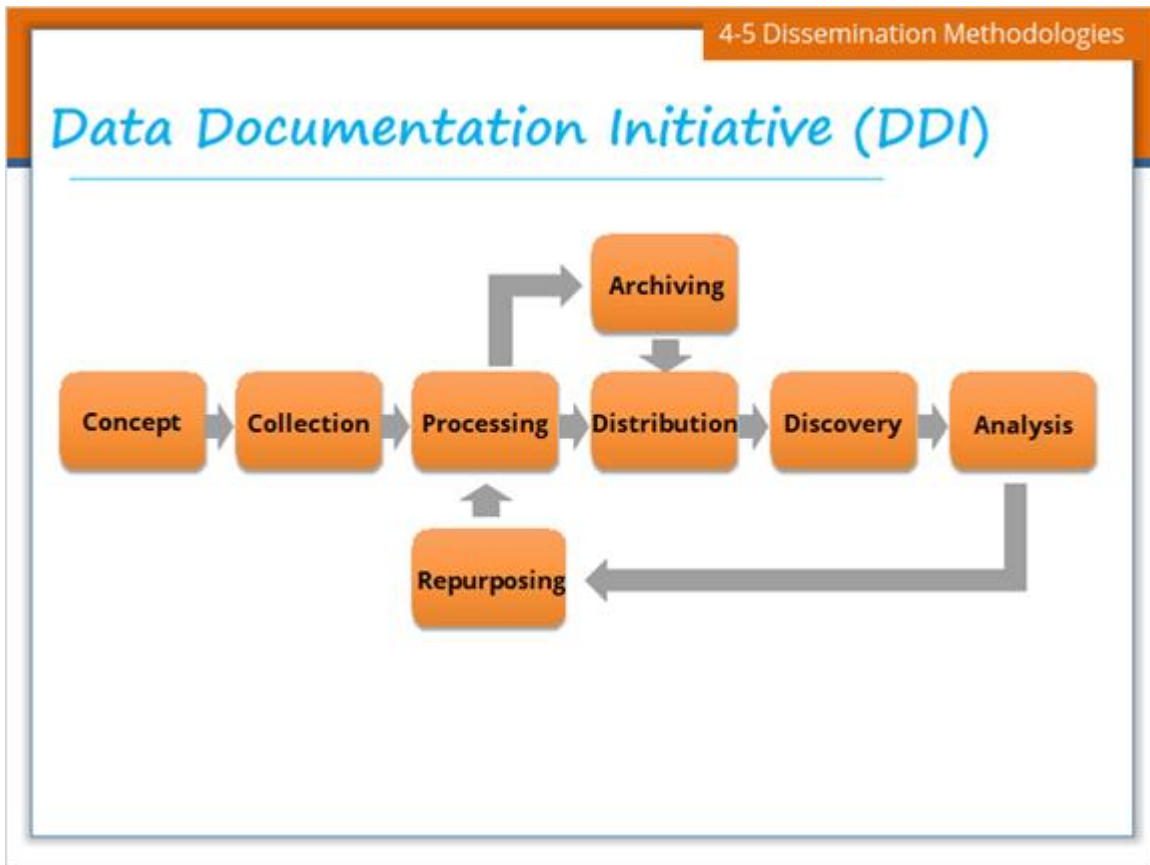
Notes:

The lifecycle model underlying the Data Documentation Initiative (DDI) - Lifecycle specification is shown in the slide. It is useful to think of DDI as supporting metadata-driven survey design. Over the life course of a survey that results in a data set - from initial conceptualization to data publication and beyond -- a huge amount of metadata is typically produced.

These metadata can be recorded in DDI format and re-used as the data collection, processing, tabulation, and reporting/dissemination take place. The DDI metadata are both documentary and also "machine- actionable" - that is, they can be used to drive processes and to support additional steps in the life cycle. This is especially important for describing how microdata are cleaned, edited, tabulated and anonymized in the production of aggregates.

4.41 4-6 Rough Sketch of Official Statistics by Scope

4-6 Rough Sketch of Official Statistics by Scope

Notes:

4.42 Social Statistics

Notes:

Specific statistics will be overviewed from this slide by four scopes: social statistics, economic statistics, agricultural statistics and environment statistics.

Please note that these four scopes overlap. For example, Household Income and Expenditure Surveys and Price statistics can be regarded both as social statistics and as economic statistics. System of Environmental and Economic Accounts (SEEA) is economic statistics and environment statistics as its name suggests. Gender statistics overarches all scopes.

This slide summarizes social statistics.

The most popular and obvious instrument for monitoring is the household survey. It provides data that can be disaggregated to show results for different population groups and has the benefit of providing information both on the beneficiaries and the non-beneficiaries of development interventions. There are a number of different household

survey models that can be used, each with its own strengths and weaknesses. The most complete coverage is provided by the Population Census. Although obviously not appropriate for day-to-day monitoring, the census is important because it provides the framework for almost all other household survey activities, including Agricultural Censuses and surveys, and provides periodic basic socio-economic information down to the smallest geographical areas. Integrated multi-topic household surveys are another form of enquiry that have become increasingly popular. They are particularly good as baseline surveys that can be used to measure poverty levels, identify potential problems in need of attention, and generally understand the way in which households establish mechanisms to cope with difficult living conditions. The big disadvantage is that they are difficult surveys to undertake, and many countries have neither the analytical nor the survey capacity to regularly carry out such large-scale complex surveys successfully. Lighter and more rapid household surveys are, however, becoming increasingly popular. Service delivery surveys have been used in market research for a long time but are relatively recent additions to an NSO's repertoire of surveys. They are very well-suited to monitoring early results. They are also easy to implement and can be repeated annually without disturbing any other survey work that the NSO may be undertaking.

4.43 Economic Statistics
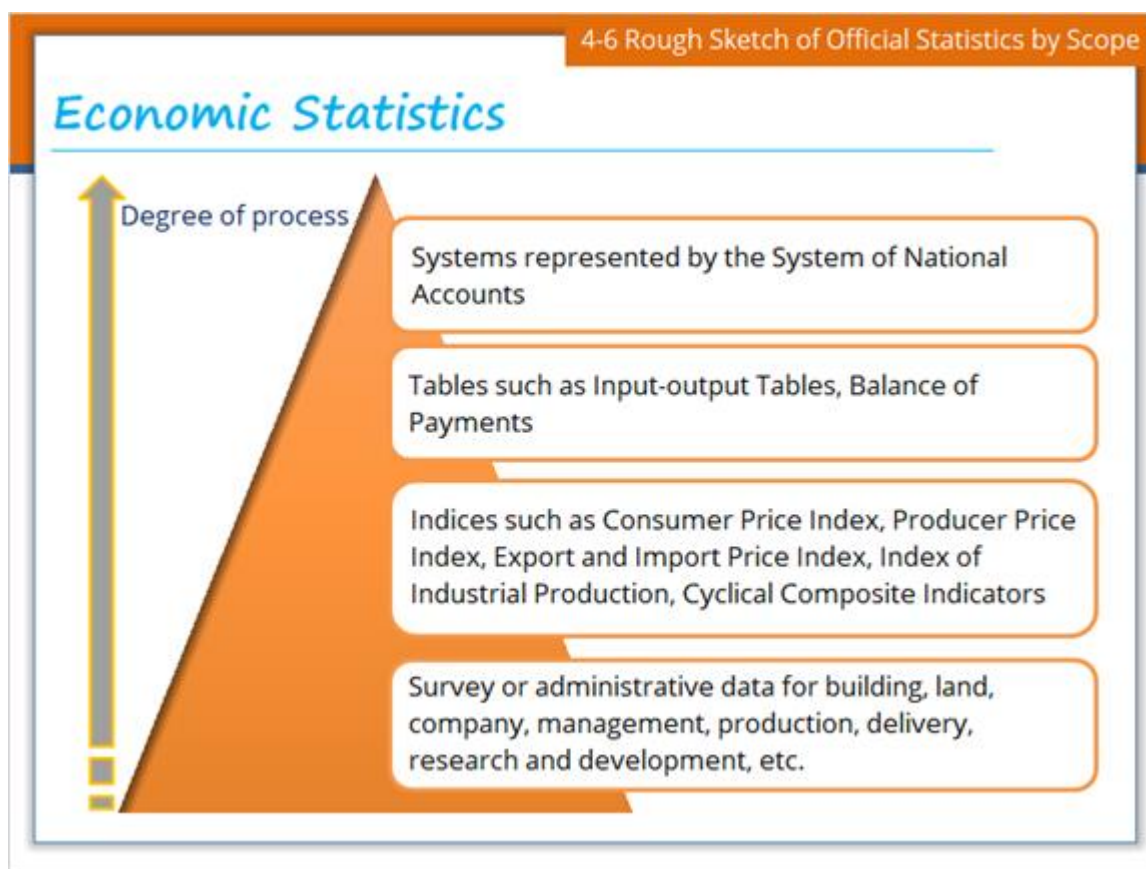
Notes:

NSOs conduct enterprise surveys to gather data on production, prices, employment, and other economic indicators. Amongst other things these data are used to produce measures of inflation and economic growth which are closely monitored and evaluated.

Index is formed from the ratio of aggregate values in the given period to the aggregate values in the base period. Index is useful for comparisons in time series.

Some economic statistics are highly processed. Input-Output Tables describe the sale and purchase relationships between producers and consumers within an economy. The Balance of Payments is a statistical statement that systematically summarizes, for a specific time period, the economic transactions of an economy with the rest of the world.

The System of National Accounts (SNA) is the internationally agreed standard set of recommendations on how to compile measures of economic activity. The SNA describes a coherent, consistent and integrated set of macroeconomic accounts in the context of a

set of internationally agreed concepts, definitions, classifications and accounting rules. SNA extends to various satellite accounts such as the System of Health Accounts (SHA), the Tourism Satellite Account (TSA) and the System of Environmental and Economic Accounts (SEEA).

4.44 Agricultural Statistics
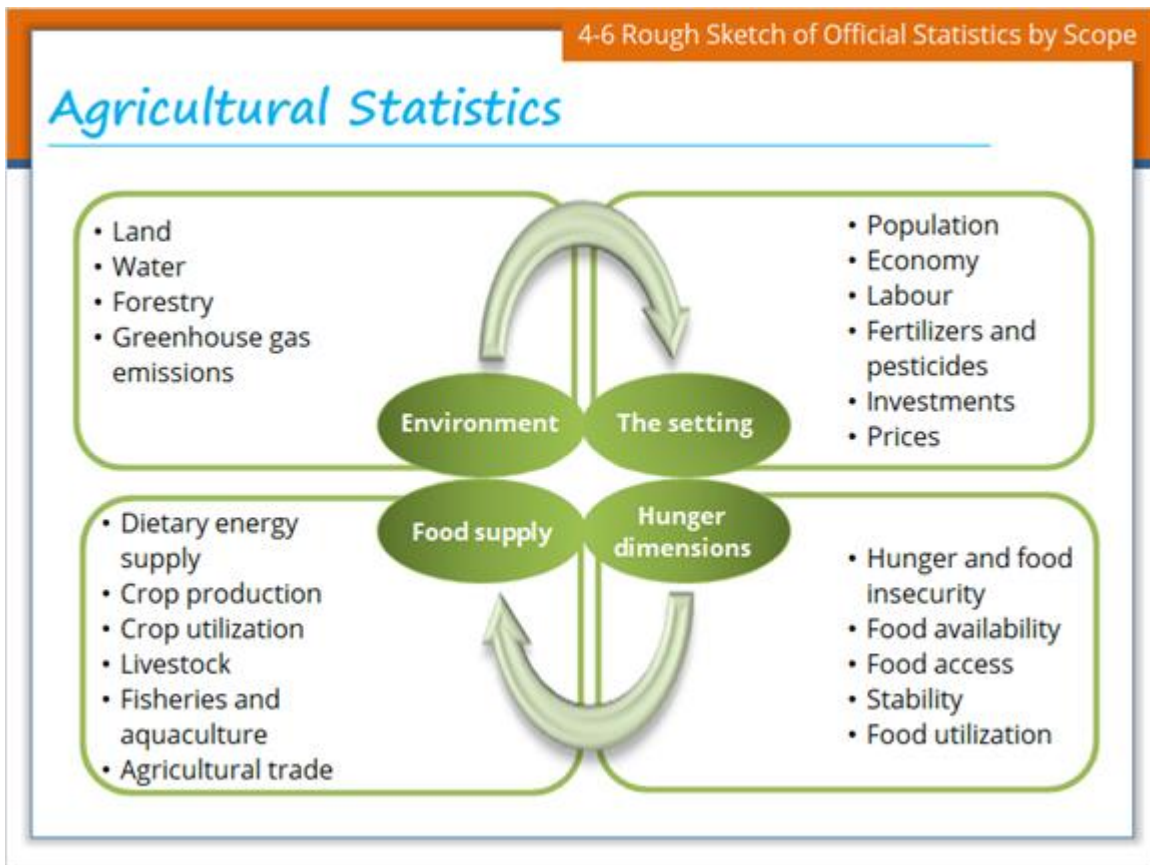
Notes:

Targets of agricultural statistics reach crops, foods, lands as well as farmers and establishments. Agricultural statistics are categorized into four groups: environment, the setting, food supply and hunger dimensions. Broadly speaking, agricultural statistics include forest statistics, fishery statistics and rural statistics. The Global Strategy to improve agricultural and rural statistics is a comprehensive framework for improving the availability and use of agricultural and rural data.

4.45 Environment Statistics
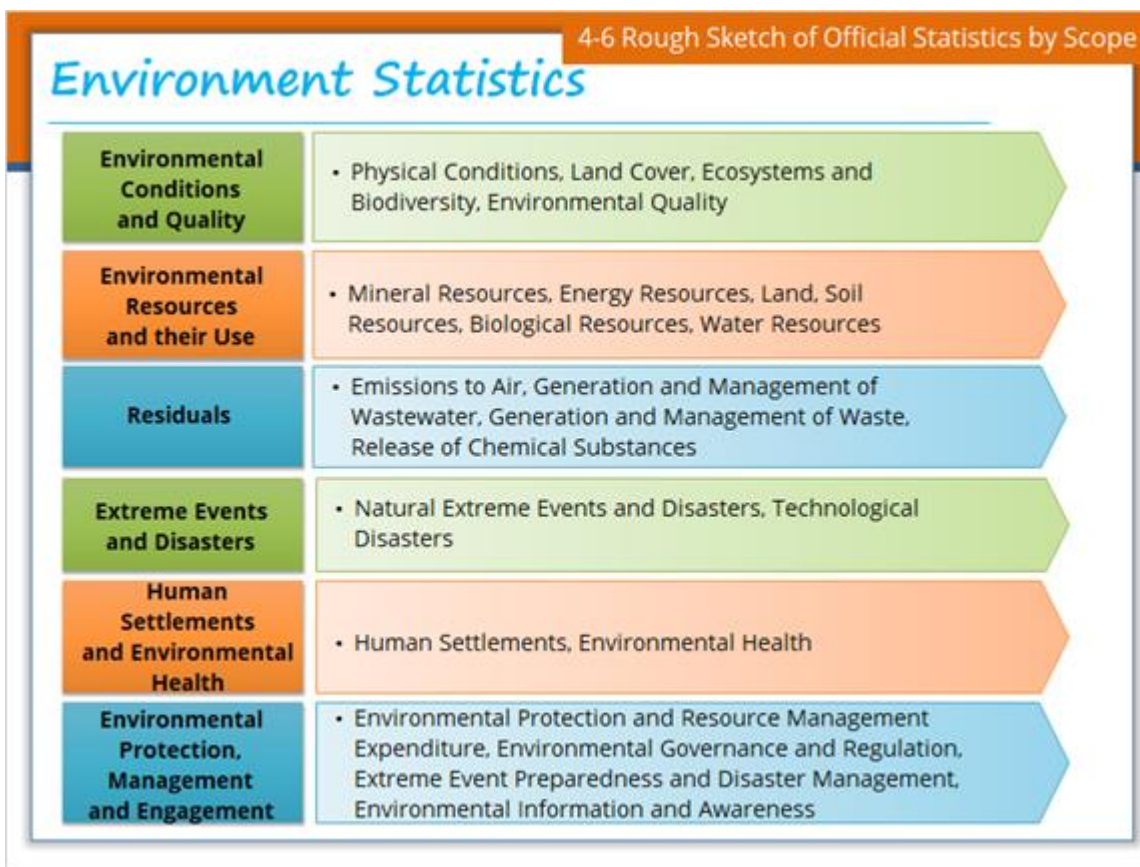
Notes:

Based on the Framework for the Development of Environment Statistics published by the UN, environment statistics extends across a wide area. The scope of environment statistics covers biophysical aspects of the environment and those aspects of the socioeconomic system that directly influence and interact with the environment.

For a highly processed system, the System of Environmental-Economic Accounting (SEEA) is a framework that integrates economic and environmental data to provide a more comprehensive and multipurpose view of the interrelationships between the economy and the environment and the stocks and changes in stocks of environmental assets, as they bring benefits to humanity. It contains the internationally agreed standard concepts, definitions, classifications, accounting rules and tables for producing internationally comparable statistics and accounts. The SEEA framework follows a similar accounting structure as the SNA.

This is the end of Lesson 4.

4.46 Summary of Lesson 4

Summary of Lesson 4

- International standards are the comprehensive body of international statistical guidelines and recommendations that have been developed by international organisations working with national agencies.
- Three factors: purpose, unit and time should be considered for census and sample surveys.
- Statisticians need to devise a methodology of sampling to achieve the utmost accuracy under small sample size.
- Civil registration provides the basis for individual legal identity but also allows countries to identify their most pressing health issues.
- A Statistical Business Register is a regularly updated, structured database of economic units in a territorial area, maintained by an NSO, and used for statistical purposes.
- Metadata should assist the user in searching for statistical information, interpreting its content and, after its potential downloading from the Internet, should help in post-processing statistical applications.

Notes:

4.47 End of Lesson

Notes:

2. Menu