ESSnet KOMUSO

Quality in Multisource Statistics

http://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics_en

Specific Grant Agreement No 3 (SGA-3)

Framework Partnership Agreement Number 07112.2015.003-2015.226

Specific Grant Agreement Number 07112.2018.007-2018.0444

Work Package 2

Quality Guidelines for Frames in Social Statistics

Version 1.5 - 2019-06-26

Prepared by: Thomas Burg (Statistics Austria), Alexander Kowarik (Statistics Austria), Magdalena Six (Statistics Austria), Giovanna Brancato (Istat, Italy) and Danuté Krapavickaité (LS, Lithuania)

Reviewers: Fabian Bach (ESTAT), Lionel Vigino (ESTAT), Li-Chun Zhang (SSB, Norway), Ton de Waal (CBS, The Netherlands), Angela McCourt (CSO, Ireland), Eva-Maria Asamer (Statistics Austria) and Christoph Waldner (Statistics Austria)

ESSnet co-ordinator: Niels Ploug (DST, Denmark), email npl@dst.dk, telephone +45 3917 3951

Table of contents

1 Purpose and Objectives of the Document

The present document is a deliverable of the ESSnet KOMUSO[1] under the margin of the ESS.VIP Vision 2020.ADMIN which consists of a project portfolio relevant for realizing the goals of the European Statistical System (ESS) Vision 2020[2]. The provision of guidelines is one of the key objectives formulated in the business case of ESS Vision 2020.ADMIN.[3] Therefore, the main goal of this document is to provide guidelines for all processes relevant to **Frames in Social Statistics**. In Official Statistics, the description of social phenomena by statistical figures is one of the main functions of national statistical systems. Frames are essential to this function. Besides this general objective, as a contribution to the implementation of the ESS Vision 2020, this document pursues three more specific objectives.

The first specific objective concerns compliance with the Code of Practice. In the last decade of the past century, the question of quality in the production of Official Statistics has been approached in a systematic way. By defining quality on a multidimensional basis and laying down the quality dimensions in the EU-Statistics regulation 223, the concept for a framework, which was later enhanced by formulating and adopting the European Code of Practice (CoP) for official statistics, was established. Principle 4 of the CoP- "*Commitment to Quality*" – and, specifically indicator 4.1states, "*Quality policy is defined and made available to the public. An organisational structure and tools are in place to deal with quality.*" The Quality Assurance Framework (QAF) suggests the development of quality guidelines for relevant statistical process steps as one of the key methods to provide evidence for compliance with principle 4. This concrete element together with the stipulation of indicator 7.3 ("*The business register and the **frame for population surveys** are regularly evaluated and adjusted if necessary in order to ensure high quality*") shows that the CoP demands the continuous development and use of good (or, if possible, best) practices with respect to frames for population statistics which have to be manifested in concrete guidelines. In line with this, one of the main objectives of the document is **to deliver a building block for safeguarding the compliance with the Code of Practice with respect to the construction, use and assessment of frames in social statistics.**

As a second specific objective the guidelines try to provide producers of Social Statistics with systematic guidance for all process steps relevant to frames in Social Statistics. The procedures of National Statistical Institutes (NSI) working with Frames in Social Statistics seem to be more heterogeneous than the procedures for Economic Statistics, where the development and maintenance of the frame (which in most of the cases is equivalent to the business registers) was in some way generic for NSIs all over the world. Looking again to indicator 7.3 of the CoP, it tells us about the business register and the "frame for population survey". Looking to various NSIs shows different scenarios of constructing, using and maintaining frames. Thus the second objective of the document is to **provide basic generic guidance regarding all relevant processes for frames in social statistics in a systematic way based on agreed definitions and standards**.

[1] KOMUSO ESSnet on quality of multisource statistics
[2] See as well http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020
[3] The https://circabc.europa.eu/sd/a/be1c01a7-a952-4e89-ba10-5417a91ae118/ADMIN%20Business%20Case%20v5.1.pdf

Historically, the idea of using frames originates from investigating social phenomena by conducting surveys. It turned out, that you need to have some kind of list of the units of interest mainly dwellings aiming to reach households and/or persons living there, from which you can draw some (a sample) and conduct a survey. As a result, the main interest of NSIs was in sampling frames. In recent years it has become clear that sample surveys - while still of significant importance - are not the only way of gathering statistical information. In this regard the role of frames as a direct source for delivering figures becomes more important. Hence, the third specific objective of this document is **widening the view on frames in social statistics to use them as a possible direct source in a multisource environment**.

2 Frames in Official Statistics

2.1 Definition of a Frame

One of the main goals of statistics is to figure out characteristics of a population and describe it in a numerical way. The investigation of these characteristics is very often done by surveying units belonging to this population. In official statistics the main focus lies on special populations, so called **target populations**. To do this in an efficient way the idea is to select units of a set, which ideally covers the target population as good as possible. Having identified a set, we talk about a **frame.**

There are several ways to exactly define a frame in official statistics. In the following we present two possible definitions for a frame. The first definition is offered by ESSnet KOMUSO, SGA-1 WP 2, *Quality measures and indicators of frames for social statistics,* prepared by the project team of WP 2. This definition is pragmatic and focuses on the frame as a picture of the target population.

Definition: A **Frame** is any list, material or device that delimits and identifies the elements of the target (survey) population. Depending on the use case, a frame may allow access to and/or provide additional characteristics of the element.

There are more descriptions and definitions which can be found. One of them refers to the standard reference, Lessler, Kalsbeek, (1992). This definition of a frame focuses more on the complexity of a frame and provides an outlook on possible use cases for a frame:

> The **frame** consists of materials, procedures, and devices that identify, distinguish, and allow access to the elements of the target population. The frame is composed of a finite set of units to which the probability sampling scheme is applied. Rules or mechanisms for linking the frame units to the population elements are an integral part of the frame. The frame also includes auxiliary information (measures of size, demographic information) used for (1) special sampling techniques, such as stratification and probability proportional to size sample selection, or (2) special estimation techniques, such as ratio or regression estimation.

There are some aspects mentioned in the definitions which deserve further consideration. First when talking about a target population the units of the frame should correspond to elements of the target population. In this regard we talk about a **frame unit.** Simply a frame is seen as a flat file in which each record corresponds to a certain basic frame unit covering one element of the target population. It might be the case that more than one unit level exists corresponding to the target population in the frame. This means that there are certain aggregations combining basic units to other units of interest (in chapter 2.2. we will talk about the case for social statistics).

A second aspect is related to the set of variables associated to a frame unit. As discussed earlier, one of the main purposes of a frame is to select units and contact them for surveying. This requires the existence of **contact variables** which allow the NSI to include the unit in a sample in an operational and usable way. Basic frame units should be unique in a frame in the sense that there is a 'one-to-one' correspondence between the basic frame unit and one object in the target population. In order to achieve this goal, **a unique identifier** for each basic unit should be available. Some of the information for a frame unit sometimes depends on external sources. Therefore the existence of **linking variables which** map the relations of elements to other - most of the time external - registers is important. Lastly, the variables contained in a frame should allow for more sophisticated use, as outlined in the Lessler Kalsbeek definition: 'The purpose should be to enrich the frame by as much information as possible. Therefore the frame shall consist also of a lot of **contextual information** often called **auxiliary variables** allowing for stratification, unequal probability sampling and/or systematic analysis of the data material. Such a frame enriched with other variables, is frequently called a **Rich Frame'.**

Summarizing the basic considerations about frames we can list the following characteristics which should ideally relate to a frame.

- The frame is available in an usable IT-format;
- The frame aims to map the target population as accurately as possible;
- The frame contains basic units corresponding to the units of the target population and assigns every unit an identifier;
- The variables in the frame include linking variables which facilitate connecting the basic units to external registers;
- The frame is enriched by auxiliary variables enabling an enhanced usage (i.e. at least with contact variables);
- If there are composite units derived from the basic units the linkage of these units to the basic units is included.

Frames versus Registers

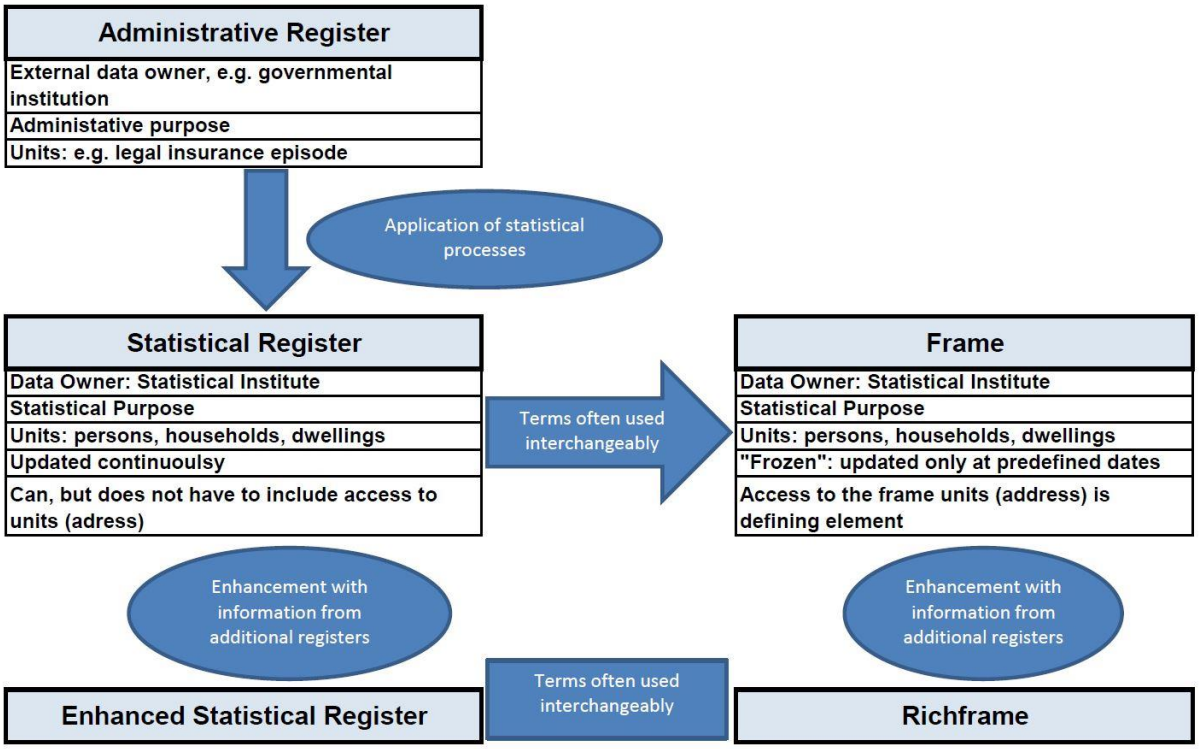


Figure 1 Differences and links between Administrative registers, Statistical Registers, and Frames

A register is a written and complete set of records containing regular entries of items and details on a particular set of objects[4].

When talking about registers one has to distinguish between administrative and statistical registers, as shown in Figure 1. Whereas an **administrative register** is maintained by an external data owner for a dedicated administrative purpose, a **statistical register** is a register created for statistical purposes, normally by NSIs. A statistical register is typically created by processing data from registers and/or other administrative data sources.[5] Two of the most important differences between statistical registers and administrative registers concern the data owner and the unit

[4] https://stats.oecd.org/glossary/detail.asp?ID=3003
[5]Glossary, worked out in the Task Force in Frames in Social Statistics, VIP Admin

of interest: in the case of statistical registers the owner is the statistical institute, and the unit is the statistical unit of interest, e.g. a person, a household or a dwelling and not an insurance episode, or enterprise and or legal unit. Statistical registers are often based on administrative registers, but diverse process steps are applied so that the statistical register fulfils the requirements of the statistical institutes.

A prominent case of an administrative register is the (administrative) population register (sometimes also called "Register of Residents"), where the records correspond to the persons living in a specific territory such as a country. It is in the interest of the persons belonging to the population to participate in the registration process because the registration is either required by law or it is a precondition for exercising rights (e.g. registration is required in order for a person to be enlisted in the voter list, in order to have access to public health services etc.). The data owner and the authority behind a population register is usually a governmental institution other than the NSI. For the population register, statistical procedures can be applied to eliminate gaps or to identify and remove persons from the register who do not belong to the target population any longer, for example because they moved away but did not deregister. A "signs of life" analysis, which is described in more detail in the Chapter 3.5 on updating frames, can also be applied here to avoid coverage errors.

Another common register beside the population register, widely used by statistical institutes, is the business register.

In the statistical world, it is often clear that one talks about statistical population registers and statistical business registers, with the statistical institute as data owner. Given that the term "statistical" is often not stated explicitly.

One of the main purposes of having a statistical register is to use it as a frame (see again Figure 1). A frame **allows access to the units** of the target population. Sometimes statistical institutes compare contact information, such as the address of the statistical population register, to other available data sources such as an address register to ensure its validity and usefulness as a sampling frame. Because the statistical population register often includes information allowing access to its elements, for example the address in a population register, the two terms are often used interchangeably. Although there seems to be no clear, internationally agreed distinction between the terms "statistical register" and "frame", the updating process seems to differ between the two terms. Whereas a statistical register is updated on a continuous basis, a frame is often seen as a "frozen version" of a register, especially when the frame serves as sampling frame.

Having in mind various use cases for frames the enhancement of the frame by as much information as possible should manifest one of the key targets. Depending on the predominating terminology in the NSI, the enrichment process leads to an **Enhanced Statistical Register** or a **Rich Frame (**see also Figure 1**).** In the case of an enhanced statistical register which maps the population of a country, the term **Enhanced Population Dataset (EPD)** has been established. An EPD consists of all relevant available information about the units, such as household, dwelling, address, classification variables and of course contact information. EPDs are based on the population registers but are enriched by additional data sources such as dwelling registers, address registers, tax registers, register of unemployed etc. If more than the unit itself and the necessary contact information are available in the frame, it can already be called EPD, e.g. the central population register with basic socio-demographic variables such as age and gender.

Typology of frames: From what was already described above, it can be seen that it is sometimes useful to assign frames with certain properties regarding different classification criteria. Within the following chapters we will on some places refer to several properties for frames. The figure below shows the most important classification types for frames.

Accessibility	Contents	Timeliness	Use
Direct frame	List frame	Frozen frame	Master frame
Indirect frame	Area frame	Continuous frame	Master sample
Single frame	Integrated frame		Sampling frame
Multiple (dual) frame	Rich frame		Rich frame

You can find further explanations of the different types in chapters where the term is used at first time. An exact definition of each frame type can be found in the glossary.

All definitions relevant for this document can be found in the glossary (annex V). At first appearance of an important item there is link which can be used in an electronic format of the document. This link leads directly to the corresponding definition in the glossary.

2.2 Frames in Social Statistics

The production of **social statistics** is one of the central functions of a National Statistical System (NSS) and particularly of the ESS. It covers a broad range of statistics about different aspects of people's lives, including statistics on poverty and living conditions as well as statistics on education, health and labour market. There are several surveys dedicated to the field of Social Statistics and a lot of statistical products complete the picture in this area. Prominent examples include:

- European Union Labour Force Survey (EU LFS)
- European Health Interview Survey (EHIS)
- Adult Education Survey (AES)
- European Union Survey on ICT usage in households and by individuals
- Time Use Survey (TUS)
- Household Budget Survey (HBS)
- European Union Statistics on Income and Living Conditions (EU-SILC)

Normally, one associates certain statistical units to social statistics. The main focus and the basic target population for an NSS is the population of the country involved. There might be some minor discrepancy between the definitions of what is exactly meant by the population (people staying in the country, people having citizenship etc. …) but, having an eye to possible frames it is understood that the basic units of a frame in social statistics l consist of persons belonging to the population of the country. With regard to composite units, one can observe that persons in the population are usually combined in households. Besides persons and households, dwellings are also occasionally units of interest. It is frequently the case that there is a one-to-one relationship between households and dwellings because it can be expected that one household lives in exactly one dwelling, but characteristics of the units will be different. Summarizing we can say that the core units in social statistics are **persons, households and dwellings**. It is worth mentioning that the concept of 'household' might cover different "household-like units". Most commonly, one can compare dwelling with living household, or household with family, where each type of unit has several definitions in use. In summary: the basic units of a frame in social statistics are persons. These basic units can be aggregated into composite units such as households and dwellings, where the following boundary conditions have to be respected:

- Every person belongs to one (and only one) household;

- Every household is located in exactly one (and only one) dwelling.

Regarding the uniqueness of units, it is necessary to have unique identifiers assigned to each person mapped in the frame (practically speaking often a person number, a household number and a dwelling number). By taking into account the other issues pointed out in chapter 2.1, we only have to slightly modify the list of characteristics for a frame to arrive at an equivalent for a frame in social statistics in the context of this document:

- The frame is a list of persons available in a usable IT-format;
- The frame aims to map the population of the country of the NSS as accurately as possible;
- The frame contains persons as basic units and households and dwellings as composite units;
- Each person household and dwelling has a unique identifier;
- The frame is enriched by auxiliary information enabling enhanced usage (i.e. at least with contact variables);
- The variables include linking variables which allow connecting persons, households and dwellings to external registers;
- The frame is enriched by sufficient information enabling enhanced usage. This includes contact variables which allow for the inclusion of persons, households or dwellings into surveys.

The issues described in the last chapter mostly refer to **list frames.** It should be mentioned that sometimes **area frames** are in use. Those are usually made up of a hierarchy of geographical units which, in turn contain units in the survey population. That is, the frame units at one level can be subdivided to form the units at the next level.

2.3 Processes involving Frames

Basically there are two main processes relevant to frames in social statistics. One concerns **the construction and maintenance** of a frame, the other one concerns the **use** of a frame.

The Generic Statistical Business Process Modell (GSBPM) mentions the term frame in two specific process steps, namely 2.4 “Design frame & sample” in the Design-Phase and in 4.1 “Create frame & select sample” in the Collect-Phase. More specifically, these two sub-processes are described as follows in the GSBPM:

> GSBPM 2.4. Design frame and sample
>
> This sub-process only applies to processes which involve data collection based on sampling, such as through statistical surveys. It identifies and specifies the population of interest, defines a sampling frame (and, where necessary, the register from which it is derived), and determines the most appropriate sampling criteria and methodology (which could include complete enumeration). Common sources for a sampling frame are administrative and statistical registers, censuses and information from other sample surveys. This sub-process describes how these sources can be combined if needed. Analysis of whether the frame covers the target population should be performed. A sampling plan should be made: The actual sample is created in sub-process 4.1 (Create frame and select sample), using the methodology, specified in this sub-process.
>
> GSBPM 4.1. Create frame and select sample
>
> This sub-process establishes the frame and selects the sample for this iteration of the collection, as specified in sub-process 2.4 (‘Design frame and sample’). It also includes the coordination of samples between instances of the same statistical business process (for example, to manage overlap or rotation), and between different processes using a common frame or register (for example to manage overlap or to spread response burden). Quality assurance and approval of the frame and the selected sample are also undertaken in this sub-process, though maintenance of underlying registers, from which frames for several statistical business

processes are drawn, is treated as a separate business process. The sampling aspect of this sub-process is not usually relevant for processes based entirely on the use of pre-existing sources (e.g. administrative sources) as such processes generally create frames from the available data and then follow a census approach.

In a broader sense, the construction is a statistical process itself, involving some sub-processes (such as 5.1, 5.5 - 5.7, 6.2. and some others)

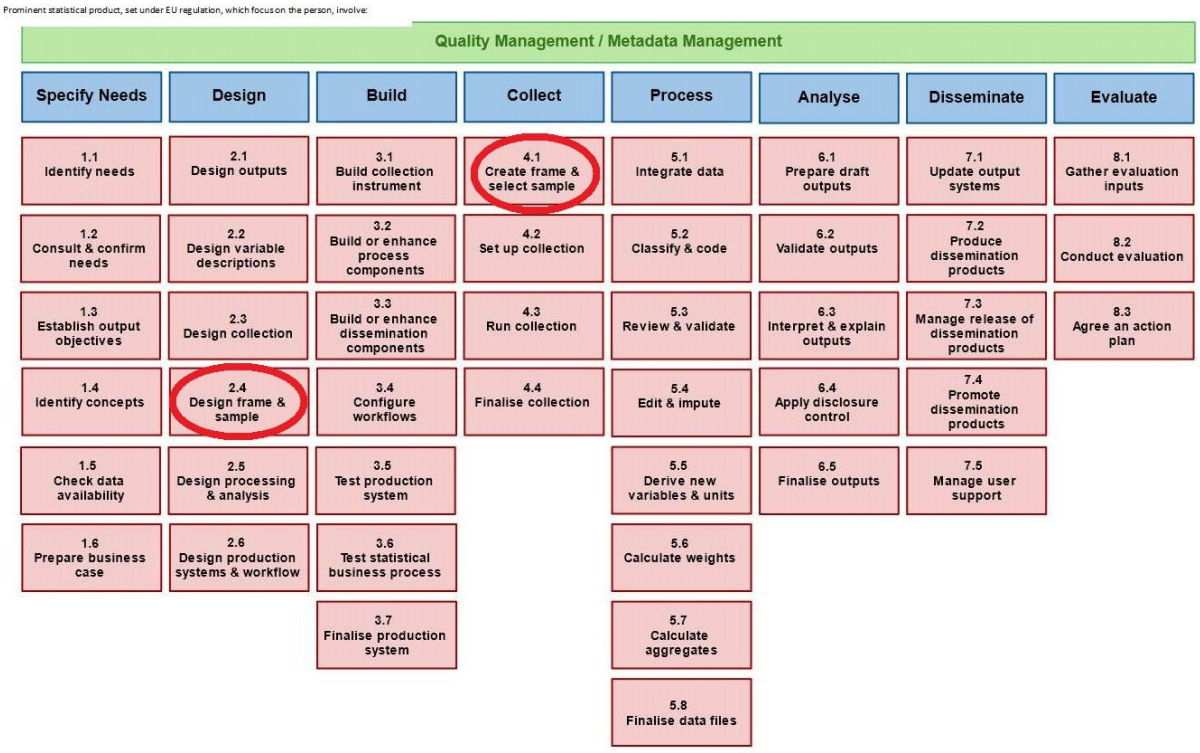


Figure 2: The Generic Statistical Business Process model – red circled are the sub-processes where frames are mentioned explicitly

As 4.2 shows a frame plays a decisive role when used for designing and drawing a sample. In this case a **sampling frame** is discussed**.**

A **sampling frame** is a frame that could be used as a basis for sampling and normally is any list, material, or device that delimits, identifies, and allows access to the elements of the survey population.[6]

The frame plays a fundamental role in **survey sampling**. Probability sampling involves selecting a subset of units from a finite collection of units in a manner that lets one determine the probability of obtaining that subset. The sampling frame is the list of units covering a finite population of which the probability sampling mechanism is applied.

Many sampling designs use auxiliary data to produce **more efficient sampling designs**. The relative efficiency of one sampling procedure to an alternative one is given by a 'design effect' defined as the inverse of the ratio of the variances under the two designs (Sukhatme and Sukhatme 1970, citied in Lessler and Kalsbeek 1992). The sampling design with the smaller variance of the estimator for a finite population parameter is said to be more efficient or to

[6] Glossary, worked out in the Task Force in Frames in Social Statistics, ESS Vision 2020 Admin

have a higher precision. A more detailed view on the use frames for sampling is provided in chapter 4.1.1. The frame can also be used after the actual sampling procedure. A rich survey frame makes it possible to enhance precision of the parameter estimates by applying complex **weight-calibration** methods (for an example see *Weighting Procedure of the Austrian Microcensus*, Meraner et al 2016). In (Meraner et al 2016), the authors say, *Theoretically, every characteristic known for the units in the sample and for the whole population can be used as a calibration variable, but only variables correlated with target variables enhance precision. Practically of course there are some restrictions, e.g., the sample size.*

Another application field is a **non-response analysis,** as auxiliary information in a survey frame is very useful to perform **non-response analysis**. This allows drawing conclusions about the question if certain population groups have a higher non-response thereby potentially causing a bias of the parameter estimates.

During processing, the frame can provide information suitable for data cleaning (Editing and Imputation). In the **editing phase**, auxiliary information of the frame can help in finding and correcting faulty values in the survey data. In case of missing or inconsistent data items**, imputation** techniques, such as k-nearest-neighbour or regression based methods, substitute the missing or rejected item with an estimated value. Auxiliary information in the rich frame can be crucial in finding the most appropriate donor or modelling target variables.

Census

A **register based census** can be regarded as one of the most important applications of the use of frame data. It replaces e.g. the population census, the housing census and the census of enterprises and their local units of employment. The information needed is not collected from the citizens themselves, but is taken directly from already existing administrative registers. In Austria, where the NSI has conducted a register based census since 2011, the Central Population Register represents the backbone of the register-based census. The other base registers are the Housing Register of Buildings and Dwellings, the Business Register of Enterprises and their Local Units, and the Register of Educational Attainment, all of which are maintained by Statistics Austria itself, as well as the Central Social Security Register, the Unemployment Register and the Tax Register (not including data about income).

For some countries – especially countries which have no central population register - it works the other way round: The census is a source for a frame in social statistics. The census information is then updated with delta information, mainly coming from registers which do not contain the whole scope of a census. In the next census round the frame is then updated with the new information for the whole country.

In this context, some countries are going to combine a population dataset (register) with some form of coverage survey.

3 Constructing and Maintaining Frames in Social Statistics

3.1 Sources for Constructing Frames

The selection of adequate data source for constructing frames is a necessary prerequisite for guaranteeing its utility. Thus, the choice and evaluation of sources are important steps in the production of the frame. In this regard the construction of a frame can be seen as a multisource scenario in which statistical information heading for a resulting database is combined on a continuous basis.

When selecting sources it is necessary to **investigate the landscape of available data** and to distinguish between sources available internally and data coming from external sources. In the internal case you will have direct access, sometimes a very good possibility of designing the production of the data or at least a direct access to all relevant metadata. On the other hand, using external sources will require efforts to establish contacts to external data holders. In this context it is very supportive to provide the holders of external sources with sufficient feedback in order to improve situations where the data sources are diagnosed as improper.

Criteria for evaluating a source before it can be used for constructing and/or enriching a frame can be defined by properties of the data provided by the source. These criteria include questions about completeness with respect to the population as well as the situation concerning certain variables in the source. **Comprehensive metadata** for assessing and understanding input sources can be seen as one of the most important ingredients for evaluating an input source correctly. This holds not only for the input data set, but also for all variables to be integrated into the frame. Each administrative file chosen for the construction of the frame should undergo soft-editing, harmonization and standardized coding, and parsing of names and addresses. Soft-editing includes checking the validity of the identity numbers based on the number of digits, the ordering and the check digit. Duplicates and out-of-scope units should be removed, including those who have emigrated or died. If the identity numbers of parents and spouses are given for most of the records in the file, the administrative family units can be generated. For those without identity numbers of family members, the algorithm for constructing households should include joining individuals with the same last name (if available) living at the same address. Also, married couples with different missing or invalid address information should be joined and placed in one of the addresses most likely to be the true address.

One important characteristic of frames is that they claim to **map the population as accurately as possible for a certain reference periods or due day.** Therefore, the availability of the input sources in a timely manner is a significant issue. The optimal solution would be that the input sources are undergoing the same update cycles as the frame does.

As we explained in Chapter 2 frames in social statistics shall cover **various kinds of statistical units**. These units - namely persons, households, dwellings and /or buildings - are connected within the frame by established links and keys. Therefore, the evaluation of the situation within a source with respect to the different levels and units should be clarified in advance.

Sometimes you encounter a scenario where a certain constructed data set (very often a central population register) is used as a **backbone,** which is later enriched by information coming from other sources. It is evident that in such a case the input data for the backbone should be of very high quality which means that the resulting outputs are sufficient with respect to the quality dimensions relevant in the ESS. Guidelines broken down by quality dimensions for statistical outputs in the light of frame data usage are provided in chapter 4.3 of this document.

When combining data from different sources the question of **linking the corresponding units** has to be considered. For that reason, the question of identifiers and/or adequate linking procedures has to be addressed, when assessing a possible source for a frame.

Challenges

The challenges below deal with the impact on quality. As it regards the direct use of statistics more information can be found in chapter 4.3

Challenge 1.The quality of the source will have an impact on the **accuracy** of statistical products, depending on the kind of use and the processes in which the frame plays a key role. When used as sample frame, contact errors originating in quality deficits of a certain frame, will have a decisive impact on the net sample size and will thereby increase sampling errors.

Challenge 2. Lack of completeness of sources as well as definitions and content of variables, which were not analysed exhaustedly, can lead to **coverage errors** that can reduce the sample size and introduce bias in the final estimates **and misclassifications** that might affect the efficiency of the sample design.

Challenge 3. When considering sources, a decisive question is the suitability of the variables and units of interest, since there is some danger of not measuring the right things and to miss the user demands. This can result in problems of **relevance and coherence**.

Challenge 4. By not taking care to access sources in a timely manner, and when the production of the frame takes a lot of time, **timeliness and punctuality** can be affected.

Quality guidelines 3.1

Quality Guidelines - Sources for Constructing Frames

Guide 3.1.1: Documentation, comprehensibility and replicability of the **decision process for source selection.**
The decision process for selecting a source for constructing a frame is well documented, comprehensible and replicable.

Minimum requirement(s):

- The decision process is mapped within the quality reporting system in the NSI.
- Responsibilities are clear assigned.

Guide 3.1.2: Regular process in order to look for new sources.
A regular process of looking for new useful sources is scheduled.

Guide 3.1.3: Assignability of **variables provided by sources to statistical units.**
Variables provided by sources used for constructing frames in social statistics are assignable to statistical units (e.g. households and/or persons) in a direct and efficient way.

Minimum requirement(s):

- Definitions of units and variables and their relationships within the sources are clearly identified and documented.
- A complete set of metadata for every source is available.

Guide 3.1.4: Contact to the data holder.
When using an external source to create the frame the contact with the data holder is well established.

Minimum requirement(s):

- There are regular meetings with the data holder
- A protocol determining the regular delivery of external data exists
- A well designed feedback process reporting all problems and shortcomings is defined

Guide 3.1.5: Format and transparency of **metadata.**
Metadata describing every source and every variable within each source are available in a standardized format and transparent and clear to all users of the frame.

Minimum requirement(s):

- All relevant variables in the source are mapped by metadata
- Every delivery of external data is accompanied by relevant metadata

Guide 3.1.6: Ability to **link sources** to the frame.
The sources used for constructing the frame are linkable to the frame. If there is no common identifier the procedures of linking and potential error rates are well documented.

3.2 Organization, support, planning and coordination

Recruiting a source or constructing and/or updating it, does not only depend on conceptual or other content related criteria. There are further issues which need to be considered and planned. It is likely that within an NSI several administrative units handle the frame. First of all a central administrative unit should run and maintain the frame. This unit can be seen as the **owner** of the frame. It is responsible for all tasks concerning updating and maintaining the frame.

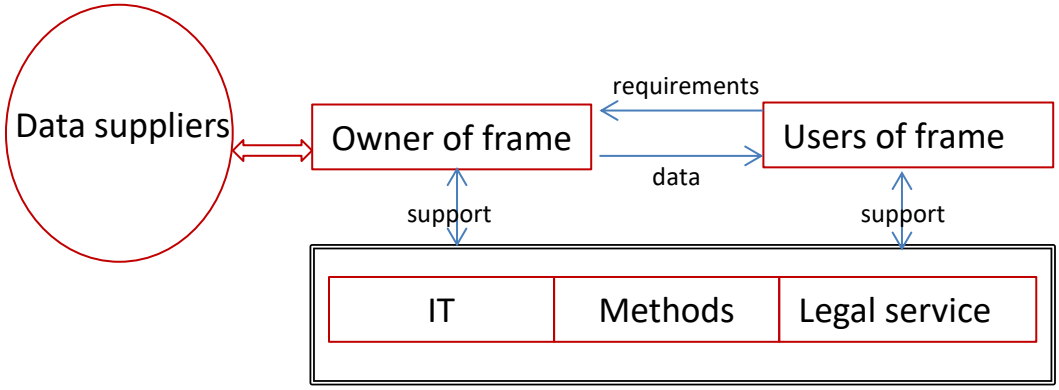


Figure 3: Coordinating and organizing a frame: roles and links of contributing instances

The presumptive **users of the frame** (subject matter units) formulate certain requirements, depending on specific use cases (sampling, direct use or support for certain processes). Based on these requirements, the frame-owner can decide if and how data can be provided for the users. However, this simple work flow might and will in most of the cases require additional actors supporting the use case in a consulting manner. As the graphic above shows, necessary **support** can come from IT, from legal service units and from the methods unit. This support may be related to the specific use case but may also be necessary for the owner in a generic way for adequately performing the duties of updating and maintaining the frame.

The role of the frame owner is a decisive one because this unit has **to organize and coordinate** all activities related to the frame. Therefore, all relevant information-requirements for use cases, feedback from the users of the frame etc. - has to be provided regularly to the frame owner.

One of the most important boundary conditions for maintenance of a frame is an **adequate legal framework**. Therefore, the aspect of **data protection**[7] in conjunction with the question of linkability is crucial. This holds especially in the context of social statistics, where individuals as the main unit of interest are involved. It is crucial for the trustworthiness of official social statistics to guarantee the identifiability of a single person is exceptionally only for contacting the respondents. Names and other forms of personal information have to be deleted as early as possible. In this regard the legal service of the NSI is an important function clarifying all questions related to data protection, data access and other legal aspects.

Particularly when it comes to sampling the **role of methodological experts is crucial. The existence of a central methodological unit** in accordance with European Quality Assurance Framework found on the organizational chart of every NSI is essential. In the case of sampling, the user rights of the frame (then often called sampling frame) are

[7] Data protection refers to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data as defined in the Regulation (EU) 2016/679.

delegated to the methods unit. Regular contact between the frame owner and the methods unit is important in order to come to a solid basis where all relevant sample designs can be realized.

Challenges

The challenges below deal with the impact on quality. As it regards the direct use of statistics more information can be found in chapter 4.3

Challenge 1. Inadequate organisation and lack of coordination will lead to an inappropriate data basis within the frame. First of all, this can distort **timeliness and punctuality.** These problems can also lead to increased **non sampling errors**.

Challenge 2. An inadequately defined work flow might lead to the situation that the requirements of use cases are not met and sampling procedures might not be implemented correctly. This, in return, can lead to a **loss in accuracy** of the resulting estimators.

Challenge 3. Non regular data delivery by external data suppliers and frequent changes of formats and contents can lead to breaks in time series, thereby distorting **comparability over time.**

Quality guidelines 3.2

Quality guidelines - Organization, support, planning and coordination

Guide 3.2.1: **Frame owner.**
Every frame is assigned to a dedicated organizational unit within the NSI called the frame owner.
Minimum requirement(s):

- The frame owner is in regular contact with all internal contributors relevant to the maintenance of the frame.
- The frame owner coordinates activities with the owners of other frames functionally or conceptually related.

Guide 3.2.2: Responsibilities of relevant bodies.
Responsibilities of all relevant bodies contributing to maintenance of the frame are well assigned.
Minimum requirement(s)

- A unit in charge of methodological issues is responsible for
  - sampling operations
  - methodology for linking.
- Legal experts take care of aspects of data protection.

Guide 3.2.3: Accessibility of **data from external suppliers.**
Accessing the data coming from external suppliers is anchored in national legislation**.**
Minimum requirement(s):

- The access to administrative sources is granted for the NSI in accordance to regulation 223/2009.
- The access to administrative data is anchored in law.

Guide 3.2.4: Cooperation with external data suppliers.
There are regular meetings and formalised cooperation with external data suppliers.
Minimum requirement(s):

- There are agreements in place about regular data delivery regarding:
  - Time of delivery
  - Method and format of delivery.
- Legal experts take care of aspects of data protection.
- IT experts are involved in all matters concerning the technical implementation of the frame.

Guide 3.2.5: Work flows and protocols regarding all relevant processes.
There are defined work flows and protocols when a use case for a frame is triggered.
Minimum requirement(s):

- Experienced experts are involved in all sampling aspects.
- All sampling aspects are approved by methodological experts.
- Legal experts are involved in all legal questions. Particularly all data protection issues are addressed.
- Regular training is accessible for staff when the work flows and protocols are introduced.

Guide 3.2.6: Data protection and **statistical confidentiality** issues**.**
Special emphasis is laid on data protection and statistical confidentiality. Ideally the degree of data protection is commensurate with the sensitivity of the data.
Minimum requirement(s):

- Data protection enjoys a first priority.
- Names of individuals are deleted from used frame data as early as possible.
- A safe repository containing names and addresses of individuals is maintained where only a limited team is allowed access.
- It is prevented that statistical units can be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all relevant means that might reasonably be used by a third party to identify the statistical unit.

3.3 Methods for Constructing Frames

This chapter covers the **initial construction of frames** in social statistics, considering methods used to meet the target of having a dataset continuously maintained over time consisting of **households and persons**. For each household **an address** should be available, which means that there should be a relationship to a **dwelling.**

Households can be seen as the most prominent example of a composite unit in social statistics. However, it should be noted that the quality and availability of certain data sources is fundamental in order to allow the composition of households. A second aspect relates to the fact that the definition of household is not unique when looking to different statistical applications.

For example for LFS we do have:

"A private household (housekeeping unit concept) is either:
1.a one-person household, i.e. a person who lives alone in a separate housing unit or who occupies, as a lodger, a separate room (or rooms) of a housing unit but does not join with any of the other occupants of the housing unit to form part of a multi-person household as defined below, or
2.a multi-person household, i.e. a group of two or more persons who combine to occupy the whole or part of a housing unit and to provide themselves with food and possibly other essentials for living. Members of the group may pool their incomes to a greater or lesser extent.

Slightly different the definition provided by the US Census 2000:

A household includes all the people who occupy a housing unit. (People not living in households are classified as living in group quarters.) A housing unit is a house, an apartment, a mobile home, a group of rooms, or a single room occupied (or if vacant, intended for occupancy) as separate living quarters. Separate living quarters are those in which the occupants live separately from any other people in the building and that have direct access from the outside of the building or through a common hall. The occupants may be a single family, one person living alone, two or more families living together, or any other group of related or unrelated people who share living quarters.

These two examples illustrate how broad household concepts – such as socio-economic links or co-residence – may be varied and/or combined to establish explicit household definitions for particular purposes. A general comparison of concepts of private households in social surveys can be found in Hoffmeyer-Zlotnik and Warner (2008).

Given this the construction and integration of households depends very much on the data situation and the related use cases. To sum up we say that as a minimal requirement for a frame in social statistics should contain either persons or dwellings. However, it is strongly recommended to integrate household as a composite unit whenever possible. It should be noted that in the context of this document **we do not provide an exact unique definition for household.** Whatever definition of household is taken the concept can - and with high likelihood will - affect survey procedures (e.g. probability of selection, weighting, non-response of sampling units etc.).

There are different scenarios when a frame is created. A huge external source (for instance a central population register) can be used to fill the frame initially. Later on this **initial backbone** is then supplemented by other satellite sources. Another scenario would be that you link several input sources. However, it could be useful to set up the process in a stepwise manner which means to declare a certain set of records as your initial data set.

Further, you will have the ambition of integrating at least a minimum set of variables. In social statistics this might be the **social core variables**. Besides this you will need a set of **contact variables.** Additional items allowing you to structure the population of interest for a certain survey might be helpful. Generally, it can be said that **enriching the frame as much as possible** with additional information - having in mind tasks like producing statistics out of the frame, supporting statistical processes and designing samples - is one of the primary goals when constructing a frame. A significant importance should be laid on **identifiers** (often called key variables) assigned to the statistical units as they play an important role in deduplication of different sources.

As mentioned in chapter 3.1., most of the time you may have a set of different sources when constructing the frame. Based on that the problem of constructing a frame can be seen as **combining information from different sources to a data set (database)** later on used as a frame.

With this in mind, **record linkage** is a key step used for constructing a frame. When applying record linkage the most convenient case is the availability of a **common identifier** which enables the producer to perform exact linkage. So having a common identifier (household number, person number, dwelling number) for every unit of the source is an ideal situation. However even then the result will not be a matching rate of 100 %. It can be expected that a certain amount of false links occur, i.e. units incorrectly linked when in reality they are not the same units. Therefore a strategy **for dealing with non-linkable records and false links** has to be developed.

When it comes to taking in a certain variable relevant for the frame, you may find that it can be drawn from more than one source. If the value for the variable varies in different sources you have to apply a kind of rule on how to decide which value you trust more. The application of **decision rules** leading to most trustworthy results may therefore be necessary. Sometimes you will only have information about a variable in an aggregated level. Therefore the application of **disaggregation methods** could be necessary. For instance, if you have a certain income component only on household level it might be useful to assign a value to each person living in the household. This can be done by applying disaggregation.

Sometimes when you try to enrich a frame by a certain variable you will not succeed to full extent. This gives the classic situation of missing values leading to the possible use **of imputation or statistical matching. Purely estimating values on a model basis or mass imputation shall only be applied in exceptional cases.**

During the construction phase of the frame, it is important to check if the resulting records are plausible and consistent. Therefore it is necessary to apply methods used in Statistical Data **Editing**. This holds for the micro as well as for the macro level. Implausible or incorrect values might be dealt with by **imputation** later on.

Challenges

Challenge 1: The choice of methods for constructing a frame will have an impact the output and the quality of the statistical products relying on the frame. Insufficient methods, or not using best or even good practices, can distort **accuracy** in all derived statistical outputs**.** Data integration is not only a potential source of errors of false non-links and false links, these types of error may also cause coverage and measurement errors.

Challenge 2: Not being able to enrich the frame by a sufficient set of variables can cause a problem when the frame is used for **sampling and estimation.** The problem can occur that samples cannot be designed in an efficient way or furthermore it might be the case that calibration procedures, when weighting a sample, cannot rely on good enough auxiliary information. Lack of completeness by not applying good practices, can lead to **coverage errors and/or misclassifications.**

Quality guidelines 3.3

Quality guidelines – Methods for Constructing Frames

Guide 3.3.1: Distinction of statistical units to the corresponding frame units.
When planning a frame in social statistics the question of the statistical units is addressed. The possible distinctions of statistical units to the corresponding frame units are specified.
Minimum requirement(s):

- A frame in social statistics contains at least dwellings or persons and their contact addresses.
- Households as composite units are integrated when the situation and the relevant use cases allow doing so.
- If a frame is used as direct source for statistics the units relevant for this statistical product are mapped.

Guide 3.3.2: Relationships between households, dwelling and persons.
The relationship of households, dwellings and persons is clearly mapped in a hierarchical way by corresponding identifiers.
Minimum requirement(s):

- Every household can be related to at least one dwelling. A household may have several dwellings, where at each moment in time one is the main residence.
- A dwelling is a housing unit that has some but not all of the essential features of a conventional dwelling. A dwelling can contain one or more households but can also contain no household (vacant dwelling).

Guide 3.3.3: Set of variables in the frame.
Before constructing a frame, the set of variables contained in the frame is predetermined. Only those variables which are covered by the available sources are included.
Minimum requirement(s):

- The quality of the contact variables in the frame allows contacting the sampling units easily when drawn into a sample (addresses, telephone numbers etc.).
- The uniqueness of identifiers for persons, households and dwellings is guaranteed.

Guide 3.3.4: Initial data set.
A starting point for constructing the frame in the form of an initial data (often called backbone) set is defined.
Minimum requirement(s):

- The initial data set is prefilled by a trustworthy source.

Guide 3.3.5: Linking by common identifiers.
The linkage procedures are based on a common identifier.
Minimum requirement(s):

- National legislation allows linking of sources for statistical purposes.

Guide 3.3.6: No identification by common identifiers.
Minimum requirement(s):

- The common identifiers do not allow disclosing (directly or indirectly) the identity of single individuals.

Guide 3.3.7: Linkage of units.
If it is not possible to link all units by using one common identifier, an attempt can be made to apply probabilistic record linkage methods based on text comparisons between, for example names and addresses. As a last resort, attempts can be made to find as similar units as possible by means of statistical matching, e.g. a nearest neighbour based on auxiliary information available in both sources.

Guide 3.3.8: Mapping of sources to the frame.
If the units in different sources are different, especially if they are on a different level, then it is clear how this is taken into account in the target data set. If it is necessary to disaggregate (e.g., break down household variables to the person level), the models used are documented.

Guide 3.3.9: Traceability of **values of variables in the frame** to the source.
When different values of variables can be found in several sources, there are clear procedures to decide which value is used.
Minimum requirement(s):

- The trustworthiness of a source is taken into account.
- Methodological experts are involved.

Guide 3.3.10: Use of **guidelines and methods**.
When constructing the frame all existing good practices are respected and state-of-the-art and scientifically evaluated methods are used.
Minimum requirement(s):

- Existing European international and national guidelines are respected.
- A complete set of metadata for every source is available.
- Methodological experts are involved.
- Methods in use are tested and these tests are documented.

3.4 Possible outputs when constructing frames in social statistics

When constructing a frame for statistical purposes you might have in mind a certain output. Basically what you try to achieve is a complete list of all persons living in the country (or owing citizenship to the country), including a possibility to contact them. Further you try to map certain relations at an aggregated level (households, dwellings). In some cases the situation does not allow for the creation of a frame of the units you would like to access for the envisaged use case, which means that a **direct list frame** cannot be created. Given this, it might be possible to create another frame of units that is indirectly related to the target population. It can then be considered to produce an estimate for the desired target population by using the links existing between these two. In this case we talk about an **indirect frame.** An example is an estimate with the target population of young children, where the available frame includes only a list of their parents. In this case, the selection process happens in two steps. First a sample of the parents is selected from the frame. Second, a sample of their children is selected. There are several scenarios for constructing frames, depending on the intended use case. Sometimes it might be sufficient if you take only information from one **single frame** to cover all requirements relevant for a use case. In other cases it might be necessary to involve more than one frame. Inferences about the target population are then based on the combined sample data. If two frames are involved, you call it a **dual frame scenario.** If you have more than two frames, you can talk about a **multiple frame use case.** The decision to use several frames can be based on several factors, such as cost considerations, the mode of data collection or the need to have a better coverage of special subpopulations. For example, if a complete frame, such as a person list, is available, it can still be more cost-effective to take a sample of reduced size from the complete frame and enhance only the sample units with additional information from another frame, such as telephone directories.

When coming to the very important aspect of units included in the frame, referring to the definition provided in chapter 2.1, it should be clear which kind of units the frame should have. Again, depending on the use case, there is the possibility of having only one kind of unit (list frame) or you can, by including composite units arrive at an **integrated frame**.

The frame aims to map the target population not only in terms of contents and units. It should also relate the population to certain time periods. This means that. In relation to reference periods (a certain year, month or quarter) all statistical products shall use the same **frozen frame**.

Another basic question is whether you should head for a **master frame** serving all statistical products assigned to an area of social statistics or for more than one frame. Having in mind coherence and consistency of definitions the scenario of one frame for all social surveys seems to be the better solution.

Sometimes NSIs use a **master sample**. A sample drawn from a frame covering a target population for use in a number of future occasions, so as to avoid ad hoc sampling on each occasion. Sometimes the master sample is large and subsequent inquiries are based on a sub-sample from it. The advantage of a master sample is that you have a well-defined set of statistical units which can be suitable for instance for panel surveys. On the other hand a master sample is a preselected set of units, which might cause bias when used as source for sample survey.

It is very important to enhance a frame with as much information as possible. This concept of a **rich frame,** where individuals, households and dwellings are augmented with information from other registers allows optimized sample designs and the possibility of providing input for statistical products.

Challenges

Challenge 1. The output of frame construction is dependent on the use case. If there are many use cases for frames serving different purposes, this can lead to coherence problems.

Challenge 2. Many problems can become evident only after the frame is used or during fieldwork. It might be difficult to obey everything during the planning phase.

Quality guidelines 3.4

Quality guidelines - Possible outputs when constructing Frames in social statistics

Guide 3.4.1: Use cases.
Before constructing a frame, the use cases for a frame predetermine the output (constructed frame).
Minimum requirement(s):
- If the frame is used as sampling frame, the main focus is laid on contact variables.
- If used for direct tabulation the focus is laid on key variables relevant for the statistical product.

Guide 3.4.2: Availability of **designated output.**
After constructing a frame a designated output (constructed frame) is made available to all users.
Minimum requirement(s):
- All relevant users of the frame are informed about its availability.

Guide 3.4.3: One frame for all products in social statistics.
The existence of one frame used for all products in social statistics shall be the preferred solution. This frame is enhanced by as much information as possible (rich frame).
Minimum requirement(s):
- The construction of special frames for single statistical products in social statistics is only constructed in justified cases.

Guide 3.4.4: Content of the **output.**
The output consists of a complete list of individuals adequately mapping the population of interest, (e.g. either persons currently living in the country or being citizens of the country). Further the relations in households and to dwellings are mapped.

3.5 Updating Frames in social statistics

One of the basic intentions of a frame in general is that it is used over longer period in time. The use is typically integrated into the production process relevant for certain statistical products. Therefore, the quality of these products relies very much on the functioning of the update processes for the frame. The central objective of the update process is the provision of a frame which is most accurate and actual in comparison to all populations covered by the frame. It is worth mentioning that, for the sake of coherence some NSIs strive for the provision of "frozen frames" where a certain set of statistical products refer to. This kind of frame is related to a certain time period (monthly, quarterly or yearly). In the case of social statistics this means that the set of persons - called the population of the country - is mapped by the frame as closely as possible to the real life situation regarding all variables identified as relevant during the construction of the frame. It seems important to distinguish between **regular updates** of the frame as part of the normal production process and updates due to some kind of bigger redesign based for instance on the integration of a new data source. We call this a **revision of the frame.** To conduct a revision it is necessary to know about the situation of the frame and therefore the update procedures and the quality of the frame should be assessed on a regular basis.

Figure 4: The process of updating a frame

As the graphics show the regular process of updating a frame can be separated into three different parts. As a first step, the inputs for the update process have to be organized and prepared to be ready for the update. One central goal is to secure the regular delivery of all input data not only by internal but as well by external data providers contributing to the update process. Very often, it might be the case that certain administrative agencies holding a **central population register** play an important role, often serving as backbone of the frame in social statistics. Since the administrative processes associated with such a register are conducted in a continuous way, the most desirable situation would be that the NSI has a direct access to it. As a general message it can be said that the delivery of data from the backbone register (if such a register exists) should be secured. Besides information coming from outside it should also be information coming from inside the NSI. First of all in social statistics there are several statistical inputs which have to be obeyed. In this regard a frame shall be updated using **information from birth, death and migration and some other possible demographic events (wedding, divorces, etc.)** on micro level. Statistics of these demographic incidents are part of the central obligations of a national statistical system.

The process of updating a frame – no matter if it is a regular update, an event-driven revision of the frame, or if it involves the update of the underlying backbone register such as the population register – often involves the attempt to "clean up" the frame (or the underlying backbone register), mostly in the sense of fighting over-coverage of the population. Over-coverage occurs, for example, when the frame should cover the residential population by definition, but the frame (or the underlying central population register) is unable to trace accurately the emigration of persons to another country**. A "signs of life"-analysis** can help to detect these supernumerary cases. This "signs of life"-analysis involves information from additional registers, such as the birth and death register, the tax register,

unemployment register, register of social welfare recipients, etc. This information from additional registers has potentially already been included as auxiliary information in the frame in the first place, but it is also possible that the signs of life analysis includes information from registers which have not played a role in the frame before and which are only included as source for the signs of life analysis (think e.g. of a register of driver licence renewals). Some variables included in the signs of life analysis will unequivocally determine the status of a frame unit (most prominently the information from the birth and death register), but it is also possible that these additional variables only give hints about the status of a frame unit. Some NSIs, such as the Estonian statistical institute, also build models combining several variables (from several additional registers) to come to a reliable decision if a person counts as supernumerary case and is therefore deleted from the frame. The Estonian model is based on a weighted sum of signs of life and assigns each person a probability of "being alive" (in the sense of being correctly in the frame) – see as well here Lehto, Maasing Tiit (Q2016.) The consequences of identifying a frame unit as (potentially) supernumerary vary. It is possible that the NSIs send a letter of clarification to the resident in question or to the respective municipality, but also the deletion of the unit in question from the frame without any further consultation is possible, given that the "signs of life analysis" provides good enough evidence.

Also, the feedback from the users of the frame can and should be integrated in the frame for future use. For instance, if a survey had determined that some variables coming from the frame are different than expected, then this should be considered to be used for an update, provided that this information is considered as more trustworthy.

The methods used during the update processes will be very similar to the ones applied during the construction. However, it is important to **assign clear responsibilities** during this step to guarantee that all necessary applications, IT-programs or IT jobs are going to run. The work flow of the regular update process should be documented in a comprehensible way.

Finally, after the processing for the update is complete, the **check of the outputs** will be decisive. It is important to generate control tables showing the number of persons, households and, if feasible, dwellings or addresses in order to check if the resulting output is correct. Additionally, whenever possible, automatized plausibility checks should be implemented in the programs after the update. Sometimes a provisional frame (or even a series of provisional frames) related to a certain population and/or time period is provided before a final frozen frame is released.

As mentioned above, in addition to the regular update procedures there could be more general updates – already called revisions - based on the fact that there is a need to redesign the frame. In most of the cases, the integration of a new data source found relevant for the frame will make it worth revising the frame. In this regard it seems very important to assess the situation regarding updating procedures and the **availability of new data sources** regularly.

Challenges

The challenges below deal with the impact on quality. As it regards the direct use of statistics more information can be found in chapter 4.3

Challenge 1. Update procedures are necessary to comply with **actuality**. Therefore not having well defined and clear update procedures can lead to a lack of consistency to reality in terms of not having up-to-date data.

Challenge 2. By not having up-to-date address data or other contact data the rate of non-contactable persons or households in a sample survey can either decrease the realized net-sample, which means an increase in sampling errors, or create the necessity of increasing the sample size, which will yield additional costs.

Challenge 3. Not performing updates in an appropriate way can lead to **coverage errors and/or misclassifications.**

Quality guidelines 3.5

Quality guidelines – Updating Frames in social statistics

Guide 3.5.1: Frame updates.
There are regular updates of a frame which are based on the statistical production needs and on data deliveries of all relevant internal and external providers

Minimum requirement(s):

- The update periodicity takes into account the internal needs, the availability of the sources and the costs.
- The update should be performed quarterly at least for the contents where information is available.
- If there is an externally maintained backbone register (for instance a central population register), the data delivery is ensured by a formal agreement.

Guide 3.5.2: Organization and documentation of the update procedure.
The regular update procedure shall be organized and documented in a comprehensive way.

Minimum requirement(s):

- There is one organizational unit responsible for the regular update procedure.
- Responsibilities for single process steps during updating are clearly assigned.

Guide 3.5.3: Completeness of the **update procedure** regarding **sources.**
The regular update procedure incorporates all relevant internal sources for updating the frame

Minimum requirement(s):

- Input from birth, death and migration statistics on micro data level is provided for the update of the frame is used
- There are regular feedback procedures from surveys resulting in the incorporation of relevant information into the frame.

Guide 3.5.4: Completeness of **the update procedure** regarding **statistical units.**
The regular update procedure takes care of all statistical units in the frame, as well as of all links between different units and of all variables included in the frame. If feasible, a sign of life analysis is conducted.

Minimum requirement(s):

- The update procedure relates to persons, households and dwellings.
- Links between households and persons, dwellings and persons are updated.
- Special emphasis is laid on the update of contact variables.

Guide 3.5.5: Results of the update procedure.
Before closing the process of regular updating, resulting outputs are checked.

Minimum requirement(s):

- Control tables regarding persons, households and dwellings are produced during the update process.
- Checking results of update is done by obeying the four eyes principle (the four eyes principle is a requirement that two individuals approve some action before it can be taken. The four eyes principle is sometimes called the two-man rule or the two-person rule).

Guide 3.5.6: Assessment of the update procedure.
The regular assessments of the update procedures take into account potential new data sources.
Minimum requirement(s):

- Relevant users, methodological experts, all data providers and quality management are included in the assessment process.

Guide 3.5.7: A major update of the frame.
A major update or revision of the frame after an assessment of the update procedures or of the frame in general is handled without distorting regular production.

Guide 3.5.8: Information about an updated frame
All relevant users of the frame are informed about the availability of an updated frame.

4 Use of frames in social statistics

Many statistical agencies have access to high quality administrative data, which can be used to supplement, augment and even replace survey data. Administrative data sources and frames built on such sources can be used both as a direct source for statistical data or as an auxiliary source to enhance and improve the quality of survey data. Currently, the main uses of frames in social statistics are:

- Direct use of the frame data to obtain statistical results,
- Defining parameters for designing samples and size variables for drawing the samples,
- Auxiliary data for calculating sample weights based on calibration estimators,
- Covariates for modelling synthetic estimates that can be used in small area estimation,
- Auxiliary data to adjust for nonresponse.

4.1 Sampling

4.1.1 Sampling Frames

In addition to the already described characteristics of a frame, a **sampling frame** has to allow for determining the inclusion probability for each unit into the sample under the chosen sampling design. Such a unit becomes a possible sampling unit. Statistical results should be based on probability samples drawn from sampling frames that allow persons, households or dwellings to be selected at random, with a known inclusion probability. This requirement can only be achieved if a high quality sampling frame is used for sample selection.
Several aspects and types of sampling frames are considered in the following:

- the processes assigned when sampling form a frame,
- the type of the units listed in the sampling frame,
- the quality of the contact information of the frame units,
- if one sampling frame is sufficient for the whole sampling procedure or if the use of multiple frames is more advantageous,
- if the frame is used as master sampling frame and
- if a master sample is drawn.

When planning a sample, the main objective is to minimize the variance of key estimators by realizing a sample design in order to comply with certain accuracy demands. So, regarding the **processes related to sampling,** the frame can be used to estimate a-priori variances for estimators to come to a first decision on a possible sample design. First of all a rough estimate of the sample size necessary which is a decisive element for the costs, can be discussed. Secondly, the frame is used for realizing sample designs by simulating a possible survey program in the background of the use of various sample designs. **Stratification** is a frequently applied procedure to lower the standard error of the estimator or to get estimates of population parameters for groups within the population (for more information about stratification procedures, see e.g. Cochran, 1977, Särndal at al. 1992). A rich frame, including useful stratification variables, is a precondition for every survey design including stratification.

Multi-stage sampling, where sampling units at each stage are sub-sampled from the composite units chosen at the previous stage require rich frames. A prominent example is given by sampling of a household in a first stage, and then surveying some (or sometimes all) members of this household. Other examples include pupils as survey unit, where first, the school is sampled. In all these cases, not only the contact information of the frame unit itself is

needed, but additional information about previous stages of a multi-stage survey design is needed also (e.g. the household, the school, etc.).

As explained in chapter 2, the sampling frames are generally of two types, area frames and list frames. A **list frame** is a list of units in the survey population. **Area frames** are usually made up of a hierarchy of geographical units which in turn contain units in the survey population. That is, the frame units at one level can be subdivided to form the units at the next level.

It is the most preferable situation if a frame of population units is based on a complete register such as an administrative population register. Electoral lists, census lists or other data sets are possible as a basis.

Another example of a list commonly used for sampling frames is the telephone directory. This list is usually not complete because of unlisted telephone numbers, households without telephone and because of the increasing number of mobile phones often not included in the directories. This is an example of frame under-coverage. A telephone directory can allow inclusion of some households with more than one telephone number and this is an example of multiple listings. When one of these two kinds of lists of individuals is used for the construction of a sampling frame, a one-stage sample design can be used, in which each final sampling unit is directly selected from the frame. Of course, even when a population register is available, a two-stage sampling design can be applied as well, which is done in some cases for efficiency gains in the survey process.

In some countries, however, a complete list of all the persons in the target population may not exist. In this case an area frame can be used as a geographical frame consisting of area units. Every target population element belongs to an area unit and can be identified after inspection of this area unit. In such a case, the total population size may be unknown. Clusters are the first stage sampling elements drawn from the area frame. The selected clusters are sub-sampled in a secondary selection step.

The quality of the contact information of the frame units has an impact on the accessibility of the sampled units. A good sampling frame should also provide sufficient information to be able to contact the selected units and uniquely identified the location (precise and up to date address). Failure to do so can result in distortions of the selection probabilities and of the sample structure because some units cannot be contacted and their inclusion probabilities become equal to zero. If the erroneous addresses cannot be corrected, they can be attributed to the frame under-coverage. For further information on contact variables, see chapter 4.1.2.

A multiple frame survey is a sample survey which is based on multiple sampling frames.

There are certain advantages associated with using multiple frames. The most important ones are listed below:

- use of administrative records more efficiently;
- useful for multiple mode sampling (for example, using independent samples from a cellular telephone frame and a landline telephone frame);
- can be used for future use of the internet for data collection. Although the internet presents many coverage challenges, it is worthy of consideration because of the potential cost savings and ease of data collection and processing;
- may improve small area estimation. A national survey is supplemented with smaller, localized surveys to obtain higher precision in those areas;
- may improve estimation for rare populations. A general population survey may be supplemented by a survey from a frame with a high concentration of members of the rare population. Multiple frame surveys can also be used in conjunction with sequential or adaptive sampling methods to improve yield of a rare or hard-to-reach population such as recent immigrants;
- can give more flexibility for design of continuing surveys. As particular frames become less expensive to sample, the relative allocation of sample size to the different frames can be modified. The modular approach also allows more flexibility in responding to changing needs for data.

Two main cases are considered here in more detail. In the first case, all involved frames are list frames, and they may be overlapping. In the second case, one frame has a different aggregation level than the other one(s). This includes the joint use of an area frame.

MF1. The multiple frames consist of the union of list frames for the same survey population. It may have such a structure: one frame (inexpensive to sample) is a subset of another larger and complete frame; two frames that are incomplete with non-empty intersection; more than two frames that are incomplete with non-empty intersection (Lohr, 2011)

A sampling frame must represent, in a statistical sense, a survey population. Sometimes, coverage of a survey population by one frame is poor. The coverage of the survey population by the *union of several frames* may also be not complete, but much higher, and the coverage error may be lower. A principal frame that provides nearly complete coverage of the target population may be supplemented by a frame that provides better or unique coverage for population elements absent or poorly covered in the principal frame. For example, an out-of-date set of listings of housing units can be supplemented by a frame of newly constructed housing units obtained from planning departments in governmental units responsible for zoning where sample addresses are located.
At times, the supplemental list frame may cover a completely separate portion of the population. It happens when the target population resides in different kinds of living quarters that are non-overlapping. For example, a survey of orphans would most likely be designed to include orphans living in institutional arrangements such as orphanages as well as those living with relatives in a household. In most cases, though, supplemental frames overlap with the principal frame. It depends on the country and the survey specific circumstances, which of the frames is considered as principal and which is seen as the supplementary frame.

MF2. Area frames together with list frames can be used *for a multistage sampling procedure.* (Turner (2003), Petterson (2003)). Multistage sampling design is used in order to reduce travel costs of the interviewers on a country side, to distribute efficiently the work load for interviewers or when a good population sampling frame is not available. For example, the sample design for a household survey can use both an area frame for the early stages and a list frame, for the last stage. Dual-frames or multiple frames designs can be used then.

A **master sampling frame (MSF)** is one frame which is used to select samples either for multiple surveys, each with different content, or for use in different rounds of continuing or periodic surveys. The sampling frame itself does not vary either from one survey to the other or from one round to another of the same survey. Instead – and this is its distinctive characteristic - the master sampling frame is designed and constructed to be a stable, established aspect for selecting sub-samples (subsampling) that are needed for particular surveys or rounds of the same survey over an extended period of time (Turner, 2003; Pettersson, 2003). MSF enables selection of different samples (including from different sampling designs) for specific purposes: individual surveys, household surveys. The MSF's distinguishing feature is that it enables samples to be drawn for several different surveys or different rounds of the same survey, which makes it possible to avoid building a special frame for each survey. MSF is a frame or a combination of frames that covers the population of interest entirely, and that enables the linkage of the individual to the household as a social unit, and both of these to the dwelling and address as an environmental unit.

It can be the case that a one stage sampling design makes use of a MSF, but it can also be the case that only the frame of the first stage of a multi stage sampling design is a MSF. In the latter case, the MSF consists of the primary sampling units, which may refer to some geographical units such as block listings or area segments.

The very diverse examples have in common the savings from sharing the start-up costs of design, stratification, listing, etc., for constructing the specific sampling frames.

The national sample survey programme defines the demands on the master sampling frame and the sample design applied to MSF in terms of, for example, the anticipated number of samples, population coverage, stratification and sample sizes. It depends on the conditions for frame construction in the country, how these demands should be met in the design work. The most important factor is the availability of data that can be used for frame construction.

From a master sampling frame, it is possible to select the samples for different surveys entirely independently. However, in many cases, there are substantial benefits resulting from selecting one large sample, a master sample (MS), and then selecting subsamples of this master sample to service different (but related) surveys. Several NSIs have decided to develop a master sample to serve the needs of their household surveys.

A **master sample** is a sample from which subsamples can be selected to serve the needs of more than one survey or survey round.

The terminology of "Master Sample Frame (MSF)" and "Master Sample" can be a bit confusing, especially if a multi-stage sampling design is involved. The defining characteristic of a master sample frame is that it serves as sampling frame for several samples (for different surveys or for different survey rounds). The defining characteristic of a master sample is that it is a sample, and different subsamples (for different surveys or different survey rounds) are drawn from it. It is important to keep in mind that not every sample drawn from a master sample frame is a master sample.

In the case of a multi-stage sampling design, and the use of a master sample frame, it is often the case that the sample drawn in the first stage fulfils the characteristics of a master sample.

There are several advantages of adopting a master sample design. It reduces costs of developing and maintaining sampling frames as more survey units share the same master sample design. It also simplifies the technical process of drawing individual samples by facilitating operational linkages between different surveys.

The use of a master sample has the following properties:

- It is cost efficient and it makes it possible for the NSI to distribute the share of costs of construction of a sampling frame between several surveys. Costs of preparing maps and subsampling frames of dwelling units or households will be shared among the surveys using the MS.
- Clear gain from using an MS in the case where interviewers need to be stationed in or close to the primary sampling unit (PSU) owing to difficulties and high costs related to travel in the field.
- Use of the same master sample PSUs for several surveys will reduce the time it takes to get the surveys started in the area and also the time it takes the interviewer to find the respondents.
- The MS facilitates quick and easy selection of samples. Subsamples from the MS can be selected quickly when needed for ad hoc surveys.
- The MS makes it possible to have overlapping samples in two or more surveys and provides a basis for integration of data from the surveys.
- Quality will usually be better than in the case of special sampling frames because it is easier to motivate investments in quality improvement in a frame that will be used over a longer period.
- Simplifies the technical process of drawing individual samples and facilitates quick and easy selection of samples for surveys of different kinds.

Challenges

Challenge 1. The realization of certain sample design demands the availability of the corresponding information in the frame.

Challenge 2. Bias for estimates of the population parameters may arise due to the population **under-coverage**. Therefore, the size of the under-coverage should be estimated.

Challenge 3. Multiple listings may result in appearance of some target population units in the frame more than once, giving them a larger probability of selection than intended. This needs to be estimated and taken into account when selecting the sample.

Challenge 4. Sampling frame information imperfections, not only coverage errors but also out-of-date information, are likely to bias or reduce the reliability of the survey estimates and to increase data collection costs. For example, over-coverage generally increases variance because it results in a reduced sample (elements which do not belong to the target population are excluded), compared with what would have been obtained under no over-coverage.

Challenge 5. Like all surveys, **multiple frame** surveys are subject to non-sampling errors:

- *Inadequate* survey population *coverage* by a sampling frame is a potential problem in a social survey, *erroneous units* included and unit *duplicates* are usually in the frame but they are often not known.
- While the union of the frames may have better coverage than a single frame, there may still be under-coverage of the target population to be taken into account.

- Estimators based on multiple frame surveys are sensitive to domain misclassification and biases that might result from different administration methods or modes in the different frames.

Quality guidelines 4.1.1

Quality guidelines – Sampling Frames

Guide 4.1.1.1: Quality of **a sampling frame.**
Before being used for concrete sampling, the quality of each sampling frame is checked and assured.

Minimum requirement(s):

- There is enough information to realize the envisaged sample design.
- The quality of contact variables are a main focus.

Guide 4.1.1.2; Stratification variables.
Special emphasis is laid on the maintenance of frame variables periodically used for stratification in samples.

Guide 4.1.1.3: **Multiple frames**
To increase the coverage of the population, the use of multiple frames for sampling is considered. Multiple frames might also be used as a part of survey design that relies on different sampling frames to help to reduce costs.

Guide 4.1.1.4: Probability samples.
When two or more overlapping sampling frames are used, a probability sample is drawn independently for each frame. If two frames are used in the construction of a final frame, and one of them just adds the new variables to the backbone, then the sampling design is applied to the final frame.

Guide 4.1.1.5: Use of **primary stage units of a master sample**.
Primary stage units (PSUs) of a master sample are used for a long time in social surveys. Therefore it is important to take care in maintaining a master sample in an appropriate way

Minimum requirement(s):

- Updating the stratification and population measures of size, and re-selecting primary sample units (PSUs) is conducted whenever necessary.

Guide 4.1.1.6: **Variables to define aggregates**
In the case of multistage sampling designs, the variables to define the aggregates in the different stages are updated periodically.

Guide 4.1.1.7: Data and materials.
When designing a master sampling frame (MSF) the available input data are assessed with respect to the quality needed, the NSI decides on the key characteristics of the MSF (area units, frame units, quality assessment, characteristics).

4.1.2 Contact variables

The main purpose of a sampling frame is to receive a sample to be handled during fieldwork. So the most important prerequisite is having decent contact variables which allow all survey units included in the sample to be reached. Technology means that today there a several possibilities (modes) to conduct a survey. Besides face to face – mostly, supported by laptops -, telephone interviews and web surveys are widely used, with the latter gaining a lot of traction in recent years. Different contact variables are needed, depending on the mode. As pointed out in chapter 2, a unique identifier should be assigned to each single unit of the frame. However, this identifier does not help in contacting the unit. Again, referring to chapter 2, we said that there is a reference to a dwelling. The dwelling has to be connected to something practical called the **dwelling address,** as one of the most important contact variables. Having a face to face mode (or paper-pencil questionnaire which is nowadays not very common in social surveys) this is sufficient for contacting the respondent. When relying on a telephone survey the telephone number will be the crucial contact information. Last but not least, since electronic survey tools like web surveys - CAWI (computer assisted web interviewing) - are becoming more and more common, electronic contact information, mainly email address, is getting more and more important. So ideally a frame in social surveys contains the following contact information.

	Person	Household	Dwelling
Address	X	X	X
Telephone number	X		
Email address	X		

As the table shows besides the mode, the unit of interest in your survey is decisive for which contact information you have to use. If you try, for instance, to survey a complete household it might be sufficient to have the phone number of one person in the household in order to contact the unit of interest. It should be mentioned that having only one address/phone number/email is in some ways an ideal situation. In reality it might (and will be) the case that there is more than one contact information for a unit. In the case of address which also provides the regional information on the lowest level there should be a mechanism to select the main address to distinguish from other ones, e.g., side residences.
To assess the quality of the contact variables in a frame, it is important to observe the contact rates while conducting the fieldwork of surveys. If there are a high number of units, where it was not possible to contact the unit due to missing or wrong contact information, one should take action to improve the situation.

Challenges

Challenge 1. It can be the case that the frame itself does not contain the current contact information directly. Therefore it can be necessary that a specific process after drawing the sample needs to be in place to obtain the contact information for the drawn units. However, since this process might involve names of persons, data protection and IT safeguards play an important role.

Challenge 2. Sometimes it might be difficult to achieve an unbiased sample for telephone surveys due to the fact that the quality of registers regarding phone numbers is decreasing. Sometimes there are (several) external commercial registers which you have to access here. Subgroups in the population exist who only have mobile phones. Further, there is some subpopulation not willing to register the phone number. This can cause bias in your sample and cause a bias in the survey results. This situation might be improved by using mixed mode data collection.

Challenge 3. It can be the case that the process used to update your frame with respect to address information is not quick enough to map the administrative reality. Sometimes, a NSI receives data deliveries from the central population register on a quarterly basis which means that you will have a certain share of contact information not completely up to date.

Quality guidelines 4.1.2

Quality guidelines – Contact variables

Guide 4.1.2.1: Existence of **contact information in each unit**.
Each unit in the frame has contact information. This includes address, phone number and email address.

Minimum requirement(s):

- Each dwelling has at least the postal address assigned.

Guide 4.1.2.2: Checking of sample.
After contact information was assigned the resulting output is checked to see if it still coincides with the sample selected according to the original sampling design.

Guide 4.1.2.3: A process to include contact information in the sample.
If contact information (e.g. phone numbers) is not integrated directly in the frame, there is a well-defined and specific process to include it to the drawn sample.

Minimum requirement(s):

- The assignment of contact information is done by one central unit.
- The process respects all legal requirements concerning data protection.

Guide 4.1.2.4: Storage of complete contact information.
If applicable, it is possible to store more than one contact information for a unit.

Minimum requirement(s):

- If more than one address is assigned to each unit one address is assigned as main address of the unit.

Guide 4.1.2.5: Briefing of **interviewers.**
Interviewers are informed that they are only allowed to use the contact information for the purpose of the survey.

Minimum requirement(s):

- This aspect is addressed in regular interviewer briefings.

Guide 4.1.2.6: Quality of **contact information**.
The quality of the contact information is assessed on a regular basis.

Minimum requirement(s):

- Every survey delivers non-contact rates and reports back the units, where the contact information was different from the expected one.
- There are regular meetings with external providers of contact information.

Guide 4.1.2.7: Landscape of administrative and other sources
The landscape of administrative and other sources is observed in order to integrate new sources of contact information.

Guide 4.1.2. 8: Maintenance process of frames takes care on contact information
The maintenance process of frames puts special emphasis to the update of the contact information.

4.2 Frames as input for processing

4.2.1 Frames supporting editing and imputation

Data cleaning -"editing and imputation" - is an essential step when producing statistics, the goal of which is to improve the quality of the statistical information. Editing is the process of **detecting errors** in statistical data. An error is the difference between a *measured* value for a datum and the corresponding *true* value of this datum. The true value is defined as objectively existing real data value without any kind of errors. Editing can be of two different types. *Logical* editing is where the data values of interest have to obey certain pre-defined rules, and editing is the process of checking to see whether this is the case. A data value that fails a logical edit must be wrong. *Statistical* editing, on the other hand, concerns the identification of data values that might be wrong. Ideally, it should be highly likely that a data value that fails a statistical edit is wrong, but there is always the chance that in fact it is correct.

Social frames can be used in the edit and imputation procedures for enhancing and improving the survey data. Survey data may consist of the statistical units or it may consist of the aggregates of the statistical units to be used for multistage sampling.

When using frame data for editing and imputation, someone can think of the following applications:

- Comparing frame data to survey data on micro and macro level. In the first phase of editing, social frame records enable error localization by identifying unexplained differences in values of the survey variables involved in failed edit-checks;
- imputing missing or rejected values by frame data;
- formulating editing rules by using frame data. Detecting possible or certain errors in a data-file requires the implementation of logical rules within or between data sets, or the existence of a ruler or a reference-file against which the values are compared;
- use of a frame as the spine for linkage to other relevant datasets for micro-level editing, i.e. each of the datasets are linked to the spine and thereby to each other, even if starting with two datasets they may not be directly linkable.

Social frame records support editing and imputation of the statistical survey data via three main mechanisms (Blum, 2003):

- the enrichment of the relevant information needed;
- the expansion of the ability to create a ruler or a relatively accurate reference-file;
- the continuous quality assurance performed throughout the statistical production process. It is more a preventive measure. Edit-checks and values comparisons between the processed data and frame data are carried out in order to identify missing values, errors in the collected data and errors added during data processing.

Challenges

Challenge 1. Combining data sources, e.g. statistical matching can be seen as a missing data problem. Each source has missing observations, variables. Editing and imputation should be engaged.

Challenge 2. If a direct record linkage is not possible due to the lack of a unique identifier in the survey data, probabilistic record linkage based on variables like sex, date of birth, etc. has to be applied.

Challenge 3. The cost effectiveness and the impact on the quality of capturing, editing and preparing administrative data for use as statistical data needs to be weighed against the cost for collecting the statistical data directly through surveys or censuses.

Challenge 4. Edit and imputation procedures can be incorporated directly into the process of combining multiple sources of data. *Inconsistencies between data sources* may occur because of the different quality and accuracy of sources, raising the possibility of obtaining conflicting values for common variables.

Challenge 5. Privacy should be protected in the editing process.

Quality guidelines 4.2.1

Quality guidelines - Frames supporting editing and imputation

Guide 4.2.1.1: Values of the frame
When information from a frame is used for the purpose of editing and/or imputation it is ensured that the values of the variables in the frame used are trustworthy.

Minimum requirement(s):

- There is evidence that the frame data used for editing and/or imputation did undergo well defined quality assurance procedures.

Guide 4.2.1.2: Use of **frame data** for checks on **macro level**
Frame data are used to conduct a first check on macro level by comparing of survey estimators to marginal totals of the frame.

Guide 4.2.1.3: Role of administrative records
In their role as an external ruler, the administrative records support the editing process in locating errors and corroborating values in variables involved in failed edit checks. They improve the ability to identify the erroneous variable. Moreover, they help to avoid false-positive errors, through a comparison between the edited file and the reference-file.

Guide 4.2.1.4: A model to define the value of a variable
Correcting errors in a data-file requires a well-defined model to predict the value of a variable, the availability of records, variables and values to enable the implementation of the model, or an external true value, or its proxy, to be imputed. While detecting errors is better off with the support of additional data sets, the correction will more often rely on them, if they are available and surpass a quality threshold.

Guide 4.2.1.5: Guidelines for data protection and confidentiality
When using microdata out of frame for editing and/or imputation all guidelines relevant for data protection and confidentiality are respected.

4.2.2 Frames supporting weighting and calibration

Weighting is a method of estimating a finite population's parameters from the data of elements belonging to the probability sample. In social surveys, this process can normally be separated into three steps, design weighting, non-response weighting and calibration. The auxiliary information - data from the frames - is used in sample surveys to realize *a sampling design,* namely to define clusters, strata, sampling stages, unequal inclusion probabilities.

The design weights are equal to the inverse of the element inclusion probability. Normally, they are calculated directly based on the frame.

Non response weighting is usually applied in the case of unit non response (while the common correction method for item non response is imputation). By making an assumption on non-response models factors inflating the design weights of the responding units are applied. Frame data play an important role when constructing the non-response inflation factors as frame data can be used as input for models (for instance, logistics regression). Also, other techniques can be used for estimating the response probability of each unit in the gross sample and then using this probability as inflation factor. In general, if there are more variables available in the frame, this will enhance the model building process. There are special forms of non-response that might be treated differently. For instance, if you have a survey conducted on the basis of sampled households surveying every person living in it you might have a phenomenon called "within household nonresponse". Coming back to the frame, it could be possible to assign weighting factors based on the composition of the households. However, the design weight inflated by non-response factors is often called base weight before calibration. In some applications, it can be the final weight for providing the estimators of the survey but in many cases it will be useful to do a final calibration.

A final step of weighting is **calibration**, where weights are benchmarked against boundary values relevant for certain breakdowns of the target population. Usually the figures relevant for this can be provided directly from the frame (or updated version of the frame after the surveying is done) or by other trustworthy external sources.

Challenges

Challenge 1. Non-response present in any survey and other non-sampling errors like under-coverage, over-coverage and other frame errors do not allow using a calibrated estimator properly. Various modifications of the calibrated estimator should be applied.

Challenge 2. Depending on the kind of a frame used, population totals for auxiliary variables may be not known, only sample totals for these variables may be known. The frame may consist of the target population aggregates instead of the units. A calibration estimator should be adapted to that.

Quality guidelines 4.2.2

Quality guidelines - Frames supporting weighting and calibration

Guide 4.2.2.1: Design weights
Design weights are based on frame data.

Minimum requirement(s):

- Every unit in the frame included in a sample is assigned a design weight immediately after the selection process.

Guide 4.2.2.2: Quality of frame information
Frame information used for the calculation of non-response factors in order to inflate design weights of the responding units is trustworthy.

Minimum requirement(s):

- The variables of the frames are checked regarding contents and definitions before used in a non-response analysis.

Guide 4.2.2.3: **Differences between the frame and reality**
If applicable for the survey under consideration the differences between the frame and reality are taken into account before starting the weighting process.

Minimum requirement(s):

- Variables, used for non-response analysis, are adjusted.
- If applicable, the composition of a household is adjusted.

Guide 4.2.2.4: Estimators based on the base weight
Estimators based on the base weight as defined as the design weight inflated by a non-response factor are analysed in order to decide if a calibration is necessary and which auxiliary information is needed.

Minimum requirement(s):

- Marginal totals of persons and/or households for important breakdowns are analysed.
- Estimators for relevant key figures of the concerned statistics are analysed (e.g. number of unemployed in LFS).

Guide 4.2.2.5: The role of frame data in weighting
Survey managers frame owners and, if applicable, methodological experts regularly discuss the role of frame data in weighting.

4.3 Frames as input for statistical outputs

Today statistical production in official statistics tries to minimize response burden as much as possible. Due to the response burden and based on the fact that the costs of conducting surveys are high compared to the use of other sources, the direct use of frames for official statistics seems to be an attractive possibility. On the other side there is a need for guaranteeing quality in a continuous and sustainable way. In most of the cases frames are composed by administrative data. As a consequence many considerations regarding quality have the origin in the use of administrative data. Generally the use of frame data as direct input for a statistical product can be seen as a multisource administrative data use project.
This subchapter provides guidelines how to guarantee the quality of outputs of direct statistic production out of the frames with respect to quality dimensions within the ESS. We consider the use of frame data as direct input for statistical products.

4.3.1 Relevance

Relevance for a statistical output is defined as the **degree to meet current and potential needs of the users**. When using data of the frame as direct input for producing statistics one of the key issues is if the definitions and concepts required for the statistical product are in line with the definitions of the data of the frame. Given this in terms of relevance the crucial question is if the definitions in a frame in social statistics used as a direct input for a statistical product do meet user needs. When coming to social statistics it is not uncommon that especially for variables of interest (for instance income) the definition can differ between certain statistical products. Therefore the use of a variable of the frame as direct input might be difficult.

For assessing relevance the **contact to users** is very important. This holds as well when the input for a statistics is based on frame data. In user oriented quality reports and other documentations available to users it should be transparent which contents of a statistical products are based on frame data and how the frame data have been processed in offer to meet the needs required by the concerned statistical product.

Statistical Institutes are conducting **user consultations** on different levels. The Code of Practice requires regular user satisfaction surveys. As well there are usually consultations on domain and even product level. In the case of the use of frame data it is important that users are in the position to provide an opinion on how far the frame data are suitable to accomplish the statistical task. Frame data are used usually to replace survey data for the sake of saving costs and/or reducing response burden. Usually this is done after a redesign of the statistical product which means in some point there is a transition from coming from survey data to the use of frame data. During the planning of the process the involvement of users seems crucial.

Quality guidelines 4.3.1

Quality guidelines - Relevance

Guide 4.3.1.1: Relevance of **variables**
If the use of data of a frame is intended for direct use, it is checked if the definitions used for the variables to be used are relevant for the statistical product.

Minimum requirement(s):

- The concepts of the units and the variables of the frame are reconcilable to the units and variables of interest for statistical production.

Guide 4.3.1.2: Content of **quality reports and other documentation available to users**
User-oriented quality reports as well as other documentation available to users provide information on concepts used by a frame if frame data are used as direct input for statistical outputs.

Minimum requirement(s):

- There is a section on input data integrated in a user-oriented quality report describing all aspects relevant for a frame if used as input.

Guide 4.3.1.3: Involvement of **key-users.**
Key-users are involved into the discussion process when a transition from survey data to the use of frame data is planned.

4.3.2 Accuracy

Accuracy refers to the closeness of estimates to the unknown true values.

This quality dimension is separated into **sampling errors** and non-**sampling errors**.

When using frame data for statistical outputs the preferred method is that you don't take a sample out of a frame but rather conduct something like a full scope survey. Given this, sampling errors are usually not a topic and the main focus here is in the non-sampling area which is again divided in several sub-dimensions. This concerns coverage errors, measurement errors, processing errors, non-response errors and model assumption errors.

The crucial question is how the fact that we use frame data as input makes the assessment of the non-sampling aspects different to a situation when we estimate based on survey data. In this context it has to be considered that during construction of the frame there have already been certain quality assurance procedures guaranteeing that there have been steps and assessments preventing from having to strong deficiencies in the data material of the frame. Although this chapter deals with the quality of statistical output by using frame data directly, the question cannot be separated completely from quality assessments of the frame. Therefore, the quality indicators provided in chapter 5.1.2 and in annex I of this document play a decisive role. It can be said that an assessment of a frame regarding accuracy is a prerequisite for using it as a direct input data source.

A crucial point when talking about accuracy is completeness and coverage. It might be necessary to complete the figures gained by the use of frame data by estimating certain parts of the population known not to be included in the frame (for instance homeless people, or persons living abroad).

Challenges

Challenge 1. Frame data might have systematic errors. To determine which kind of bias and the magnitude of the bias is challenging.

Challenge 2. It is necessary to get evidence about quality assurance processes for frames used as input for statistics.

Quality guidelines 4.3.2

Quality guidelines - Accuracy

Guide 4.3.2.1: Assessment on the sources.
Before a frame is used as direct input, there is an assessment of the sources regarding frame errors which have an impact on the accuracy of the estimates produced using the frame data.

Minimum requirement(s):

- There is an assessment regarding completeness and coverage and measurement errors of the frame.

Guide 4.3.2.2: Knowledge about incompleteness of the frame.
Knowledge about incompleteness of the frame and under coverage is taken into account when producing statistics directly out of the frame.

Minimum requirement(s):

- There are suitable estimators for unknown parts of the population.

Guide 4.3.2.3: Reporting on accuracy.
All aspects on possible accuracy problems caused by the use of the frame are included in quality reports.

Guide 4.3.2.4: Quality assurance steps.
It is assured that frames intended to be used as source for direct statistics experience exhaustive quality assurance steps.

Minimum requirement(s):

- All guidelines for constructing, coordinating and updating frames as formulated in chapter 3 have been considered.

Guide 4.3.2.5: Indicators for frame errors.
The indicators for frame errors are presented in the WP2 final deliverable of the SGA1 KOMUSO (and annex to section 5.1 here). They are quantitative, and can be recalculated to the relative measures in comparison to the frame size, and thresholds in the terms of these measures should be defined. If the relative number of errors exceeds the threshold level admissible, the frame is not used.

4.3.3 Timeliness and Punctuality

Timeliness refers to the period between the availability of the information and the event or phenomenon it describes. When using frame data, it is important to have information as actual as possible as described in chapter 3.5. One of the key problems in this regard lies in the fact that a frame can be seen as a periodically maintained data base.

Punctuality refers to the delay between the date of the release of the data and the target date (the date by which the data should have been delivered). In the case of frame data, this can only be distorted if the frame data is not available in a punctual manner. We can refer here to chapter 3.2, where we talked about organization and coordination. What is important is that the project management responsible for statistical product integrates the frame owners.

Sometimes it is necessary to deliver results on a provisional basis or even on final basis when, for instance, obliged by legal provisions. In this case, it can be necessary to rely on a provisional frame. In this case, some might give priority to punctuality over accuracy. It can result in a need to revise the results when the final frozen frame is available.

Challenges

Challenge 1. Statistical products are often bound to some delivery dates (for instance based on EU-regulation). Therefore the delayed availability of frame data can have an impact on punctuality.

Challenge 2. To align frame data with the requirements of mapping, determining a reference period relevant to a statistical product can be challenging.

Quality guidelines 4.3.3

Quality guidelines - Timeliness and Punctuality

Guide 4.3.3.1: Check on reference period of the frame.
There is a check if the availability of frame data allows for using it with respect to the reference period and/or due days for a statistical product.

Minimum requirement(s):

- The units provided with variables of the frame correspond to the units in the frame (For social statistics: households, persons, dwellings).
- All variables of interest to be used for a statistical output are checked.

Guide 4.3.3.2: Responsibility for project- and time planning
There is an instance responsible for the maintenance and coordination of the requirements concerning delivery dates or other matters with respect of punctuality.

Minimum requirement(s):

- Frame owners are integrated into project and time planning.

Guide 4.3.3.3: A policy for provisional versions of the frame.
If functional to the statistical production, the NSI can establish a policy for having provisional versions of the frame with a shorter update periodicity respect to the frozen version.

4.3.4 Accessibility and Clarity

Accessibility and Clarity refer to the conditions and modalities by which users can obtain, use and interpret data.

Clarity in statistical products has to do a lot with the availability of metadata which enables the user to understand statistical figures to full extent and to interpret it in a correct way. Two aspects should be covered with respect to the case of the use of frame data.

First there should be information about the fact that frame data serve as an input for certain variables. Points of consideration would be:

- which frame is used for which variable,
- what kind of processing was necessary to transform the frame information into a suitable output on micro level.

Eventually, users should be informed about potential problems when this process takes place.

The second aspect relates to information about the frame itself. There should be comprehensive documentation about all aspects of the frame, for instance in relation to the procedures described in chapter 3. Within the metadata related to a statistical product, there should be a link to the frame documentation.

Summarizing, we can say that providing clarity in the case of frame data usage deals with **quality reporting.**

Most national statistical institutes have instances where users can ask certain questions regarding statistical products. It seems important that, when such questions arise in the context of the use of frame data, the feedback is provided to the project leader and the owners of the frame.

Challenges

Challenge 1. It is not always evident for users that frames are used as a direct input for statistical product. Users have to keep informed.

Challenge 2.: Metadata follow sometimes complex standards. It can be challenging to fit in metadata for frames into metadata standards relevant for an NSI.

Challenge 3. Obtaining an overview of all use cases for frames can only be achieved when there are regular processes investigating the inputs for statistical products.

Quality guidelines 4.3.4

Quality guidelines - Accessibility and Clarity

Guide 4.3.4.1: Metadata concerning the use of a frame

If frame data is used as an input source for statistical products metadata concerning the kind of use accompany the figures of the concerned statistical product.

Minimum requirement(s):

- It is clarified which variables of the product are based on the input of frame data and which frame serves as input source.
- The method of processing frame data to gain a sufficient input for the concerned variables is described.
- There exists a link to the comprehensive documentation of a frame when used as input source.

Guide 4.3.4.2: Questions about the use of frame date

There is a regular process that questions by users concerning the use of frame data are answered appropriately.

Minimum requirement(s):

- Questions regarding used frame data are discussed by project leaders and the owners of the frame.

4.3.5 Comparability

***Comparability* refers to the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas, sectoral domains or over time.**

Regarding comparability over time in the case of frame data use it is important to mention that whenever there is change in the frame this might lead to a potential **break in the time series** for resulting outputs. When this occurs, it might not be necessarily harmful but the point is that users have to be informed. So a direct information link from the frame owners to the project management of the statistical products and later on to users regarding the effects of the change should be in place.

One of the goals of official statistics is to provide **comparable time series as long as possible**. Given this, when something, is changing methods in order to backcast values on micro level might be applied.

Regional comparability of statistical products can be distorted if the data material for regions varies. Concerning data frames it is the preferred solution that they cover the whole territory relevant to a national statistical system. Nevertheless, it is possible that differences can arise in quality when some of the frame information is gathered by input sources provided by regional instance. This might be the case when a national statistical system is decentralised or when information from registers maintained by regional administrations has to be combined.

Regarding **sectoral domains** in social statistics, it can happen that a frame covers only certain parts of the population, for instance, having only employed people. Having this situation and using a frame as input for a statistical product and other parts of the product are based on other sources there might be a problem in comparability (for instance if you compare employed to unemployed). Another aspect that can distort comparability between groups is that different parts of the population can have different quality, different concepts or different sources with respect to a certain variable in the frame. For instance, gathering information for children under a certain age is sometimes mainly based on proxy information.

Challenges

Challenge 1. All frames can and will have errors: under-coverage, over-coverage for the survey population and for domains, erroneous values, duplicates, missing values, errors in matching or linking etc. The frame errors affect also comparison of the derived statistics.

Challenge 2. The frames of the same type may differ in the reference time, implying differences in coverage of the population domains by the frames. The domains used for frame comparison should not differ or changes available should be taken into account.

Challenge 3. When frames are used as input sources the coherence within the system of social statistics hast to be respected. Population figures and other relevant indicators referring to the same time period and geographical units should be consistent.

Quality guidelines 4.3.5

Quality guidelines - Comparability

Guide 4.3.5.1: Changes in concepts, sources and processes concerning the frame.
When there are changes in concepts, sources and processes used in constructing a frame, internal users making use of the frame as direct source for statistical products are informed.

Minimum requirement(s):

- It is investigated if the changes do lead to a break in the time series. If so, users of the concerned statistical product have to be informed.

Guide 4.3.5.2: Backcasting techniques.
With regard to the requirement of providing time series comparable as long as possible the application of backcasting techniques is considered.

Minimum requirement(s):

- There are, at least, backcasting estimations on a macro level for relevant aggregates, in compliance with European demands.

Guide 4.3.5.3: Comparability of **frames over time.**
Comparability of the frames over time is assured by minimising the frequency of changes to standards, classifications and coding of variables. The frame changes over time can be measured by the number of changes over the population and over its domains in time, denoted by PM1-PM4, and presented in WP2 of the SGA1 and here in annex I of this document.

Minimum requirement(s):

- The possible consequences of changing methodology and/or classifications are addressed when planned.
- Users of the outputs are informed about the consequences of changing methodology.

Guide 4.3.5.4: Harmonization of concepts and methods
If input sources from regional administrations contribute to the construction of a frame the NSI as a central coordinator of the statistical product, strives for harmonization of concepts and methods.

Guide 4.3.5.5: Comparability between frame data and data from other sources.
If only parts of a statistical product are based on a frame as an input source, the aspect of comparability to other parts of the population is addressed.

Guide 4.3.5.6: Availability of **quality reports on frames.**
Frames like population registers – which are usually shared by many statistical products – have quality reports of their own. This puts less pressure on every particular survey to examine frame quality.

Minimum requirement(s):

- For every frame used in social statistics, a standardized quality report is available.
- Any update of a frame triggers an update of the corresponding quality report

4.3.6 Coherence

Coherence' refers to the adequacy of the data to be reliably combined in different ways and for various uses.

So, *coherence* of statistical information reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework. The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys. Coherence cannot always be measured. Sometimes, we can only speculate about the differences in the methodology and incoherence, but the size of incoherence may not make any sense. The most evident characteristic when being coherent is numeric consistency. To give an example if you look at the number of unemployed persons delivered by the estimation out of LFS and to the same number delivered by an administrative register and observe differences you call it incoherent, but both could be correct if based on different definitions.

When looking at the system of social statistics containing a variety of statistical products, some of them referring to the same characteristics using frames as input source. The three following scenarios can arise:

1. More statistical products rely on the same frame as input source;
2. Some statistical products referring to certain figures (population totals, number of households etc.) some of them are based on a frame as input some are based on other sources (survey, administrative sources, or another frame);
3. Some statistical products use one frame, some use another frame.

Regarding the first scenario, it is desirable that the population statistics for different statistical products is consistent (for instance, the number of persons living in the country at a certain time should be the same). Therefore, when using frame data as input the selection of the time stamp for which the relevant frame is frozen is relevant question. This leads to a concept called **frozen frame,** already explained in chapter 3.4. The advantage of a frozen frame is that every product using the frame for a certain reference period will access the same data material. It needs to be born in mind that this provides coherent figures between statistical products while there might be a loss with respect of having actual data.

When statistics is using a frame for direct tabulation regarding a specific statistical product and compare the figures to other ones produced by survey data differences might occur simple to the fact that survey estimators are subject to sampling errors. Taking this into account, it should be checked if differences exceed certain confidence limits. Applying methods from statistical testing theory (if prerequisites of that are given) can be considered.

Looking at the third scenario, the following aspects are of interest:

a) Frames for the same population of *the same aggregation level* can be studied how they cover the population domains. Cross-classification by domains can be used to study their coverage, and measures of the relationship for categorical variables based on their cross-classification may show the degree of the coherence between the frames.
b) Frames for the same population but different although hierarchical aggregation level can be checked for numeric consistency: if the number of units for domains in the low aggregation level frame equals to the domain sizes in the higher aggregation level frame.
c) If several frames are supplementary to each other, they cover different parts of the same population and their union increase the population coverage, this affects coherence. The common variables in the frames should be checked for their definitions. If they agree, then the frames can be also considered coherent.
d) Let us assume that a rich frame / enhanced population data set is constructed. It may include categorical and quantitative variables. The internal consistence of such a frame can be checked.

e) Frames from various countries for international comparison should use the same coverage rules and other methodological aspects.

Ideally, we would use the same definitions for variables in all the frames. This is not always possible, but any differences should be explained. If the data used for frame construction is the responsibility of different agencies, it is likely that requirements will be different. There may be neither opportunity nor incentive for harmonisation. Then coherence of the frame variables suffers. The greater coordination role in statistics production in the NSI provides incentives for *harmonisation.*

Even if the definitions are the same, collecting the "same" information about events from different sources will, with probability one, give different results. The totals for the same variables from different sources by domains will give different results. (Holt, 1998).

It should be emphasized the in a national statistical system for each statistical product, the fact that it is a part of the system is a potential – a potential to pay attention to. Comparisons of the frame data on both macro and micro levels should be performed on a regular basis. If, for example, there is a sudden upwards change in the number of registered unemployed, there should be a change in the number of employed as well. Changes can be "real" or effects of re-organisations. It is important to observe and check the statistical unit used (Elvers, Nordberg, 2001). These connected changes show *coherence*.

Challenges

Challenge 1. Coherence in social statistics relates to a whole sub system of official statistics. There is a need to identify all products where a frame is used a direct input source

Challenge 2. Sometimes differences can occur due to different concepts used in the frame that are needed for an application by a statistical product. Other reasons can be population coverage and random effects. It is demanding to separate the phenomenon of incoherence correctly in detail.

Challenge 3. Changes in the methodology, classifications or changes in laws may influence a level of the data from two processes. For example, number of unemployed with respect to labour exchange and with respect to the Labour Force Survey over domains or in time.

Challenge 4. Variables may be different, but correlated, like number of employed and number of unemployed by county in the Labour Force Survey. Changes in the coherence may be not observable directly. Coherence measurement may show the change.

Challenge 5. Attention should be paid if the statistics for international comparison is coherent.

Quality guidelines 4.3.6

Quality guidelines - Coherence

Guide 4.3.6.2: Documentation about the frame and its variables
There exists a well prepared documentation for the frame and its variables in order to take decisions concerning coherence.

Guide 4.3.6.2: Check on consistency of statistical products.
When a frame is used as an input source for a statistical product, there is conducted a check of the differences to other statistical products delivering the same key statics.

Minimum requirement(s):

- Differences are listed and possible reasons are investigated

Guide 4.3.6.3: Investigation of differences to survey estimators.
When compared to a survey and the difference to a survey estimator exceeds the confidence limit of an estimator further investigations follow.

Guide 4.3.6.4: Frames related to a specific reference period.
When using frame data all statistical products making use of a frame as input use one frozen frame related to the same reference period.

Minimum requirement(s):

- For each frame a final frozen frame covering a reference year (yearly frame) is produced;
- For each frame a final frozen frame covering every reference month (monthly frame) is produced;
- For each frame a final frozen frame for every quarter (quarterly frame) is produced.

Guide 4.3.6.5: Coherence of **the frame system.**
If the frame is part of a system of frames, coherence among the relevant units and variables included in other frames is pursued.

5 Assessing and evaluation the quality of frames in social statistics

The following chapter addresses the quality of frames and its main focus is the quality of the data provided by the frame. It will not tackle the aspect of the quality of statistical products using the frame as an input what was already been done in prior chapters of this document.

All guidelines and remarks about quality assessment in the first subchapter are very much based on the findings of work package 2 related to ESSnet KOMUSO SGA 1. If somebody would like a deeper insight into the errors, quality measures and assessment methods they can have a look into the final deliverable of this work package. In this regard SGA1-WP2 can be considered as the methodological supporting documents for those who want to further look into the matter.

Within each guide, the guidelines are ordered from the less demanding to the most complex and costly ones. According to the available resources and priority settings the readers can set the minimum requirements.

5.1 Methods to assess the quality of a frame

5.1.1 Quality assessment

This subchapter focuses on how to assess the **quality of data included into a frame** by separating the quality into several components, what we call error types.

The following types of errors are considered:

- **Coverage errors** due to missing, erroneous and duplicated frame units
- **Domain classification errors** of frame units
- **Contact information errors** of frame units
- **Alignment errors** between different types of frame units
- **Unit errors** of composite frame units

These errors are part of non-sampling errors and thereby resorting to the quality dimension of accuracy.

The coverage of the frame is one of the most important quality aspects. First of all, **under-coverage,** where certain parts of the target population are not integrated in maybe a systematic way, can lead to problems for the use of the frame. For instance if you want the frame for a survey, and you cannot cover the population to a full extent by the frame, you have to look for other solutions or procedures to estimate it out of other sources. With respect to social statistics, frequent examples of that are persons living abroad or homeless people not included in administrative registration procedures and therefore missed in the frame. In the case of estimators, that are linear functions of the observed data, the impact of under-coverage on the quality in terms of increase of variability can be considered low, whereas the concern is on the potential bias (Biemer, P.P.; Lyberg L.E., 2003). Generally you can distinguish between two types of under-coverage. **Design under-coverage** comes from the population groups or geographical areas that are excluded a priori for practical reasons of feasibility. They could be e.g., population elements in areas that are remote or difficult to access. The design under-coverage is normally measured as a fraction of the frame population elements not covered by the frame. On the other hand error **under-coverage** originates from those frame population elements that are not excluded by design and should be included in the frame but which are actually missing. The error under-coverage is normally measured as a fraction of the frame population elements that are

expected to be in the frame but not present Frame over-coverage is the case if it entails duplicated, non-existent or out-of-scope elements. It is common to distinguish between two main types of over-coverage: duplicate listing and erroneous enumeration. Duplicates are the population elements which are referred to by at least two elements of the frame. The duplicates cause error for a frame used for tabulation of statistics on the basis of auxiliary variables. For a sampling frame, they increase the cost of the data collection and processing. The duplicates are normally measured by the fraction of records in the frame that are redundant The units found not eligible for the survey will be discharged with a consequent (probably small) reduction of the sample size and increase of the variability of the final estimates.

Domain classification errors represent under-coverage in a domain and over-coverage in another. Domain misclassification corresponds to incorrect auxiliary information. To have a decent frame means as well to be able to implement a broad variety of sampling designs. This means that the quality of domain and classification variables has to be checked regularly. In social statistics, the so-called 'social core variables' including geographic information are of significant importance. By domain we refer to population domain as a segment of the population, for which separate statistics are needed. It could consist of a geographical area such as a region or major population centre. It could also comprise a specified population category, such as a major national or ethnic group.

As pointed out in previous chapters, the **contact variables** in a frame play an important role. Assessments and monitoring of indicators like the missing of contact variables or the number of wrong addresses or the mismatch to telephone numbers are important measures in order to have a continuous improvement of the data material. The technological progress has an impact on those matters in so far as not only addresses have to be considered but also phone numbers, e-mail addresses and for dwellings GPS- (Global Positioning System)-coordinates.

The **statistical units** in frames of social statistics are, as already mentioned, persons, households and dwellings. They should be compared to other sources. This relates to the units in terms of existence and numbers but also to the relation between basic and composite units. Also, relationships between units are important. **Alignment** and **units errors** are closely related. For example, alignment errors between persons and addresses may cause unit error in the identification of the statistical unit "household".

A set of measures for each error type has been defined in SGA1-WP2 final deliverable and can be seen in Annex I of this document. They are distinguished into **quality indicators** which provide evidence of potential problems and **quality measures** allowing for an end-point of quality assessment.

It is worth mentioning general methods of obtaining information about quality of frames. One possibility is to evaluate a frame regularly by a certain process like a quality audit or designed ex post quality studies or quality surveys assigned to a certain topic, as for example post enumeration surveys to estimate under-coverage. On the other hand, one of the most important set of sources for information is formed by surveys. There you can **gather feedback on a broad variety of aspects** (number of non-contacts, number of wrongly classified units and so on). Therefore an institutionalized process, which incorporates feedback from the field operations, is obtained by the frame owners is essential.

In the following the quality guidelines for assessment on quality are presented. Different to other chapters there are no minimum requirements because the issues enumerated below the individual guidelines are regarded as an exhaustive list of aspects which should be considered when assessing the quality of a frame. With respect to measurement of errors and quantitative indicators we can refer to the guidelines 5.1.2 of the next chapter.

Quality guidelines 5.1.1

Quality guidelines – Quality assessment

Guide 5.1.1.1: **Assessment on under-coverage**.
There is a regular assessment on under-coverage within the frame.

- Knowledge about design under coverage, which is related to sub-populations not included in the frame, exists.
- Delays and corrections causing under-coverage are quantified and analysed.
- The distribution of the time between the date of registration in the frame and the occurrence date of the event is analysed.
- Under-coverage is roughly assessed by aggregated comparisons with external sources, also known as net or gross discrepancy checks.
- Additional sources are used to identify Sign-of-Life data.
- Under-coverage is estimated at least every five years by either comparing with other independent sources or organizing a targeted survey
- Impact of under-coverage in terms of variance and bias on the final estimates is estimated.

Guide 5.1.1.2: Quality of **contact information**.
The quality of the information allowing for the contact of the units is assessed.

- Frequency of missing values in contact information is computed.
- Incoherencies among the variables allowing for the contact of the units are discovered.
- Feedbacks from the on-going surveys to quantify and monitor nonresponse rates due incorrect contact information are used.
- Additional sources to identify units at risk of incorrect contact information are used.
- Analyses of potential bias of the final estimates attributable to the reduction of the sample/population due to missing contact information are carried out.

Guide 5.1.1.3: Quality of **classification variables.**
The quality of the domain classification variables is assessed**.**

- Frequency of missing values in domain classification variables is computed.
- Incoherencies among the domain classification variables are discovered.
- Feedback from the on-going surveys to quantify and monitor out-of-scope rates is used.
- Domain classification error is estimated by means of coverage surveys and modelling approaches.

Guide 5.1.1.4: Errors in alignment between base and composite units.
Errors in the alignment between base and composite units are assessed.

- Frequency of base units not uniquely (or unambiguously) assignable or unassignable to composite units is computed.
- Aggregated comparisons with external sources are considered.
- Quality surveys based on an audit sample from the frame are carried out.

Guide 5.1.1.5: Errors in the derivation of statistical units.
Errors in the derivation of statistical units are assessed.

- Frequency of units not uniquely (or unambiguously) assignable to the eligibility status and units with suspicious composition or rare types is computed.
- Aggregated comparisons with external sources are carried out.
- Quality surveys based on an audit sample from the frame are carried out.

- ❖ Sensitivity analyses when the unit derivation is assisted by experts, based on statistical models or assisted by the integration among sources (e.g. subjective auditing approach) are considered.

Guide 5.1.1.6: Assessment of over-coverage.
Over-coverage is assessed:

- ❖ Frequency of duplications and units not belonging to the target population(s) is computed.
- ❖ In addition to duplicates error over-coverage is assessed.
- ❖ Feedbacks from the on-going surveys to quantify and monitor over-coverage rates are used.
- ❖ Quantification and analyses of delays and errors causing over-coverage are quantified and analysed.

Guide 5.1.1.7: Signs of life.
Additional sources are used to assess "signs of life". This is done at dwelling unit level (electricity or water consumption, phone calls, etc.) and/or for individual persons (tax data, medical records, etc.).

5.1.2 Quality indicators

In the previous chapter, a framework consisting of several important error types was outlined. When it comes to the aspect of measurement of the error types in order to evaluate the quality of frame there are several questions which should be raised.

1. Is it possible to find suitable quantitative indicators for the various error types?
2. Is it possible to combine the measures of the different error types in order to arrive at a single quality indicator related to a frame?
3. Have the use case and/or the statistical product which uses the frame an influence on the selection of quality indicators?

The first question is basically answered by annex I and more detailed by the final report of WP 2 SAG 1 of KOMUSO https://ec.europa.eu/eurostat/cros/system/files/wp2-framequality-finalreport.pdf). The concepts there allow for the derivation of measures for specific error types. However these are theoretic concepts and it might be the case that to bring in the information necessary will turn out to be burdensome.

In the following considerations we use the term population domain 8see as well 5.1.1.). A population domain is a segment of the population for which separate statistics are needed. It could consist of a geographical area such as a region or major population centre. It could also comprise a specified population category, such as a major national or ethnic group.

The following table shows the scenario at a frame;

Population Domain	Frame Domain				Missing (error under-coverage)	Excluded (design under-coverage
	1	2	…	H		
1	N_{11}	N_{12}	…	N_{1H}	M_1	E_1
…			…		…	…
H	N_{H1}	N_{H2}	…	N_{HH}	M_H	E_H
Erroneous (over-coverage)	R_1	R_2	…	R_H		
Duplicates	D_1	D_2	…	D_H		

In this matrix H represents any domain classification related to the population represented by the frame. For instance H can be “size of household”. Looking at the matrix some can argue that an ideal frame is given when only the elements of the main diagonal in the central part are greater than zero. All other elements contribute to the **domain classification error**.

As already outlined on various places within this document the relation between base units (BU) and composite unfits (CU) is very important. The most typical example in social statistics is that you have persons as BU and household as CU. If one or several BU’s are incorrectly associated with a CU we call it an **alignment error**.

Taking table 4 of annex 1 we have:

Base Unit Classification (Total N)	Composite Unit Classification (Total M)				Base Unit Total
	1	2	…	H	
1	(N_{11}, M_{11})	(N_{12}, M_{12})	…	(N_{1H}, M_{1H})	N_1
2	(N_{21}, M_{21})	(N_{22}, M_{22})	…	(N_{2H}, M_{2H})	N_2
…			…		…

G	(N_{G1}, M_{G1})	(N_{G2}, M_{G2})	…	(N_{GH}, M_{GH})	N_G
Composite Unit Total	M_1	M_2	…	M_H	

The N_{ij} are the totals of the BU and the M_{ij} are the totals of the CU in a cell of a cross table between domain classifications of BU and CU. In each cell you can have correct and incorrect aligned base units (erroneous). So $N_{ji}=N_{ji,c} + N_{ji,e..}$ Alignment error is closely related to **unit error**. A unit error for a composite unit is given when the CU is formed erroneously despite the information for the involved base units being correct.

When an individual value of a frame variable of a population element is erroneous, we call it a **value error**. We can distinguish here between **contact error** and errors in other values (auxiliary variables or variables used for direct tabulation). In the first case we call it contact error in the latter we talk about **value error in its closer sense**.

The following table proposes quality indicators for the various error types taking into account the notations developed above. It should be noted that the indicators result from summation over a certain domain classification H. however this is not absolutely necessary. In case this is not available or for the sake of simplification if no classification is used the summation is taken only over one element (H=1 and/or G=1). The proposal lists only one indicator per error type in order to achieve a relative measure for each error. If some is interested in more indicators you can look at annex I.

Error types with proposed indicators and interpretations

Error type	Proposed Indicator	Interpretation
Coverage error	Error under-coverage: $U_E = \sum_{i=1}^{H} X_i/N$	Share of missing basic units in a frame
	Design under-coverage: $U_D = \sum_{i=1}^{H} E_i/N$	Share of basic units excluded from a frame
	Error over-coverage: $O_E = \sum_{i=1}^{H} R_i$	Share of basic units which are included in the frame erroneously
	Duplicates: $O_D = \sum_{i=1}^{H} D_i/N$	Share of duplicate basic units in the frame
Domain misclassification	$D = \sum_{i \neq j} N_{ij}/N$, i,j from 1 to H	Share of misclassified basic units when using a certain domain classification H
Alignment error	A=$\sum_{j=1}^{G} \sum_{i=1}^{H} N_{i,j;e}/N$	Share of wrongly aligned base units to a predefined composite unit
Unit error	$U = \sum_{i \neq j} M_{ij}/M$, i,j from 1 to H	Share of composite units which have been constituted erroneously
Contact error	$C = \sum_{i=1}^{H} C_{ei}/N$ C_{ei} is the number of basic units in domain *i* holding invalid or missing contact information.	Share of base units where the contact information is incorrect or missing

Value Error	$V = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{N}\sum_{j=1}^{H} V_{i,j;e}$ $V_{i,l;e}$ is the number of basic units in domain *j* with invalid or missing value for variable *i*.	Average share of incorrect or missing values when k variables are considered.

Combined frame quality indicator

Taking the indicators defined before we can define a global indicator by taking a weighted sum over the various quality error indicators defined in the table before. So we define the global frame quality indicator Q_F

$Q_F=W_{UE}*U_E+W_{DU}*U_D+W_{OE}*O_E+W_{OD}*O_D+W_D*D+W_A*A+W_U*U+W_C*C+W_V*V$ where

$W_{UE}+W_{UD}+W_{OE}+W_{OD}+W_D+W_A+W_U+W_C+W_V=1$

Since the single error components deliver all values between 0 and 1 this property holds for the combined indicator as well and 0 means an ideal frame and 1 reflects the worst possible situation.

There are some remarks which should be made by looking at this indicator:

- The selection of weights for the individual components will influence the indicator to a high degree. It seems not to be suitable for every use case to use equal weights
- The use case itself will have the most decisive impact on the selection of the weights and even on the design of the indicators of the single error components. For instance when not used for sample but for other purposes (direct tabulation) the value for W_C should be zero.
- The information for calculating indicators for the single error components might be difficult to obtain. about the document does not discuss how measure or estimate but refers to the final report of KOMUSO WP2 of SGA 1:
  https://ec.europa.eu/eurostat/cros/system/files/wp2-framequality-finalreport.pdf
- One can observe quality over time to develop a weighting scheme for an indicator, as suggested above related to important applications in social statistics, for instance for the LFS you can design a Q_{F_LFS} and follow it over time (see e.g. simple combined indicator below).
- The weights can also be set to zero for practical reasons, e.g. unavailability of some information or comparability over several frames.
- The indicator as it is constructed does not allow drawing any conclusion regarding the share of error-free units.

Simple combined indicator:

As a first simplification step, it can be proposed to combine the four components of coverage errors (coverage as sum of error under-coverage, design under-coverage, error over-coverage and duplicates) into a single, simplified combined indicator. It is in line with the methodology defined above by adjusting the weights appropriately as explained and equally weights the four error components by ¼ (weights of other errors may be set to 0). So we have

$C_S=\frac{1}{4}*(U_E+U_D+O_E+O_D)$

Most of the time there will be some knowledge about U_D whereas U_E can only be assessed with independent information (dedicated surveys, other sources than used for the frame, etc.) so that it is typically more difficult to

obtain. Therefore, as a second simplification step, it can be proposed to drop error under-coverage U_E from the simplified combined indicator arriving at:

$O_S=\frac{1}{3}*(U_D+O_E+O_D)$

This combined quality indicator may be reduced to a specific application in social statistics. Therefore, as a temporary first approach, it can be proposed to calculate this simplified combined indicator O_s in the context of national LFS, as it is normally the largest social survey with quality being a particular concern.

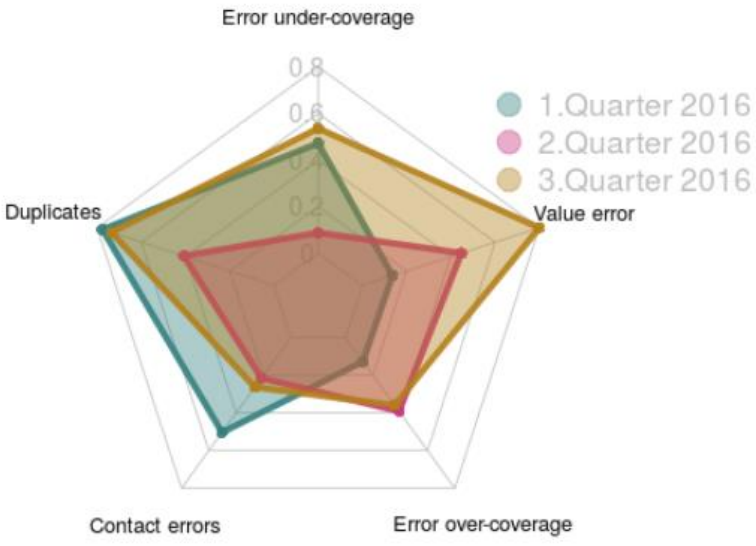
Concluding we can say that there could be a stepwise approach when developing quality indicator for a frame. Starting by calculating O_S (for a specified application or generally) some of the mentioned simplification steps could be gradually lifted as enough information is available to calculate less simple quality indicators.

Coming back to the remark regarding the impossibility of drawing conclusions on the amount of error-free units, we could define another possible simple indicator by

$E_S = 1 - \frac{E_F}{N}$, where E_F stands for the number of error-free basic units in the frame.

Having this definition we should define what is meant by an error-free unit. Looking at the different error types defined before we see that they are fractions where the nominator is a count of all units contributing to the error type. So a unit contributes either by 1 or 0 to the error (except for the value error *V* where each unit contributes *l*/*k*, with *l* denoting the number of erroneous variable values for this unit and *k* the total number of variables). If a unit contributes to all error types by 0 we call it error free. By this definition, the error-free units measure excludes errors caused by error under-coverage U_E. Of course you can again simplify matters when only looking at some of the error types. For instance if you look at O_S you consider all units not contributing to U_D, O_E and O_D as error-free.

A possibility of showing the total quality of frame taking into account all error types graphically would be by a so called radar plot. The graphics below shows a radar plot for five error types observing the quality for three consecutive time periods (values are hypothetical)



Quality guidelines 5.1.2

Quality guidelines - Quality indicators

Guide 5.1.2.1: Calculation of q**uality indicators**.
Quality indicators for all relevant error types should be calculated and evaluated regularly.
Minimum requirement(s) :

- All error types described in Chapter 5.1.1. should be considered

Guide 5.1.2.2: Inclusion of quality indicators into quality reports.
Quality indicators for frames should be included in quality reports of frames and all statistical products using frames.

Guide 5.1.2.3: Methods for estimating input for quality indicators**.**
The methods for obtaining information necessary for calculating quality indicators should be assessed regularly.

Guide 5.1.2.4: Stability of quality indicators over time.
The concepts, the design and all methods used in order to calculate quality indicators for frames should be stable overt time. Time series allowing for the observation of the quality of frames should be available.

Guide 5.1.2.5: Using a combined quality indicator.
When using a combined quality indicator, the selection of the weights as well as the design of the estimation of the single error components should be adequate for the use case.
Minimum requirement(s):

- There is at least one combined quality indicator for each frame.

5.2 Quality and metadata management, quality improvement and quality reporting

As mentioned in section 3.2, the owner of the frame is in charge of the coordination and management of the frame. Frames in social statistics do not normally constitute a self-standing statistical product. They are used for production of statistics. Internal users take information for designing and drawing samples out of the frame or use the information for certain statistical products or for direct estimation out of the frame. Given that, it seems important to have some overview about the internal use cases. There should be a regular monitoring of internal users and uses of the frame allowing assessing its **relevance** against its implementation and maintenance costs. Assessing the "satisfaction" of the internal users with respect to the quality of the frame can provide hints to improve the quality itself. As already mentioned, the surveys using the frame obtain a great amount of information that can be used by the owner of the frame to improve the quality of the data contained, e.g. by imputing/correcting microdata where necessary in case of missing/incorrect information, provided the data from the survey is considered trustable.

It is deemed important that the owner of the frame takes into account the documentation of the content of the frame, also considering the changes in the administrative legislation that may have an impact on the frame built using administrative data. The metadata is vital for the users of the frame and include unit and variable definitions, classifications used and quality indicators produced in the assessment phase.

Effective and efficient process management contributes to a better quality frame. The process should be mapped in each step, from the acquisition of the administrative sources, to their loading in the IT systems of the NSI, to the checks on the format of the data, to the preliminary quality control on the completeness and correctness of the administrative data, to the data integration and identification of the statistical units of interest, to the coding of some variables with standard classifications, to the delivery of final version of the frame, finally to the computation of quality indicators. Each step of the process should have assigned times and resources. The process should be governed and monitored. Metadata should be properly documented so as to make the whole process reproducible.

Nevertheless, the quality reporting system of the NSI providing metadata for external users shall take into account all relevant information available for frames which are involved in the production process of statistical products. For instance, if there exists a user oriented quality report for a survey based statistical product in an area of social statistics, the description of the sampling process shall also take into account the frames involved and give a picture of how the frame is used and of potential error sources caused by the frame. Every statistical product using the frame should include in the standard documentation information on it and its quality. Coherently with the Eurostat standards for documentation (ESS Handbook for Quality Reporting and SIMS) the documentation for external users includes information about reference period of the frame, updating procedures, reference to other quality documents, sources used to build the frame and size of the population in the frame.

Metadata can be seen as the basis for quality reporting. Therefore the standardized collection of metadata is an important process to observe the quality for frames for sufficient time series in a comparable way. The questionnaire presented in Annex IV approved by an ESS Task Force on frames[8] provides a format which should be used to collect information of frames on a regular basis.

[8] https://circabc.europa.eu/sd/a/fade8124-801e-4bcf-a70f-181fce69dfbb/FRAMESTF_Final_Mandate.pdf

Quality guidelines 5.2

Quality guidelines - Quality and metadata management, quality improvement and quality reporting

Guide 5.2.1: Relevance of the frame.
There is a regular assessment of the internal relevance of the frame.
Minimum requirement(s) :

- It is monitored who the internal users of the frame are and for which purposes the frame is used.
- The satisfaction of the internal users with respect to the quality and suitability of the frame is monitored.
- Feedback from the users on how to generally improve the frame is gathered.

Guide 5.2.2: Quality of **the frame at microdata level.**
The improvement of the quality of the frame at microdata level is a key objective.
Minimum requirement(s) :

- The feedback from ongoing surveys is used to correct microdata, by imputing missing information or editing incorrect information.
- Master samples to improve the quality of a strategic part of the frame are used.

Guide 5.2.3: Coherence and comparability of the data of the frame.
Ensuring coherence and comparability of the data of the frame is a key objective.
Minimum requirement(s) :

- Definitions and concepts (units, variables) are comparable over time.
- The frame is internally consistent (arithmetic identities observed).
- Coherence between outputs derived from the frame is addressed.

Guide 5.2.4: Timeliness and punctuality of **deliveries.**
Timeliness and punctuality of internal deliveries is monitored.
Minimum requirement(s) :

- Punctuality, i.e. time between the actual and the planned date when the frame is delivered for the internal uses (if applicable, i.e. if a frozen dataset is created) is computed and monitored.
- Timeliness, i.e. the time between the reference period of the frame and the date the frame is released for the internal uses (if applicable, i.e. if a frozen dataset is created) is computed and monitored.

Guide 5.2.5: Standardized **collection of metadata**.
For every frame in social statistics metadata should be collected by using the standardized questionnaire for metadata on frames data which are used by the social and population statistics.

Guide 5.2.6: Availability of **documentation for external users**.
For every frame documentation for external users is available.
Minimum requirement(s) :

- The following contents are included in the external documentation:
  - Reference period of the frame.

- Procedures used for updating the frame and their frequency.
- References to other quality documents.
- Sources used to build the frame.
- Size of the population in the frame.

Guide 5.2.7: Availability of **documentation for internal users**.
For every frame there is available documentation for internal users.
Minimum requirement(s) :

- The following contents are included in the internal documentation:
  - All the aspects included in the external documentation.
  - Unit and variables definitions.
  - Classifications adopted.
  - Quality indicators computed

Annex I: Frame quality assessment: items, approaches and methods

17 items that measure frame accuracy are specified in the final report of the WP2 SGA1. Each item may consist of one or several target parameters that need to be estimated. Five approaches are outlined in the quality assessment. The *most readily applicable* methods of assessment are presented in the WP2 SGA1 final report.

The persons are considered as basic units, and their groups, like households and dwelling-households are considered as composite units.

A description of the 17 items is provided in Table 1. Here, CM is a shorthand for "coverage including domain classification measure", PM for "progressiveness measure", AM for "alignment measure", "UM" for "unit error measure", and IM for "(contact) information measure".

Table 1. List of frame accuracy measurement items

Coverage	CM1. Total under- and over-coverage for the target population CM2. Total correct domain classification CM3. Domain-specific population under- and over-coverage CM4. Domain misclassification (i.e. cross-domain under- and over-coverage)
	PM1 – PM4 Counterparts of CM1 – CM4, specifically due to delays in source data
Accuracy, completeness, Quality	AM1. Total of correctly aligned base units (i.e. persons typically) AM2. Domain totals of correctly aligned base units AM3. Distribution of correctly aligned base units by composite unit types AM4. Total of correctly aligned composite units (e.g. household, address, etc.) AM5. Domain totals of correctly aligned composite units
	UM1. Total number of population composite units UM2. Domain total numbers of population composite units
Contact information	IM1. Total of frame units of given type with (correct, invalid, missing) contact IM2. Domain totals of frame units with (correct, invalid, missing) contact

CM1 – CM4 for frame coverage and domain classification measures are specified based on the notations presented in a following table:

Table 2. Notations for CM1 – CM4

	Frame Domain				
Population Domain	1	2	…	H	Missing
1	N_{11}	N_{12}	…	N_{1H}	M_1
…			…		…
H	N_{H1}	N_{H2}	…	N_{HH}	M_H
Erroneous	R_1	R_2	…	R_H	

The in-scope frame total is $\sum_{i=1}^{H}\sum_{j=1}^{H} N_{ij}$ and the population total is $\sum_{i=1}^{H} M_i + \sum_{i=1}^{H}\sum_{j=1}^{H} N_{ij}$.

CM1u. Total under-coverage: $M = \sum_{i=1}^{H} M_i$

CM1o. Total over-coverage: $R = \sum_{i=1}^{H} R_i$

CM2. Total correct domain classification: $N_0 = \sum_{i=1}^{H} N_{ii}$

CM3u. Domain-specific population under-coverage: $\{M_i; i = 1,...,H\}$

CM3o. Domain-specific population over-coverage: $\{R_j; j = 1,...,H\}$

CM4. Domain misclassification: $\{N_{ij}; i \neq j, i = 1,...,H, j = 1,...,H\}$

The coverage measures CM3 are for the whole population, hence they may be referred to as *genuine* coverage error measures, whereas the coverage measures CM4 are applicable between the domains and the misclassified frame units are still inside the population, hence they may be referred to as *spurious* coverage errors. In principle, domain-specific coverage can be assessed using the same data and method as for population coverage, treating each domain as a separate population. However, there may not be enough data to produce reliable estimates in this way, as for population coverage.

PM1 – PM4 are counterparts to frame coverage error measures CM1 – CM4, but they are used when the coverage problems arise due to the frame progressivness. The measures are based on the notations in the following table:

Table 3. Notations for PM1 – PM4

Target population at time *t* Frame at t_1	Frame at t_2				Not in at *t* (according to frame at t_2)
	1	2	…	H	
1	N_{11}	N_{12}	…	N_{1H}	N_{10}
…			…		…
H	N_{H1}	N_{H2}	…	N_{HH}	N_{H0}
Not in at *t* (according to frame at t_1)	N_{01}	N_{02}	…	N_{0H}	

The frame for target population at time *t* constructed at time t_1 is being assessed at time t_2, where $t \leq t_1 < t_2$. We assume progressive frame data for *t* converges by t_2, after which there will be no changes about the state-of-affairs at *t*. In practice, $t_2 - t$ can sometimes be many years.

PM1. Total under- and over-coverage: $M = \sum_{h=1}^{H} N_{0h}$ and $R = \sum_{h=1}^{H} N_{h0}$

PM2. Total correct domain classification: $N_0 = \sum_{i=1}^{H} N_{ii}$

PM3. Domain-specific population under- and over-coverage: $\{\mathrm{N}_{0h}; \mathrm{h} = 1,...,H\}$ and $\{\mathrm{N}_{h0}; \mathrm{h} = 1,...,H\}$

PM4. Domain misclassification: $\{N_{ij}; i \neq j, i = 1,...,H, j = 1,...,H\}$

We observe that in order to assess PM1 - PM4, it is necessary to be able to distinguish in the frame *at least two dates*: one for the relevant demographic event (e.g. birth, death or change of status), one for the registration of that event. Assessment is then possible without additional data. When the convergence time point t_2 is known and feasible, in the sense that the observations of N_{ij} are available *in retrospect*, the measures PM1 - PM4 can be calculated almost surely. However, when t_2 is either unknown or infeasible, e.g. when some of the data sources are completely new, or when the frame at t_1 is assessed at some $t' < t_2$, more sophisticated methods are needed.

AM1 – AM5 are used for alignment between frame base units and any type of frame composite units with the notations specified based on the following table:

Table 4. Notations for AM1 – AM5

Base Unit Classification (Total N)	Composite Unit Classification (Total M)				
	1	2	…	H	Base Unit Total
1	(N_{11}, M_{11})	(N_{12}, M_{12})	…	(N_{1H}, M_{1H})	N_1
2	(N_{21}, M_{21})	(N_{22}, M_{22})	…	(N_{2H}, M_{2H})	N_2

…			…		…
G	(N_{G1}, M_{G1})	(N_{G2}, M_{G2})	…	(N_{GH}, M_{GH})	N_G
Composite Unit Total	M_1	M_2	…	M_H	

When person is BU, typical examples of CU are address, household, family, building, etc. But the table above is generic and (BU, CU) can be defined for the situation at hand. For example, one may set buildings as the BU and spatial grids as the CU. In any case, according to the frame, N_{gh} type-*g* BUs are aligned with (or belong to) M_{gh} type-*h* CUs. Let $N_{gh;t}$ and $M_{gh;t}$ be the number of *correctly aligned* BUs and CUs, respectively, and $N_{gh;e}$ and $M_{gh;e}$ that of the *incorrectly aligned* units. We have

$$N_{gh} = N_{gh;t} + N_{gh;e} \text{ and } M_h = M_{h;t} + M_{h:e}$$

AM1. Total of correctly aligned base units: $N_t = \sum_{g=1}^{G}\sum_{h=1}^{H} N_{gh;t}$

AM2. Domain totals of correctly aligned base units: $\{N_{g;t} = \sum_{h=1}^{H} N_{gh;t}; g = 1,...,G\}$

AM3. Distribution of correctly aligned base units: $\{N_{gh;t}; g = 1,...,G, h = 1,...,H\}$

AM4. Total of correctly aligned composite units: $M_t = \sum_{h=1}^{H} M_{h;t}$

AM5. Domain totals of correctly aligned composite units: $\{M_{h;t}; h = 1,...,H\}$

Next, the measures UM1 – UM2 are specified for any type of composite units subjected to unit errors.

Table 5. Notations for UM1, UM2

	Frame CU Classification				
Population CU Classification	1	2	…	H	Missing
1	$M_{1;t}$	--	…	--	Z_1
2	--	$M_{2;t}$	…	…	Z_2
…		…	…		…
H	--	--	…	$M_{H;t}$	Z_H
Erroneous	$M_{1;e}$	$M_{2;e}$	…	$M_{H;e}$	

Here, h = 1, 2, …, H denotes some classification of the CU, such as dwelling household by the number of residents. An erroneous frame CU cannot become another true CU in the population. For example, if a dwelling household of two residents constitutes a unit error, then either these two person live in different addresses, or there are other persons at the address. In either case, this frame dwelling household is not a dwelling household in the population. The absence of spurious over- and under-coverage is a key difference between CUs and BUs.

UM1. Total number of population composite units: $N = \sum_{h=1}^{H} M_{h;t} + \sum_{h=1}^{H} Z_h$

UM2. Domain total no. population composite units: $\{N_h = M_{h;t} + Z_h; h = 1,...,H\}$

Quite often it is possible to represent the relationship between frame units and associated contact information data as alignment between suitably defined BU and CU. Nevertheless we have retained contact information error separately, and phrased IM1 and IM2 accordingly.

Table 6. Notations for IM1, IM2

	Contact Information			
Frame Unit Domain Classification	Correct	Invalid	Missing	Total
1	$N_{1;t}$	$N_{1;f}$	$N_{1;m}$	N_1
2	$N_{2;t}$	$N_{2;f}$	$N_{2;m}$	N_2
…	...	…	…	…
H	$N_{H;t}$	$N_{H;f}$	$N_{H;m}$	N_H

Total	N_{+t}	N_{+f}	N_{+m}	N

IM1: Total number of frame units with correct, invalid and missing contact information are, respectively, $N_{+t} = \sum_{h=1}^{H} N_{h;t}$, $N_{+f} = \sum_{h=1}^{H} N_{h;f}$ and $N_{+f} = \sum_{h=1}^{H} N_{h;m}$.

IM2: Domain totals of (correct, invalid, missing) contact information are (N_{ht}, N_{hf}, N_{hm}).

Annex II: References

Biemer, P.P. and Lyberg L.E. (2003). Introduction to survey Quality. John Wiley & Sons (2003).

Blum Olivia (2006), Evaluation of Editing and Imputation Supported by Administrative Files, in *Statistical data editing. Impact on quality*, V. 3, UN, Geneva, 302-311.

Chambers Ray (2006). Evaluation Criteria for Editing and Imputation in Euredit, in *Statistical data editing. Impact on quality*, V. 3, UN, Geneva, 17-27.

Delden van Arnout, Jeroen Pannekoek & Li-Chun Zhang. *Measuring coherence*. (Presentation on the Vienna meeting 2018-01-11).

Deville J.-C., C.-E. Särndal. Calibrated estimators in survey sampling. *Journal of the American Statistical Association,* 1992, 87, 376-382.

Elvers, Eva and Lennart Nordberg (2001). *A Systematic Approach to Quality Measurement and Presentation*. Q2001, European Conference on Quality and Methodology in Official Statistics. Stockholm, Sweden.

European Statistics Code of Practice http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice

EUROSTAT, ESS handbook for quality reports, 2014 Edition
http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

HSCO. *Coherence in Multi-Source Statistics – Literature Review by HCSO*. Version 3. (Working paper).

Gasemyr Svein (2006). Editing and Imputation for the Creation of a Linked Micro File from Base Registers and Other Administrative Data, in *Statistical data editing. Impact on quality*, V. 3, UN, Geneva,312-320

Groves Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, Roger Tourangeau (2004). *Survey Methodology.* John Wiley & Sons.

Hoffmeyer-Zlotnik, Jürgen H. P and Warner, Uwe (2008). Private Household Concepts and their Operationalisation in National and International Social Surveys. Survey Methodology, Volume No. 1, GESIS – ZUMA, Mannheim. https://www.gesis.org/uploads/media/SM1_Gesamt.pdf

Holt, Tim and Tim Jones (1998). *Quality Work and Conflicting Quality Dimensions*. 84th DGINS Conference. Stockholm, Sweden.

Kish Leslie. *Selected papers.* (eds. Graham Kalton, Steven Heeringa). Wiley & Sons, 2003.

Kish, Leslie, Survey Sampling. Wiley & Sons (1965).

Krapavickaitė Danutė, *A glance to coherence*. (Working document)

Lehto Kristi, Maasing Ethel, Tiit Ene-Margit *Determining permanent residency status using registers in Estonia,* Paper presented at the European Conference on Quality in Official Statistics 2016 (Q2016) http://www.ine.es/q2016/docs/q2016Final00155.pdf

Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. Wiley.

Lohr Sharon. L. Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology*, December 2011 197 Vol. 37(2), pp. 197-213, Statistics Canada, Catalogue No. 12-001-X

Meraner et al (2016), Weighting Procedure of the Austrian Microcensus using Administrative Data, *Austrian Journal of Statistics*, Volume 45, 3-14

Myrskylä, P. (2004). Use of Register and Administrative Data Sources for Statistical Purposes. Statistics Finland.

Pettersson H.(2003) Design of master sampling frames and master samples for household surveys in developing countries. Chapter 5. In *UN Handbook on designing of household sample surveys*.

OECD Glossary of Statistical Terms. https://stats.oecd.org/glossary/

Särndal, Carl-Erik. The calibration approach in survey theory and practice. *Survey Methodology*, December 2007, Vol. 33, No. 2, Statistics Canada, Catalogue no. 12001X.

Särndal, Carl-Erik, Sixten Lundström. *Estimation in Surveys with Nonresponse*. Wiley & Sons, 2005.

Särndal C.-E., B. Swensson, J. Wretman. *Model-assisted Survey Sampling*. Springer, 1992.

Statistics Canada. *Quality Guidelines*. (12-539-X) 2009.
http://www5.statcan.gc.ca/olc-cel/olc.action?objId=12-539-X&objType=2&lang=en&limit=0

Statistics Canada Quality Guidelines. Catalogue no. 12-539-XIE, 2003
http://www.statcan.gc.ca/pub/12-539-x/12-539-x2003001-eng.pdf

Sukhatme P. and Sukhatme B. (1970), Sampling Theory of Surveys with Applications, 2nd rev.ed.Ames, IA: Iowa State University Press

Natalie Shlomo (2006) The Use of Administrative Data in the Edit and Imputation Process. In *Statistical data editing. Impact on quality*, V. 3, UN, Geneva, 321-333.

Turner A. G. (2003) Sampling frames and master samples. Chapter 3. In *UN Handbook on designing of household sample surveys*.

Marco Di Zio, Ugo Guarnera, Orietta Luzi, and Antonia Manzari, (2006) Evaluating the Quality of Editing and Imputation: The Simulation Approach, in *Statistical data editing. Impact on quality*, V. 3, UN, Geneva, 44-59.

Waal de Ton, Jeroen Pannekoek, Sander Scholtus. Handbook of Statistical Data Editing and Imputation. Wiley & Sons, 2011.

Wallgren, A. & Wallgren, B., Register-based statistics: Statistical methods for administrative data. Wiley. & Sons (2014)

Zhang, L-C. (2015). On modelling register coverage errors. Journal of Official Statistics, 31(3), 381-396.

Zhang, L-C. (2011). A unit-error theory for register-based household statistics. Journal of Official Statistics, 27(3), 415-432.

Annex III: Requirements for frame contents

The following standards for minimal and optional requirements have been selected in order to be able to compile all products which are assigned to European social statistics whenever the use of frames plays a role.

Regarding the definition of the variables recommended below for inclusion into frames, it is recommended to refer to the current standardisation process of social variables within the ESS.[9]

Table 1 frame units:

Unit	Minimal	Optional
Dwellings or persons*)	X	
Person and dwelling		X
Household		X

*) Since it might be difficult for some countries to include persons and dwellings simultaneously the minimum requirement is set to at least one of the basic units. However it is strongly recommended to include both.

Table 2 variables:

Variable	Minimal	Optional
Dwelling		
Unique dwelling identifier	X	
Regional NUTS 2	X	
Regional NUTS 3		X
City	X	
Address- "Street"	X	
Address "House number"	X	
Address "Door number"		X
Address – "Location within the house		X
Postal code	X	
Size of dwelling		X
Type of dwelling (house, dwelling in a house, etc..)		X
Tenure status		X
Person		
Unique person identifier	X	
Unique household identifier	X *)	
Unique dwelling identifier	X	

[9] Work in progress at the time of drafting these guidelines; see for instance: https://circabc.europa.eu/sd/a/7039be8c-a45a-493f-bc49-987e0ba8f798/DSS-2017-Mar-4.2%20Standardisation%20of%20social%20variables%20%20progress%20report.pdf

Variable	Minimal	Optional
Relation to dwelling		X
Relation to household		X
Sex	X	
Date of Birth	X	
Citizenship or country of birth	X	
Citizenship and Country of birth		X
Occupation		X
Income		X
Highest level of education		X
Marital Status		X
Email address		X
Phone number		X
Household		
In table 1 household are not listed as minimal standard. Therefore the table below suggesting minimal and optional standards for household related variables has to be seen under the precondition that households as composite units are integrated into a frame.		
Unique household identifier	X	
Institutional Yes/No	X	
Size of Household	X	
Type of Household		X
Household income		X

*) Household items are only seen as minimum requirement if households as composite units are integrated

Annex IV: Standardized questionnaire for metadata on frames data which are used by the social and population statistics

This is a questionnaire about the frames which are used by the European social surveys. The questionnaire has several parts. The first one (Part A) identifies the frame(s) used by each survey and the suitable contact person who is the most qualified to fill the other parts. Then, the following parts request further information on each frame:

Part F.i (repeated for each frame #i mentioned in Part A) asks for the sources used to build the frame.

Part S (repeated for each source mentioned at least once in part F) provides more details about the source and its quality.

Part C.i (repeated for each frame #i) deals with the way the source(s) is/are processed to construct frame #i.

Part Q.i (repeated for each frame #i) describes the quality of the frame.

Part I.i (repeated for each frame #i that is used for the LFS survey as indicated in part A) contains basic frame quality indicators to construct the simplified quantitative quality indicator of QGFSS chapter 5.

Part O.i (repeated for each frame #i) contains more comprehensive frame quality indicators to construct the complete frame quantitative quality indicator of QGFSS chapter 5.

An abridged version of the questionnaire will be asked for frames concerning less than 5% of the population. Questions that are kept for these small frames are indicated by an asterisk* appended to the question number in the leftmost column below.

Fields will be prefilled as available from the previous replies.

Questionnaire

Explanations and filling instructions

Part A Identifies the sampling frames for each survey and the suitable contact person who is the most qualified to fill the other parts

A.1 * Surveys and population census

	1. European Union Statistics on Income and Living Conditions (SILC)	2. European Union Labour Force Survey (EU-LFS) (LFS)	3. European Health Interview Survey (EHIS) (EHIS)	4. Adult Education Survey (AES) (AES)	5. Community survey concerning statistics on the Information Society (ICT)	6. Harmonised European Time Use Surveys (HETUS)	7. Household Budget Survey (HBS)	8. Census (Census)
National Name of the survey	*[free text]*	*[free text]*	*[free text]*	*[free text]*	*[free text]*	*[free text]*	*[free text]*	*[free text]*
Statistical units targeted	Individual *Y/N* Household *Y/N* Other *Y/N* if Y, specify *[free text]*	Individual *Y/N* Household *Y/N* Other *Y/N* if Y, specify *[free text]*	Individual *Y/N* Household *Y/N* Other *Y/N* if Y, specify *[free text]*	Individual *Y/N* Household *Y/N* Other *Y/N* if Y, specify *[free text]*	Individual *Y/N* Household *Y/N* Other *Y/N* if Y, specify *[free text]*	Individual *Y/N* Household *Y/N* Other *Y/N* if Y, specify *[free text]*	Individual *Y/N* Household *Y/N* Other *Y/N* if Y, specify *[free text]*	Individual *Y/N* Household *Y/N* Other *Y/N* if Y, specify *[free text]*
Number of frames used	*xx*	*xx*	*xx*	*xx*	*xx*	*xx*	*xx*	*xx*

The **statistical unit** is the entity on which the statistics are calculated, which is not necessarily the sampling unit (see glossary definitions). The same survey might address several statistical units, e.g. statistics may be produced on persons and households.

Sometimes **more than one frame** is used for a single survey, e.g. two frames (census + dwelling register), where each frame is used to produce a separate sample.

A **random route** method is a particular way of sampling where an initial geographical point is sampled with a random path to find dwellings to survey.

Use of random route method?	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*n.a.*
Uses of quota method?	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*n.a.*

The **quota method** is a way of making a sample, controlling some marginals (quotas) like sex and/or age groups.

A.2* **Use of frames in surveys and population census**

	Frame name	SILC	LFS	EHIS	AES	ICT	HETUS	HBS	Census	Contact
Master Sample	…	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	…
Frame #1	…	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	…
Frame #2	…	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	…
…										
Frame #i	…	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	…
…										
Frame #N	…	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	*Y/N*	…

This overview table will organise the rest of the questionnaire. For each survey (column), the table identifies the frames (rows) used. The column 'contact' identifies the person most qualified to answer the rest of the questionnaire for each frame.

For each frame #i created in this table, the parts F.i, C.i, Q.i, I.i and O.i will be sent to the indicated contact person.

A **master sample** is a common (first stage) sample for several surveys. Usually the master sample is used to set the same geographical units for all samples. It is useful to geographically settle a network of interviewers in order to reduce the cost for face-to-face interviews (cf. ' master frame' in the glossary).

A.3* **Average annual mobility rate of households in your country over last 5 years?**

Average over last 5 years of:
(number of households that moved during the year) / (number of households) [%]

The **mobility rate of households** is a regular and foreseeable process that introduces errors in the contact variables and the auxiliary variables, irrespective of the quality of the frame construction. A household that moved should be counted only once and it can be a nuisance for survey takers if the contacts are not updated promptly enough.

A.4* **Average share of new dwellings per year over last 5 years?**

Average over last 5 years of:[10]
(number of new dwellings) / (number of dwellings) [%]

New dwellings are a regular and foreseeable process that introduces errors in the contact variables and the auxiliary variables, irrespective of the quality of the frame construction. The yearly share of new dwellings is defined as the fraction of dwellings that the first residents, after the construction finished, moved in during a given year, and it is occupied as principal residence.

[10] If not available, to be filled with latest available estimate incl. year

A.5*	**Is there national legislation basis for NSI / other statistical authority to access and use of administrative data for statistical purposes?** *Y/N; If Y, specify: [free text]*	As the legal situation in Member States may change over time, this question is to keep the information up to date.

Part F.i Characteristics of each frame #i and sources used to build it (to be filled for each frame listed in A.2)

F.i.1* Type of this frame:

List frame, List frame with area variables, Other: specify … [free text]

A **list frame** is a list of units which may be sampled. A **list frame with geographical variables** is a list of units with geographical auxiliary variables. It allows selecting areas as units of the first stage of the sample design.

If the proposed classification does not fit, use the option 'other' and describe the situation using free text.

F.i.2* Sampling units of this frame used by at least one of the European surveys?

F.i.2.a* Individual:

Y/N

F.i.2.b* Household:

Y/N

F.i.2.c* Dwelling:

Y/N

F.i.2.d* Area:

Y/N

F.i.2.d* Other:

Y/N; if Y, specify … [free text]

The **sampling unit** is the unit which is sampled in the frame with a known probability of selection (see glossary definition).

E.g.,, if the statistical units are households and individuals, the dwellings (or main residence) may be selected to contact the statistical units. Then, the survey could collect information about the household living in the selected dwelling. As well, the individuals may be selected. In this example, the sampling unit drawn from the frame is the dwelling.

F.i.3* Sources used for this frame

F.i.3.a* Population and Housing Census:

Y/N

F.i.3.b* Population register:

This section establishes a list of frame sources.

For each source x with answer 'yes' for one or more frames, a part S.x sheet will be asked to the contact person named in A.2.

Note that sources comprise administrative sources, statistical registers and other statistical products, or other sources to be

Y/N

F.i.3.c* Building/dwelling/housing register(s):

Y/N

F.i.3.d* Social security register:

Y/N

F.i.3.e* Income tax register:

Y/N

F.i.3.f* Dwelling tax register:

Y/N

F.i.3.g* Other tax register:

Y/N

F.i.3.h* Utilities (e.g. water, electricity supply) - either subscription or consumption data:

Y/N

F.i.3.i* Vital events register:

Y/N

F.i.3.j* Employment/occupation/unemployment register(s):

Y/N

F.i.3.k* Education register:

Y/N

F.i.3.l* Business register:

Y/N

F.i.3.m* Other administrative register:

Y/N; if Y, specify …[free text]

F.i.3.n* Other source:

specified: such a multisource picture is more and more the standard situation for many social statistics frames in the ESS.

Y/N; if Y, specify …[free text]	

Part S	**Source characteristics, to be filled for each source F.i.3.[b-n] that is mentioned at least once in F.i.3**	Not applicable to the census of population and housing (F.i.3.a) because the census metadata will be reused.
S.[b-n].1*	**Name:** *[Free text]*	
S.[b-n].2*	**Main unit:** *[Free text]*	Describe the concept of the main unit – defined normally as the one identifying the individual records contained in this source: e.g. person, dwelling, etc.
S.[b-n].3*	**Presence of unique identifier:** *Y/N*	
S.[b-n].3.a*	If *N*, list the variables used for unique identification and/or linkage: *[Free text]*	
S.[b-n].4	**Administrative purpose:** *[Free text]*	The administrative purpose of a statistical register, owned by the national statistical office or another owner, is production of official statistics.
S.[b-n].5	**Organization in charge / owner:** *[Free text]*	
S.[b-n].6	**Regular meetings with the owner / organization in charge:** *Y/N or n.a.*	Applies to **administrative sources** – for other types of sources, the answer option 'not applicable' (*n.a.*) is available.

S.[b-n].7	**Access for statistical purposes guaranteed by law:** *Y/N or n.a.*	Applies to **administrative sources** – for other types of sources, the answer option 'not applicable' (*n.a.*) is available.
S.[b-n].8	**Internal update frequency of the source (inside the responsible department or entity):** *Continuously, Monthly, Quarterly, Annually, No updates, Other: specify … [free text]*	It is important to distinguish this question addressing source updates from the question C.i.3 on frame versions. Moreover, this question addresses the update frequency of a source inside the department or entity responsible for maintaining the source, irrespective of how often source updates are made available to the source users (e.g. the department responsible for maintaining the frame – question S.[b-n].9).
S.[b-n].9	**Frequency with which source updates are made available to the department responsible for frame production/maintenance:** *Continuously, Monthly, Quarterly, Annually, No updates, Other: specify … [free text]*	It is important to distinguish this question addressing source updates from the question C.i.3 on frame versions. Irrespective of the internal update frequency inside the department or entity responsible for maintaining the source (S.[b-n].8), it is important information how often source updates are actually made available to the department responsible for maintaining the frame. This is relevant both for administrative sources from outside the NSI as well as for statistical registers or other sources inside the NSI: statistical sources may be maintained by other departments than the one maintaining the frame in question even within the same office, so it is important information how often source updates are made available for frame maintenance.
S.[b-n].10	**Frequency of the assessment of the quality of the source by the statistical office:** *Monthly, Quarterly, Annually, Other: specify … [free text]*	Quality assessment means any action intended to measure the quality of the source.
S.[b-n].11	**Complete geographical coverage of the entire territory of the country:**	Indicate which main units on the territory of the country are not covered,

	Y/N; if N, specify …[free text]	e.g. overseas territory may or may not be included.
S.[b-n].12	**Describe (as applicable) dwellings not covered by this source:** *[Free text]*	If the question is not applicable, just answer 'n/a'.
S.[b-n].13	**Describe (as applicable) persons not covered by this source:** *[Free text]*	If the question is not applicable, just answer 'n/a'.
S.[b-n].14	**Is the source the result of a sampling?** *Y/N*	'Yes' when this source is already a sample. E.g., re-use of the LFS sample as a first stage to further sample for other surveys; or French rolling census, etc.
Part C.i	**Construction of the frame #i (to be filled for each frame listed in A.2)**	A frame can be constructed either from **a single source** or from **multiple sources**. This distinction is made depending on whether the answer 'yes' is given in F.i.3 either once (single source) or more than once (multiple sources).
C.i.1*	**[*Single source:*] Describe how this frame is constructed and updated from the source mentioned in F.i.3:** **[*Multiple sources:*] Describe how the sources mentioned in F.i.3 are combined to construct and update this frame:** *[Free text]*	Describe how the frame is constructed from a single source or combining various sources, depending on the situation. **If the frame is based on multiple sources, address the following:** • Presence of a common identifier and its degree of harmonisation between the different sources • Presence of common units and their degree of harmonisation between the different sources • Presence of common auxiliary variables and their degree of harmonisation between the different sources • Linkage procedures and linkage difficulties • Share of correct linkage between sources • Share of false links • Share of units not linked
C.i.2	**Frequency of the quality assessment of the frame**	

construction/updating process:

Annually, No assessment, Other: specify … [free text]

C.i.3 Frequency with which new versions of the frame are released:

Continuously, Monthly, Quarterly, Annually, No updates, Other: specify … [free text]

It is important to distinguish this question addressing release schedules of frame versions from the questions S.[b-n].8-9 on source updates.

Updates of (some of) the source(s) used to construct the frame are normally used to update the frame itself following predefined schedules or procedures. For instance, new frame versions may be released on a daily basis (i.e. continuously) based on the latest available source information (irrespective whether sources were updated or not), or whenever (partial) source updates become available, or in fixed longer intervals (cf. 'frozen frame' concept in the glossary).

If the frame versions follow a predefined schedule (i.e. versions of the frame are released regularly to its users), the appropriate reply option should be selected. If, on the other hand, the frame is updated according to a more complex procedure, or in irregular intervals following e.g. the availability of source updates, the particular update procedure should be explained in the free text under 'other'. Also if different parts of the frame follow different release schedules (e.g. weekly versions of contact information, quarterly versions of demographic information, …), this should be explained under 'other'.

Part Q.i Quality of the frame #i (to be filled for each frame listed in A.2)

Describe the current NSI practice in sufficient detail.

Q.i.1-5 are open free-text questions to accommodate the various situations regarding the specific methods for quality assessment and calculating quantitative indicators.

Q.i.1* **Describe the knowledge of over-coverage:**

[Free text]

See definition of over-coverage in the glossary. The free text could address e.g.:

- details on over-coverage studies;
- methods and sources used for assessment and values of estimate obtained;
- if it is randomly distributed or if it concerns particular groups of the population;
- steps implemented to correct over-coverage;
- extent of over-coverage after the corrective steps.

If the over-coverage is partially studied, give relevant details.

Q.i.2* **Describe the knowledge of under-coverage:**

[Free text]

See definition of under-coverage (incl. distinction between design/error under-coverage) in the glossary. The free text could address e.g.:

- details on under-coverage studies;
- methods and sources used for assessment and values of estimate obtained;
- if it is randomly distributed or if it concerns particular groups of the population;
- steps implemented to correct under-coverage;
- extent of under-coverage after the corrective steps.

If the under-coverage is partially studied, give relevant details.

Q.i.3* **Describe the knowledge of duplicates:**

[Free text]

See definition of duplicates in the glossary. The free text could address e.g.:

- details on duplicates studies;
- methods and sources used for assessment and values of estimate obtained;
- if it is randomly distributed or if it concerns particular groups of the population;
- steps implemented to correct duplicates;

- extent of duplicates after the corrective steps.

If the duplicates are partially studied, give relevant details.

Q.i.4 **Describe the knowledge of completeness:**

[Free text]

The free text could address e.g.:

- if you have all necessary information for sample design for each frame use;
- difficulties detected in the process of sampling and/or in the calculation of inclusion probabilities.

Q.i.5* **Describe any methodology not covered by Q.i.1-4 to assess the quality of the frame, including benchmarking:**

[Free text]

Q.i.6* **Content of contact information**

Q.i.6.a* Postal address:

Y/N

Q.i.6.b* Email address:

Y/N

Q.i.6.c* Phone number:

Y/N

Q.i.6.c* Other:

Y/N; if Y, specify …[free text]

Q.i.7* **Contact information error estimates**

Provide error estimates on all types of contact information indicated in question Q.i.6.

Contact information errors [%]	Person	Dwelling
Postal address	*xx*	*xx*
Email address	*xx*	*n.a.*

Phone number	*xx*	*n.a.*
Other*: specify …*	*xx*	*xx*

Q.i.8 Variables used in sample design and estimation

Q.i.8.a List variables used in stratification or for defining stages of the sampling design for

i. EU-SILC: *[free text]*
ii. EU-LFS: *[free text]*
iii. EHIS: *[free text]*
iv. AES: *[free text]*
v. ICT : *[free text]*
vi. HETUS: *[free text]*
vii. HBS: *[free text]*

List variables used for the preparation of each survey (i.e. before field work).

Q.i.8.b List variables used in post- stratification, calibration or non-response treatment for

i. EU-SILC: *[free text]*
ii. EU-LFS: *[free text]*
iii. EHIS: *[free text]*
iv. AES: *[free text]*
v. ICT : *[free text]*
vi. HETUS: *[free text]*
vii. HBS: *[free text]*

List variables used for procedures after the survey field work.

Q.i.8.c Describe quality assessments of variables mentioned in Q.i.8.a-b, if any, and their results:

[Free text]

Q.i.8.d Describe comparisons of marginal distributions of frame variables mentioned in Q.i.8.a-b to an external data source, if any:

[Free text]

Q.i.9 **Use of frame as data input for other (non-survey) uses**

Y/N; if Y, describe the use: …[free text]

Describe non-survey uses of the frame, for instance: editing and imputation, formulating editing rules, as the spine for linkage to other relevant datasets for micro-level editing, modelling, data matching and especially direct tabulation of statistical outputs.

Q.i.10* **Delay between the date of an event in the population affecting the principal contact variable and the date of the respective frame update:**

The average/maximum number of months needed to include a change of a principal contact variable in the frame (the span of time from the date of change to the date of frame update).

Q.i.10.a* Average:

xx [months]

Q.i.10.b* Maximum:

xx [months]

Part I.i Basic quantitative indicator, to be filled for each frame listed in A.2 that is used for LFS

Basic quantitative quality information about frames used for LFS is intended for calculation of the **simplified quantitative frame indicator** described in chapter 5 of the QGFSS.

I.i.1* **Percentage of duplicates:**

xx %

See definition of duplicates in the glossary.

I.i.2* **Percentage of over-coverage:**

xx %

See definition of over-coverage in the glossary.

I.i.3* **Percentage of design under-coverage:**

xx %

See definition of design under-coverage in the glossary of the QGFSS.

Part O.i Additional (optimal) set of quantitative metadata information (to be filled for each frame listed in A.2 as available)

All necessary information to calculate the **complete quantitative frame indicator** introduced in chapter 5 of the QGFSS.

Provide individual indicators only if available.

O.i.1* **Quantitative frame quality indicators:**

	[%]
design under-coverage	*xx*
error under-coverage	*xx*
duplicates	*xx*
over-coverage	*xx*
non-contact variable error rate	*xx*
rate of errorfree units	*xx*

See relevant definitions of the indicators and error types in chapter 5.1.2 of the QGFSS.

Annex V: Glossary

Administrative data source: A data holding that contains information collected primarily for administrative (not research or statistical) purposes. This type of data is collected by government departments and other organizations for the purposes of registration, transaction and record keeping, usually during the delivery of a service. They include administrative registers (with a unique identifier) and possibly other administrative data without a unique identifier.

Address: A number or similar designation that is assigned to a housing unit, business or any other structure. Addresses mainly serve postal delivery, but are also important for administrative purposes, for example in civil registration systems and in census taking.

Area frame: An area frame is a collection of well-defined land units that is used to draw survey samples. Common land units composing an area frame include states, provinces, counties, zip code areas, or blocks. An area frame could be a list, map, aerial photograph, satellite image, or any other collection of land units. An area frame can be seen as an example of an indirect frame.

Auxiliary variable: A variable of the sampling frame which is neither the contact variable, nor the identifier. Auxiliary variables are used to optimize the sample, or to compile detailed tabulation when a frame is used for producing statistics directly, or to enhance other processes like editing and imputation.

Base weight: A factor usually the product of the design weight and a non-response factor assigned to each sampling unit before calibration.

Composite frame unit: A unit which consists of a certain number of smaller, so-called 'micro units' is called a composite frame unit. For example, a group of persons might form a household.

Contact variable: A variable whose values allow contacting a statistical unit in surveys. Normally contact variables used in surveys are postal address the phone number and the email.

Continuous frame: A frame, which is maintained continuously. All units and the variables are updated simultaneously to the occurrence of the change.

Data source: In the context of frame, by data source we mean a structured material from which input for a frame can be taken. The data source can be an internal one (an internal register maintained by a statistical office) or of external nature, like an administrative data source.

Design under-coverage: Design under-coverage comes from the population groups or geographical areas that are excluded a priori for practical reasons of feasibility. They could be e.g., population elements in areas that are remote or difficult to access.

Design weight: The design weight in a probability sample corresponds to the reciprocal value of the probability of being included into the sample. It depends on the sample design.

Due day: A time stamp, normally a fixed day, on which a certain stock of statistical units is taken is called a due day**.**

Duplicate: A statistical unit belonging to the target population that is referred to by at least two units on the frame. An example of duplicates is given by the sampling of phone numbers in the phone book, as some individuals have several phone numbers.

Dwelling: A dwelling is a room or suite of rooms - including its accessories, lobbies and corridors - in a permanent building or a structurally separated part of a building which, by the way it has been built, rebuilt or converted, is designed for habitation by one household all year round. A dwelling can be either a one-family dwelling in a stand-alone building or detached edifice, or an apartment in a block of flats. Dwellings include garages for residential use, even when apart from the habitation or belonging to different owners.

Editing: An application of checks that identify missing, invalid or inconsistent entries or that point to data records that are potentially in error.

Error under-coverage: error under-coverage originates from those frame population elements that are not excluded by design and should be included in the frame but which are actually missing.

Four eyes principle: The four eyes principle is a requirement that two individuals approve some action before it can be taken. The four eyes principle is sometimes called the two-man rule or the two-person rule.

Frame: Any list, material or device that delimits and, identifies, and allows access to the elements of the target (survey) population. Depending on the use case, a frame may allow access to and/or provide additional characteristics of the element.

Frame Error: Frame errors are non-sampling errors caused by the insufficiencies of a frame used in statistical production. It comprises all use cases: As sampling frame, as support in statistical processes and when used for direct production.

Frame owner: The frame owner is a unit within the statistical organization responsible for the construction and maintenance of a frame.

Frame user: A unit of as part of the organisational structure of or a physical person working at a statistical office who is going to use frame data in order to support any process relevant for a statistical product.

Frozen frame: A frame is called frozen when the units and their assigned values of its variables are stable for a fixed reference period (e.g. month or quarter). The attribute frozen refers to the values in the frame as they were taken in for a certain time stamp. It does not hamper the continuous update if useful (for instance of contact variables in order to reach a certain frame unit).

Identifier: An identifier is any variable or set of variables which are structurally unique for every population unit, for example a population registration number. If the formal identifier is known to the unauthorised user identification of a target individual is directly possible for him or her, without the necessity to have additional knowledge before studying the microdata. Some combinations of variables such as name and address are pragmatic formal identifiers, where non-unique instances are empirically possible, but with negligible probability.

Imputation: Imputation is a procedure for entering a value for a specific data item where the value is missing or unusable.

Indirect frame: The units of an indirect frame are related to the units in the target population. It can then be considered to produce estimates for the desired target population by using the links from the indirect frame to the units of interest.

Integrated frame: A frame holding not only one kind of frame units but composite units as well is called an **i**ntegrated frame. In the case of social statistics, households and sometimes dwellings are composite units composed by persons.

List (direct) frame: A frame where the units belonging to a target population are presented directly in a list. Sometimes it is called a 'direct frame'. A typical property for a list frame is that it covers only one type of frame unit.

Master frame: A frame used for a connected group or a family of statistical products.

Master sample: A common first stage for sampling several surveys. Usually the master sample is used to set the same geographical units for all samples. It is useful to geographically settle a network of interviewers in order to reduce the cost for face-to-face interviews.

Multiple frame: A multiple frame scenario is present when the information necessary for a use case of frames is available in more than one frame.

Multiple frame sampling: Refers to surveys in which two or more frames are used and independent samples are respectively taken from each of the frames. A special case is a dual frame scenario.

Over-coverage: Frame over-coverage is the case if it entails duplicated, non-existent or out-of-scope elements. It is common to distinguish between two main types of over-coverage: duplicate listing and erroneous enumeration.

Population register: A statistical register of residing persons normally in a given country. Additionally, it often provides some characteristics of individuals.

Primary Sampling Units (PSUs): Primary sampling unit refers to sampling units that are selected in the first (primary) stage of a multi-stage sample ultimately aimed at selecting individual elements. In selecting a sample, one may choose elements directly; in such a design, the elements are the only sampling units.

Record linkage: Record linkage is the task of finding records in a data set which refer to the same entity across different data sources. Record linkage is necessary when joining data sets based on entities that may or may not share a common identifier, which may be due to differences in record shape, storage location, or curator style or preference. A data set that has undergone RL-oriented reconciliation may be referred to as being 'cross-linked'. Record linkage is called data linkage in many jurisdictions, but is the same process.

Reference period: A certain time period normally a month a quarter or a year – where incidents of statistical interest are counted is called a reference period**.**

Register: A systematic collection of unit-level data organized in such a way that updating is possible. Updating is the processing of identifiable information with the purpose of establishing, bringing up to date, correcting or extending the register, i.e. keeping track of any changes in the data describing the units and their attributes.

Rich frame: A frame consisting of a lot of contextual information (auxiliary variables) allowing for a broader and/or more shaped use of the frame:

- for sampling:- implementing more sophisticated sampling designs (e.g. stratification)
- for direct tabulation: broad variety of cross tabulations

- for weighting and editing: enhancing

Sampling frame: A frame that could be used as a basis for sampling (allows determining probability of selection) and normally is any list, material or device that delimits, identifies, and allows access to the elements of the survey population.

Sampling unit: An object that can be selected with known probability from a sampling frame.

Single frame: A single frame scenario is present when all information of a use case for frames is available in one single frame.

Social statistics: Social Statistics comprises the collection, processing and presentation of data focusing on individuals and their conditions of life and work.

Statistical matching:

Statistical matching (also called data fusion or synthetical matching) aims to integrate two (or more) data sets characterized by the fact that:

(a) the different data sets contain information on a set of common variables and variables that are not jointly observed;

(b) the units observed in the data sets are different (disjoint sets of units)

Statistical product: Statistical products are, generally, information dissemination products that are published or otherwise made available for public use that describe, estimate, forecast, or analyze the characteristics of groups, customarily without identifying the persons, organizations, or individual data observations that comprise such groups.

Statistical register: A register created for statistical purposes normally by statisticians. They are typically created by transforming data from registers and/or other administrative data sources.

Statistical unit: A statistical unit is the unit of observation or measurement for which data are collected or derived.

Survey population: The part of target population from which information can be obtained in the survey.

Target population: The universe about which information is wanted and estimates are required. The target population is the set of the statistical units**.**

Under-coverage: The number of statistical units of the target population which should be included in the frame but do not appear. For example, homeless people which are part of the population are not included in the frame. Under-coverage can be decomposed into design under-coverage and error under-coverage.

Variable: Characteristic of a unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, weight, etc. and "occupation", "industry", "disease" etc.).

Weighting: Act of assigning weights to sampling units, which are then used to obtain estimates of population parameters by calculating weighted sums of observed values.