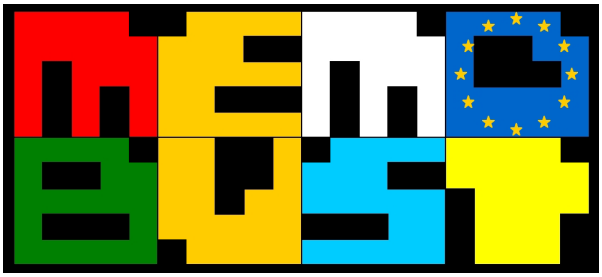

This module is part of the

# Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

# Theme: Collection and Use of Secondary Data

**Contents**

# General section

## 1. Summary

National Statistical Institutes (NSIs) aim to produce undisputed and up-to-date statistics about their society. This requires up-to-date and reliable data. These could be data that the organisation itself collects (primary data) or data that are available in the outside world (secondary data). The latter can, for instance, be administrative sources maintained by other governmental organisations, and sources nowadays identified as 'Big data', such as data available on the internet and data generated by sensors. Mindful of the costs and response burden involved in the collection of primary data, more and more NSIs aim to maximise the use of secondary data for statistics production. The entire process of collecting already existing data is generally referred to as the collection of secondary data. This chapter discusses the advantages and disadvantages of this approach from an official statistics point of view.

In order to be in a position to use data from secondary sources, NSIs need to know which secondary sources exist with respect to their country and if they are allowed access them on a regular basis. Next, the 'fitness for use' of the data source for official statistics needs to be determined. There are many ways to determine this. The most important approaches focus on the metadata quality of the source, on the data quality of the input data, and on the data quality of the statistics produced. When a secondary data source is found suited for use, delivery agreements with the data provider need to be set up. It is considered good practice to assign an NSI-employee as the contact person for the source and the data provider. For important statistics that are dependent on the availability of the secondary data, ways to deal with any interruption or delay in the delivery need to be set up. These so-called fall-back scenarios may range from very simple actions, such as directly contacting the data provider, to the use of complex models that are able to cope with any data missing.

Apart from administrative data, some more recent work also focuses on the use of innovative secondary sources, so-called Big data, for statistics. Since a lot of these projects are still going on and these sources are not used for statistics yet, the focus of this chapter is limited to what is already known on the use of secondary sources for statistics.

## 2. General description

National Statistical Institutes (NSIs) that want to produce undisputed and up-to-date statistics need recent and reliable data. These could be data that the organisation itself collects, primary data, or data that is available in the outside world, so-called secondary data (Hox and Boeije, 2005). Secondary data may be data gathered and maintained by other organisations for administrative purposes (Statistics Denmark, 1995; Wallgren and Wallgren, 2007), or data that is generated by an increasing number of electronic devices surrounding us and on the internet; so-called 'Big Data' (UN Global Pulse, 2012). The latter sources constitute a new and rapidly developing area of the use of secondary data for statistics (Glasson et al., 2013). However, since these sources are not used for statistics yet, the main focus of this chapter is on the – more established – use of administrative sources for statistics.

The remainder of chapter 2 is organised as follows. First, in section 2.1 we will discuss research strategies with secondary data. In section 2.2 we give a classification of secondary data types,

followed in 2.3 by an overview on the different types of use of secondary data by NSIs. We close this chapter by summing up the dependencies of secondary data use.

### *2.1 Research strategies with secondary data*

The technique of acquiring and using secondary data sources is not unique to the field of official statistics. It evidently has multidisciplinary appeal, with extremely diverse academic fields drawing on the information included in secondary sources. All methods used belong to the academic discipline known as secondary research (Golden, 1976; Stewart and Kamins, 1993), which involves using existing data for a purpose different from the one for which they were originally collected.

In general, three different secondary research strategies can be discerned (’t Hart et al., 2005; Golden, 1976): content analysis, secondary analysis, and systematic review. The focus in content analysis is on extracting or summarising the content of various forms of human communication. Frequently used sources include newspapers, books, TV images, websites and paintings. A problem with content analysis is how to satisfactorily categorise and code what is often a large volume of unstructured data. Secondary analysis is about using quantitative data that were previously collected by other people for a different purpose. The general methods of secondary analysis differ very little from those used for primary data sources (Golden, 1976; Wallgren and Wallgren, 2007). Systematic review (sometimes referred to as meta-analysis) combines and investigates the output of multiple studies concerned with the same or a similar phenomenon.

Many NSIs may apply all three secondary research methods. However, without doubt the most commonly used method is secondary analysis, since usually the data content of secondary sources provides input for official statistics. Examples of secondary analysis from official statistics practice are processing of Value Added Tax data, from the tax office, for the short-term business statistics (Constanzo, 2011) and the use of administrative sources containing (human) population-related data for the Virtual Census. The other above mentioned two secondary research methods (content analysis and systematic review) may be less frequently used. A typical example of content analysis is a historical review of an NSI statistic or statistics. Examples of a systematic review are a publication in which time series of trade statistics are compared between various countries and an investigation into the relationship between cancer and nutrition by combining all data published on the subject in the scientific literature over the past 15 years.

### *2.2 On terminology and types of secondary sources*

There is a wide range of terminology and definitions concerning ‘secondary sources’ and ‘registers’ which can be quite confusing. The first two terms we want to clarify is the distinction between a source and a register. SDMX (2009) defines a source as “a specific data set, metadata set, database or metadata repository from where data or metadata are available” and a register as a “data store where registered items are recorded and managed”. The crucial point of a register here is that it is *managed*. The context of the SDMX(2009) definition of a register clarifies this further and explains that a (statistical) register is “a continuously updated list …”, which is also found in the definition of UNECE (2007). In summary, we follow the ideas of SDMX (2009) and use the term source as a general notion for a data set whereas we use the term register as a special case where data are stored and structured in such a way that they can be managed and continuously updated.

To clarify the term register further, we provide some more context. A selection of registers are specifically devoted to maintaining a population of objects by updating any changes in the properties of the objects. Examples are the so-called base registers (UNECE, 2007) that hold lists of objects that are used by public institutions (see below for more explanation) and a business register where statistical units with identifying variables are derived according to Eurostat recommendations. In addition to this however, there are also statistical registers (SDMX, 2009) where a number of data from different sources is integrated and continuously updated for statistical purposes. Thus, in the present paper, the term register is not synonymous to an updated list of objects, since in some registers also many variables are integrated, to be used for statistical output.

We will now further specify secondary data sources, since NSIs exploit a very diverse range of secondary sources. Examples of these are base registers, data on taxes, survey data from another survey oriented organisation in the country, price scanner data of supermarket products, and airline ticket prices from the internet. Some of these sources may be deemed to constitute an administrative source, but the distinction between administrative and other types of secondary data is unclear in some cases. Price data given on a website clearly do not constitute an administrative source, and neither are they maintained for administrative purposes.

From the viewpoint of NSIs information requirement, three main categories of secondary data sources can be discerned: statistical sources, administrative sources and organic sources (slightly modified from Daas and Arends-Toth, 2009). This categorisation is based on assessing the sources against their various characteristics. Figure 1 shows the various categories of secondary sources distinguished. In UNECE (2012) a more detailed list is included. The way by which the sources in each category can be integrated in the statistical process varies (Daas and Arends-Toth, 2009; Groves, 2011; UNECE, 2012) as will be explained below. Some examples are also included in figure 1 for clarification.

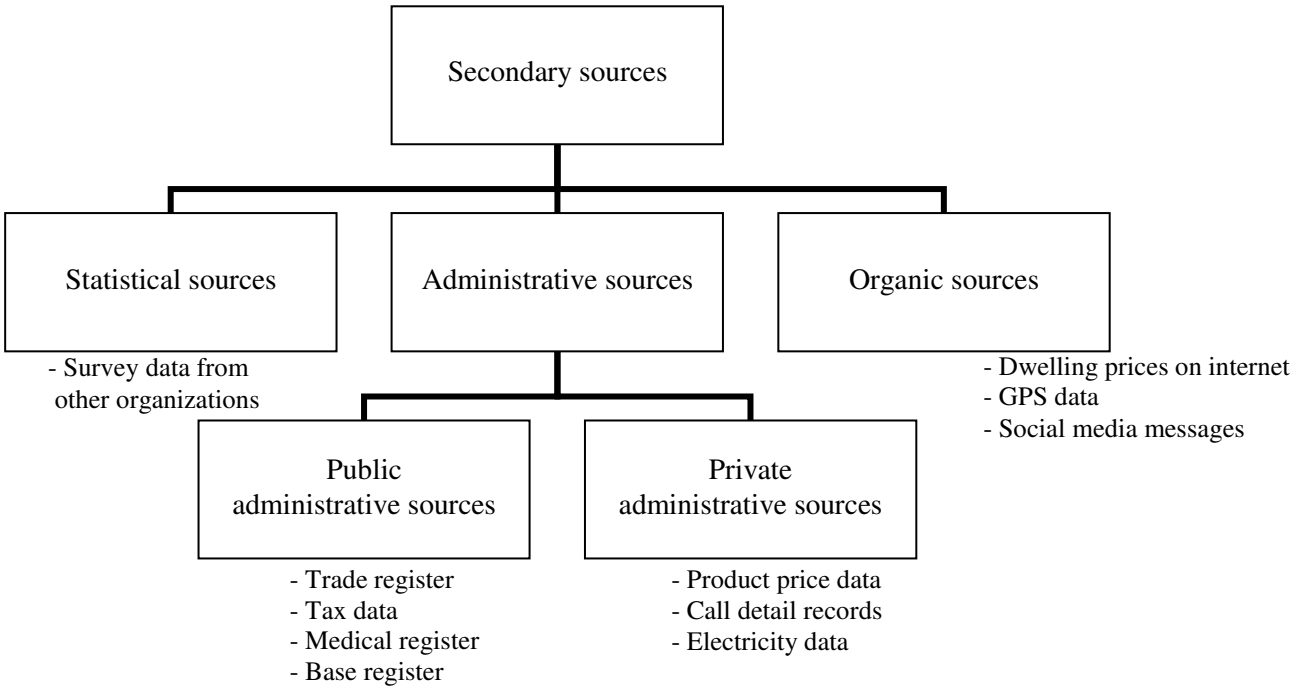


*Figure 1. Distinguished categories of secondary sources with examples*

The first main category concerns statistical sources. Statistical sources consist of statistical objects and statistical data, that can 'easily' be processed for official statistics. Among the statistical secondary sources used by NSIs are survey data collected by other (survey-oriented) organisations, such as those collected by market research organisations, or by government research bodies.

The second category are administrative sources, i.e., secondary sources that have an administrative purpose. Figure 1 shows that the administrative sources can be split into two subcategories namely public administrative sources and those from private (including companies) organisations. Examples of source in the first subcategory are the Trade Register, a National Medical Registration and the Population Register. Examples of the second subcategory are a database with prices of supermarket products, mobile phone call-detail records, and data collected by smart electricity and gas meters. They all serve an obvious administrative purpose.

A special group of administrative data from public organisation are so called 'Base registers'. Base registers are special data sources that form the foundation of many government's implementation tasks in the Nordic countries and in the Netherlands (UNECE, 2007). Base registers contain data that is frequently used by the government in policy, implementation and enforcement. The typical function of base registers is to keep stock of the population of objects at any given time. In addition, they have to maintain identification information to be used by other sources (UNECE, 2007). For instance there is a base register on individuals with data on address, age, sex etc. and a base register on legal units ('firms'), with ownership, kind of economic activity. Since governmental organisations are obliged to use the data in base registers and report any suspected errors in the data, its use will improve the quality of the data. Storing data in a system of related base registers is expected to help improve quality (Wallgren and Wallgren, 2007). Next to the base registers, other (public) administrative sources may hold data that could be useful for some governmental organisations. When exploratory or feasibility studies reveal that using one of these sources might help to reduce response burden, the potential reduction in burden is substantial and that the data will be heavily used; it may be earmarked as a future base register.

Public administrative data can usually be related to a large proportion of the statistical target population. Administrative data from private organisations however, can often be related to only a subset of the target population, with the risk of being selective. Administrative data from private organisation may have to be complemented by other data in order to cover the target population. Examples of private administrative sources are call detail records of a mobile phone provider, electricity data of an energy company and product price data of a supermarket chain. For both subcategories of administrative sources it holds that the concept of the variables may be different from the target one, so derivation or estimation rules are needed to delineate the target variable.

The last group of secondary sources is composed of organic sources (Groves, 2011), indicating the fact that these sources contain data created in a more unstructured way. In fact, a considerable part of these sources can be identified as 'Big Data' (Glasson et al., 2013). Examples of organic sources are a dataset with dwelling prices collected from a website, satellite-based navigation system (GPS) data, and a collection of social media messages. For all these sources no immediate administrative use is foreseen nor do they cover an exactly defined population. The potential application of organic sources for statistics is the focus of several studies currently employed (Daas et al., 2013). Processing these sources poses a number of challenges. First of all, it is often difficult to link those data to a statistical

population, since no identification numbers are available and there is little to no auxiliary data available for those units (such as age, sex, etc.). One of the main advantages of these sources is their volume and their near real-time availability.

We now return from secondary sources that are maintained by external organisations to those maintained by NSIs. Statistical registers (UNECE, 2007), such as the Business Register (BR) and the Social Statistical Database (SSD), are *internal* NSI products. They are compiled from primary and secondary sources, and as such can*not* be considered to be secondary data sources. A characteristic of statistical registers is that they contain an enumeration of statistical object, together with statistical data (properties) of those objects.

### *2.3 Consequences of using secondary data*

NSIs that want to increase the use of secondary data sources for statistics usually aim to lower the response burden of respondents and/or the costs of data collection. Needless to say, the cost aspect is also affected by an NSIs secondary data acquisition expenses and the amount of work needed to transform these data to their requirements. Furthermore, some secondary sources tend to have data about a complete population, which enables the publication of extremely detailed statistics. Moreover, if integration of one or more secondary data sources is successful, new and detailed statistics can be published with no additional response burden (UNECE, 2007; Wallgren and Wallgren, 2007). The various ways in which secondary sources are used in statistics production (section 2.4) clearly show how advantageous the use of the types of data sources can be.

Downside of an increased use of secondary data, is that it makes NSIs become more dependent on:

1) the existence of and access to secondary sources;
2) the fitness for use (i.e., quality) of the secondary sources available;
3) the timely and stable delivery of secondary sources.

Problems in one or more of these dependencies can have serious implications on the production of statistical output. In the most extreme case an NSI might no longer be able to produce some of its statistics, when it solely depends on a single administrative source. The above mentioned three dependencies and the ways developed to cope with them are discussed in section 3.

### *2.4 Types of use of secondary data by NSIs*

The benefits that secondary data sources offer makes them very interesting for statistics production. NSIs accordingly use secondary sources for the following statistical applications:

1) in statistics production as a replacement for primary data;
2) as a sample framework and source of auxiliary information in sample design (see also the topic “Statistical Registers and Frames”);
3) as a source of additional variables to be used for estimates;
4) as auxiliary information to support processing of primary data (e.g., data editing, imputation, calibration of estimates)
5) as input for statistical registers (such as the Business Register).

Also, the data in secondary sources may be ideal for some specific statistical applications, in particular when these data sources cover an almost complete population. These data sources can be used for:

6) detailed publications (such as regional statistics);

7) publications about special (infrequently occurring) events.

Secondary sources that cover multiple time periods and maintain a stable composition, over a relatively long period of time, are also very suited for:

8) detailed longitudinal studies.

The above mentioned uses make secondary data ideally suited for statistics production in our modern world, which demands statistics at a very detailed level without an increase in perceived response burden. Big data sources have the additional potential to produce very timely statistics (Glasson et al., 2013) and may enable the creation of so-called leading or even (nearly) real-time indicators.

### *2.5 Some practical issues concerning collection of secondary data at NSIs*

When an NSI is using secondary data for statistics production some practical processing steps need to be taken. Firstly, the data needs to be transferred in a secure fashion from the data holder to the NSI. Data can, for instance, be transferred on a physical storage medium, such as a hard drive or DVD, to the institute or send electronically (web, email). If this is the case serious data protection measures need to be taken. Data should be encrypted and the decryption key should be send separately. Furthermore, it is good practice to store the data in a general file format, such as XML or CSV.

Secondly, the NSI needs to check whether the data received meets the quality standards agreed upon. This can be done by applying some elementary technical checks, such as whether the format is correct and the total number of columns agrees to the number expected. Thirdly, the data received needs to be uploaded into the data storage system of the NSI. When the data is uploaded and no problems have occurred it is good practice to check the input quality of the data, for instance the completeness of the records. This is described in more detail in section 3.3. This is also done in subsequent processing phases, for instance when different data sources are combined during the creation of statistical registers by micro-integration or other data integration methods. These processing steps however are beyond the scope of the this chapter.

## 3. Design issues

NSIs that use or start to use secondary data become more dependent on the availability of secondary data. Unavailability of a part of the source or the source as a whole can have serious implications on the statistical output. NSIs need to take measures to deal with the consequences of this dependency.

The remainder of this chapter is organised as follows. First, in section 3.1 a way to obtain an overview of existing secondary data sources in the country of interest is discussed. Next arrangements that enable structured access to secondary sources are listed. In section 3.3 data quality issues are discussed, followed by an overview of ways to deal with an interruption in the availability of secondary data. The chapter ends with guidelines on maintaining good relations with the data providers.

### *3.1 Existence*

An NSI that wants to use secondary data needs to know what secondary sources are available in its country. The data protection law offers a good starting point. Countries that have set up a personal data protection act (DLA piper, 2013) generally have an organisation that registers all data sources and organisations that process data in which personal identifiers are included. The official authority responsible for data source registration is – very likely – able to provide a list of all data sources reported to them. For example, in the Netherlands, the website of the Dutch Data Protection Authority (www.dutchdpa.nl) has such a list available on their website. This list consists of all sources in which personal identifiers are included in the country and only lacks of i) sources that are exempted, such as membership and payroll records, and ii) databases used by the police and judicial authorities.

### *3.2 Access*

To enable structural access to secondary data sources by an NSI, special arrangements may need to be made. The statistical offices of the Nordic countries have created an overview of their best practices that facilitate the large-scale use of data from secondary sources in their countries (UNECE, 2007; Statistics Finland, 2004). In summary, these are:

*Legal basis*: Legislation provides a key foundation for the use of secondary data sources for statistical purposes. Data protection arrangements must be part of these provisions.

*Public approval*: The general public must have no objection to the use of 'their' data for statistical purposes. The reputation of a statistical institute as a reliable and eminent user of secondary sources is an important factor in acquiring and preserving public consent.

*Unified identification codes*: It is vital that unified identification codes are used (for the various object types) across different sources. The identifiers enable fast data processing and give rise to fewer linkage errors. Sources without such identifiers can still be used, but costs are higher and their use will result in an increased number of errors (because of incorrect and missing links).

*Reliable secondary data*: The secondary sources used must contain reliable data covering as much of the target population as possible. The use of these sources by multiple official organisations and the population itself increases data reliability and decrease the chance of units missing from the target population.

*Cooperation among administrative authorities*: Effective liaison between the authorities involved in using and maintaining the sources helps in the development of a stable and reliable system of secondary sources. It is important that this is supported up to the highest management level in the organisations involved.

The reader needs to be aware that for the use of individual secondary sources specific agreements need to be made with the data provider regarding the delivery and other issues, such as the possibility for feedback or assistance.

### *3.3 Usability (Fitness for use)*

NSIs will, very likely, use the data in a secondary source for a purpose different from that for which it was originally collected (see also the theme module "Data Collection – Techniques and Tools"). This may give rise to problems. For instance, a source may define an important variable, such as turnover,

(slightly) differently from the one used in official statistics leading to a reduced validity (Scholtus and Bakker, 2013). It is important that an NSI is able to access the fitness of use of a secondary source for official statistics, and to pinpoint the cause of the problem. These aspects are all related to the quality of secondary data. For an overview of the sources of error in secondary sources, the reader is referred to the paper by Zhang (2012).

In recent years quite a number of projects have been (partly) devoted to the study of the quality of secondary data used for statistics. Most noteworthy projects are the BLUE Enterprise and Trade statistics project (BLUE-ETS, 2013) and the ESSnet on the use of Administrative and Accounts data for Business statistics (ESSnet Admin Data, 2013). As such a whole range of possible ways to get grip on the 'fitness of use' of secondary sources is available. Main difference between the approaches developed is their focus. Three general approaches can be discerned which specifically focus on: 1) the quality of the input data of a secondary source (Daas et al., 2012), 2) the quality of the output of the statistics based on secondary data (Frost, 2011; Laitila et al., 2011; Burger et al., 2013) and 3) the metadata quality of secondary sources (Daas and Ossen, 2011).

Since the approaches suggested above complement each other, they – as a whole – constitute a more complete framework with implications of potential value for use in other contexts than first intended (Laitila, 2012). Below a short overview is provided of each of three general approaches discerned. The goal is to enable the reader to quickly decide which approach best covers his or her needs.

*Input oriented data quality*

When a secondary source enters an NSI, assessing its quality as early on as possible may be important for an NSI. When this is the case, the user has an *input oriented view* on the quality of secondary data.. Daas et al. (2013) have developed an evaluation procedure and a report card to structurally note the findings. The quality indicators used are grouped into five dimensions, these are: 1) Technical checks, 2) Integrability, 3) Accuracy, 4) Completeness, and a so-called 5) Time-related dimension. These dimensions contain indicators that specifically focus on: 1) the technical usability of the file and data in the file, 2) the extent to which the data source is capable of undergoing integration or of being integrated, 3) the extent to which data are correct, reliable and certified, 4) the degree to which a data source includes data describing the corresponding set of real-world objects and variables, and 5) the indicators that are time and/or stability related, respectively. To ease the use of the procedure, the indicators have been incorporated in the 'dataquality' package for the open source statistical programming environment R (R core team, 2014). Because the time required to thoroughly evaluate secondary data is a serious issue, a visualisation based approach (a 'tableplot') has also been developed; as a quick and general applicable alternative. This allows the creation of data 'pictures' of sources and subsequent deliveries, enabling a comparison of these 'pictures' for a selected number of variables over time. The reader is referred to the paper of Tennekes et al. (2013) for more details on this topic. The approach has been applied to various administrative sources used in business statistics in several countries.

*Output oriented data quality*

The ultimate intention of secondary data is its use for the production of statistical output. The quality of such output is obviously affected by the quality of the secondary data in the source, by the combination of sources used, and by the quality of the production process itself (Laitila, 2012). This

makes the assessment of the quality of the output based on secondary data a difficult task. In recent years, two ways to determine this have been independently developed.

The first one is described by Frost (2011). This work is based on the dimensions of quality proposed by Eurostat (Eurostat, 2003). The dimensions discerned are: a) Accuracy, b) Timeliness and punctuality, c) Comparability, d) Coherence, e) Cost and efficiency, and f) Use of administrative data. The dimensions each contain indicators that particularly focus on: a) the closeness between an estimated result and the unknown true value, b) the lapse of time between publication and the period to which the data refer and the time lag between actual and planned publication dates, c) the degree to which data can be compared over time and domain, d) the degree to which data that are derived from different sources or methods, but which refer to the same phenomenon, are similar, e) the cost of incorporating admin data into statistical systems, and the efficiency savings possible when using admin data in place of survey data, and f) background information relating to admin data inputs. The framework has been applied to various administrative sources used in business statistics in several countries.

The other approach is based on the framework developed in Sweden (Laitila et al., 2011). This general framework contains quality indicators divided into four groups. The groups discerned are i) Metadata, ii) Accuracy, iii) Integration with a base register, and iv) Integration with other data sources. The indicators in each group report evaluation findings on: i) the information available from the data provider, ii) the results of analysis and data editing of the source, iii) the results of integrating the source with the relevant base register, and iv) the results of integrating the source with relevant other primary and secondary sources. Evaluation starts with the metadata contents of the source followed by the accuracy of its content. The integration steps focus on the incorporation of the source into the statistical system by first relating it to a base register, followed by addresses the issue of how the source can be utilised for improving other relevant statistics produced by the NSI. In the end, the findings are summarised in a quality report card. This framework has been applied to several Swedish sources (Daas et al., 2013).

*Metadata quality*

Apart from the quality of the data, evaluating the metadata quality components of secondary sources is very important (Daas and Ossen, 2011). Next to the general Swedish approach described above – covering both metadata and data quality –, specific metadata quality specific alternatives are available for secondary sources. It is highly recommend that quality evaluation of secondary sources starts with the evaluation of metadata quality. Advantage is that it i) enables the identification of important issues very early on in the process that ii) not immediately a great deal of attention and work is put into the evaluation of quality of the data aspects. The latter is often the case in practice.

Metadata quality evaluation approaches have been described by Daas et al. (2009) and by Verschaeren (2012). In both approaches two different views on metadata quality are discerned, notably: 1) those essential for the delivery of the source and 2) the conceptual metadata quality indicators. The quality indicators in the first view are related to the stable delivery and the continuation of the access to the source by the NSI. The indicators in this view focus on the provider of the source, the relevance of the source, privacy, security and delivery issues, and procedures. In the second view the availability and comparison of conceptual metadata definition of the units, variables, and reporting period(s) in the source with those of the NSI are evaluated. Here the description of the metadata of the provider is

evaluated and compared to those of the NSI. In addition the inclusion of unique keys, and any data checks performed by the provider of the source is studied. The latter is very important process-related meta-information because it highly affects the quality of the secondary data source. Evaluation of both views is guided by and the findings are summarised in a specific metadata checklist. The checklist is regularly applied in the Netherlands (Daas et al., 2009).

An important addition included in the work of Verschaeren (2012) is the explicit mentioning of the keeping a repository of evaluation information on secondary data sources. This assures findings are structurally stored For this approach a pre-evaluation checklist has been created. The list has been tested in Belgium and in the Netherlands.

*Other alternatives*

Apart from the overview provided above, there are also several interesting alternatives suggested by others. Two of them are particularly interesting and they will both be briefly mentioned here. The NSI of New Zealand has proposed a quality framework on administrative sources from a business statistics perspective (McKenzie, 2009). Both a preliminary assessment of data quality and a process management oriented way on quality are discerned. A very different way of dealing with data quality is to increase co-operation with the data provider who could, for instance, implement additional checks upon request of the NSI. Statistics Norway is pursuing this approach (Hendriks, 2012).

### *3.4 Coping with interruptions*

When an NSI starts using secondary data for statistics production, it seriously needs to consider the effect of an interruption or delay in the delivery of the source. The problems that may occur and their effects on the statistics that use the source need to be identified and scored by using risk analysis. Depending on the importance of the statistics based on the secondary sources, the risk analysis may indicate the need for implementing measures to cope with the potential (temporary) loss of secondary data. The combined set of measures constitutes a so-called fall-back scenario. In all situations maintaining good contact with the data provider is essential.

*Risk analysis*

The standard process specification used by NSIs, such as those applied for the required availability of information systems (including databases), can also be used to assess the risk of unavailability of a secondary source. In the Netherlands, a template has been created to determine the need of developing a fall-back scenario for a given statistic (table 1). The risk assessment component of the template estimates whether there is any need to create a fall-back scenario. Among the components considered are the assessment of problems with the delivery of the source, the stability of the delivery, and the impact on the statistical output. If delivery problems are likely to occur, with severe consequences for the NSI, it is advised to draw up a fall-back scenario. In all other cases, the NSI manager responsible for the statistics produced needs to decide whether or not a fall-back scenario has to be developed.

*Fall-back scenarios*

It is unrealistic to prepare fall-back scenarios for all imaginable situations. In our experiences, fall-back scenarios are often tailored to specific situations. The best solution in any given situation will depend on what exactly has occurred, what part of the data is missing and the quality of the data available. The chosen solution *must* also address the costs and time available, which will usually be

short. It is therefore advised to draw up fall-back scenarios only for statistics for which the unavailability of secondary data will have serious consequences. The early detection of potential problems increases the chance of a satisfactory response. This is why active relationship management, contact with the data provider, is very important. For sources on which several statistics depends, more than one fall-back scenario may have to be drawn up.

*Table 1. Dutch evaluation template for fall-back scenario*

*Which statistics are involved?*
- Name
- Division, sector, task force
- Uses the following secondary sources: …

*General information about each secondary source*
- Name of source
- Name of data provider
- Contact person at provider
- NSI contact person for the source/provider
- Other NSI contacts (if any)
- What regular contacts are there between the data provider and NSI?

*Risk assessment*
- How great is the estimated risk of the data provider being unable to deliver the source?
- What are the consequences for the NSI?
- How stable is the delivery of the source?

*Process information of the statistic*
- Are there any alternative sources, or does any research exist which indicates that the data could be derived from a model if the source or any of the required variables are unavailable?
- Possible fall-back scenarios: 1. wait; 2. model-based approach; 3. use alternative source

*Summary*
- Risk of untimely publication or non-publication of the statistic
- Consequences for NSI
- Available alternatives

*Meta-information checklist*
- Update frequency of the checklist
- Date of last update
- Drawn up by:
- Signed (name and position)

No fall-back scenario has to be drawn up for a source that becomes permanently unavailable. In this case, a new statistical data collection process, needs to be started or re-organised in order to satisfy the statistical output obligation.

The transition period may be lengthy. External pressure and publication obligations may necessitate the introduction of other 'creative' temporary solutions in the meantime, such as a completely model-based figure, a nowcast, an expert 'guess', or even the use of the Delphi method. It goes without saying that the use of such temporarily solutions must be communicated clearly to the outside world. The emergency measure applied in the transition period can be viewed upon as a temporary fall-back scenario.

The following general approach is recommended for developing a scenario for dealing with the temporary unavailability of an important secondary source:

1. determine whether it is feasible, in terms of time and costs, for NSI-employees to obtain – preferably via alternative external sources – the missing data elsewhere;

2. apply a model-based approach if there is no alternative for the missing data and some of the data about the reporting period are available. Application is subject to the plausibility of the quality of the results provided by the model, so develop and test the model in advance;
3. notify the important users of the potential consequences of unavailability of the source;
4. postpone publication or decide not to publish at all if the above options are impossible.

Postponement is not an option for very important statistics, such as unemployment or economic growth. In such cases alternatives, for instance other sources or a model, *must* be available. Since it may take considerable time to develop a model, it should be created shortly after the need for a fall-back scenario was identified.

We provide an example to illustrate the advantage of having fall-back scenarios available. In the Netherlands quarterly turnover levels and changes are estimated for populations of enterprises, classified by kind of economic activity, for the Short Term Statistics. These estimates are obtained by combining sample survey data with Value Added Tax (VAT) data, provided by the Tax office, leading to observations for nearly all population units. Because there had been some irregularities in the delivery of the VAT data, a fall back scenario was developed. In this we distinguished two situations (a) the delivery problem is discovered very shortly before the planned publication date or (b) the delivery problem is known well in advance.

In situation (a) we use a model-based approach in which the growth rate is estimated as a weighted combination of growth rate of the sampled and non-sampled units in the population. For the non-sampled units normally quarterly VAT values are available. However, in case of delivery problems only the values from monthly VAT emitters are available for the first two months of the quarter. The VAT-data of those two months are subsequently used to estimate the growth rate of the non-sampled population combined with survey data obtained by directly contacting crucial enterprises and requesting their quarterly turnover. Crucial enterprises are those missing units that have the largest (historical) turnover, up to a certain threshold which is determined by the desired accuracy of the estimate.

In situation (b) we send out a small sample survey, for which stratified random sampling is used. The sampling design has been made in advance and can be used directly when needed.

### *3.5 Contact management*

An increase in the use of secondary sources necessitates good relations with the suppliers of the sources: the data provider. This is an activity that is in the field of relationship management. A way to deal with this is appointing supplier managers for the most important data providers, such as the Tax Administration, the Chambers of Commerce and the owners of the Population Register. These managers are required both to provide and to gather information to and from the sources under the responsibility of their contacts. The duties also may include making and monitoring agreements, managing expectations and detecting new developments. For instance, an appointment for exploratory talks will be made with a view to establish the statistical usability of a potential source. Clear agreements must be drawn up with data providers for sources that an NSI decides to use, covering the delivery of the source (including metadata), the use of the data in the source, and the mutual obligations involved. The agreements must be recorded in a formal contract.

The supplier managers are involved in contacts with the data providers at a strategic level: to enhance the long-term relationship between the NSI and the data provider. It is good practice to also make the supplier manager the NSIs internal contact person for any questions and problems regarding the source, its delivery, and the data provider. As a consequence, nearly all contacts with the data provider are channelled through, or follows consultation with, the supplier manager.

A part of the tasks of the suppliers manager, especially those concerning contacts at operational level, concerning daily production issues, may be delegated to other representatives within the NSI. For instance, an NSI may appoint someone who monitors and verifies whether the agreements on delivery and other quality requirements are met. This representative contacts the data source holder in case the delivery of administrative data is too late or incomplete. At tactical level, consultations between data source provider and NSI may concern desired delivery schemes, and objects, variables and classifications within the data source. These consultations may, for instance, include annual meetings at a high (administrative) level, three-monthly user meetings, or two-monthly bilateral meetings of technical experts. The owners of the statistical process that uses the secondary data will usually join those consultation meetings; supplier managers will not necessarily attend all meetings of this kind. Needless to say, the supplier manager needs to be kept informed on the outcome of all meetings.

## 4. Available software tools

## 5. Decision tree of methods

## 6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

## 7. References


BLUE-ETS (2013), Project description on the BLUE Enterprise and Trade Statistics website. (http://www.blue-ets.eu)

Burger, J., Davies, J., Lewis, D., van Delden, A., Daas, P., and Frost, J-M. (2013), *Quality guidance for mixed-source statistics*. Deliverable 6.3 of ESSnet Admin data, February 2013.

Costanzo, L., Di Bella, G., Hargreaves, E., Pereira, H. J., and Rodrigues, S. (2011), An Overview of the Use of Administrative Data for Business Statistics in Europe. Paper for the 58th Session of the International Statistical Institute, Dublin, Ireland.

Daas, P. J. H. and Arends-Tóth, J. (2012), Secondary Data Collection. Statistical Methods 201206, Statistics Netherlands, The Hague/Heerlen.

Daas, P. J. H. and Ossen, S. J. L. (2011), Metadata Quality Evaluation of Secondary Data Sources. *International Journal for Quality Research* **5**, 57–66.

Daas, P. J. H., Ossen, S. J. L., Tennekes, M., and Burger, J. (2012), Evaluation and visualisation of the quality of administrative sources used for statistics. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.

Daas, P., Ossen, S., Vis-Visschers, R., and Arends-Tóth, J. (2009), Checklist for the Quality Evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Daas, P. J. H., Puts, M. J., Buelens, B., and van den Hurk, P. A. M. (2013), Big Data and Official Statistics. Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium.

Daas, P. J. H., Roos, M., de Blois, C., Hoekstra, R., ten Bosch, O., and Ma, Y. (2011), New data sources for statistics: Experiences at Statistics Netherlands. Discussion paper 201109, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Daas, P. J. H., Tennekes, M., Ossen, S. J. L., Di Bella, G., Galiè, L., Laitila, T., Lennartsson, D., Nilsson, R., Wallgren, A., and Wallgren, B. (2013), *Guidelines on the usage of the prototype of the computerized version of QRCA, and Report on the overall evaluation results*. BLUE-ETS deliverable 8.2, March 2013.

DLA piper (2013), *Data Protection Laws of the world*. Second edition, March 2013. http://www.dlapiper.com/files/Uploads/Documents/Data_Protection_Laws_of_the_World_2013.pdf

ESSnet Admin Data (2013), Project description on the web site of the ESSnet on the use of Administrative and Accounts data for Business statistics. (http://essnet.admindata.eu)

Eurostat (2003), *Quality Assessment of Administrative Data for Statistical Purposes*. Working group on assessment of quality in statistics, Luxembourg, 2-3 October.

Frost, J. M. (2011), Development of Quality Indicators for Business Statistics Involving Administrative Data. Paper for the 58th Session of the International Statistical Institute, Dublin, Ireland.

Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., and Khan, A. (2013), What does "Big Data" mean for Official Statistics? Paper for the High-Level Group for the Modernization of Statistical Production and Services, March 10.

Golden, M. P. (1976), *The research experience*. F.E. Peacock Publishers Inc., Itasca, Illinois, USA.

Groves, R. M. (2011), Three Eras of Survey Research. *Public Opinion Quarterly* **75**, 861–871.

Hendriks, C. (2012), Input Data Quality in Register-Based Statistics: The Norwegian Experience. Paper for the Joint Statistical Meeting, San Diego, U.S.A.

Hox, J. J. and Boeije, H. R. (2005), Data collection, Primary vs. Secondary. *Encyclopaedia of Social Measurement* Vol. 1, 593–599.

Laitila, T. (2012), Quality of registers and accuracy of register statistics. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.

Laitila, T, Wallgren, A., and Wallgren, B. (2011), Quality Assessment of Administrative Data. Research and Development – Methodology reports from Statistics Sweden, 2011:2, Stockholm/Örebro, Sweden.

R Core Team (2014), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Scholtus, S. and Bakker, B. F. M. (2013), Estimating the validity of administrative and survey variables by means of structural equation models. Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium.

SDMX (2009), Statistical Data and Metadata eXchange content-oriented guidelines, Annex 1: Cross-domain concepts. SDMX website. http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf

Statistics Denmark (1995), *Statistics on persons in Denmark, a register-based statistical system*. Office for Official Publications of the European Communities, Luxembourg.

Statistics Finland (2004), *Use of registers and administrative data sources for statistical purposes*. Best practices of Statistics Finland, Handbook 45.

Stewart, D. W. and Kamins, M. A. (1993), *Secondary research, information sources and methods*, second edition. Sage publications, Newbury Park, Ca., USA.

McKenzie, R. (2009), Managing the quality of administrative data in the production of economic statistics. Paper for the 57th Session of the International Statistical Institute, Durban, South Africa.

Tennekes, M., de Jonge, E., and Daas, P. J. H. (2013), Visualizing and Inspecting Large Datasets with Tableplots. *Journal of Data Science* **11**, 43–58.

't Hart, H., Boeije, H., and Hox, J. (2005), *Research methods*, 7th impression. Boom, Amsterdam.

UNECE (2007), *Register-based statistics in Nordic countries – review of best practices with focus on population and social statistics*. United Nations Publication, Geneva.

UNECE (2012), *Using Administrative and Secondary Sources for Official Statistics. A Handbook of Principles and Practices*. (http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf)

UN Global pulse (2012), *Big Data for Development: Challenges & Opportunities*. White paper, May. (http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf)

Verschaeren, F. (2012), Checking the Usefulness and Initial Quality of Administrative Data. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.

Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology, John Wiley & Sons, Ltd., Chichester, England.

Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistics Neerlandica* **66**, 41–63.

## Interconnections with other modules

### 8. Related themes described in other modules

1. Statistical Registers and Frames – Main Module
2. Data Collection – Techniques and Tools

### 9. Methods explicitly referred to in this module

1.

### 10. Mathematical techniques explicitly referred to in this module

1.

### 11. GSBPM phases explicitly referred to in this module

1.

### 12. Tools explicitly referred to in this module

1.

### 13. Process steps explicitly referred to in this module

1.

# Administrative section

## 14. Module code

Data Collection-T-Secondary Data Collection

## 15. Version history

| Version | Date | Description of changes | Author | Institute |
|---|---|---|---|---|
| 0.1 | 29-03-2013 | first version | P. Daas, A. van Delden | Statistics Netherlands |
| 0.1.1 | 22-04-2013 | first revision | M. Murgia | ISTAT |
| 0.2 | 31-05-2013 | second revision | P. Daas, A. van Delden | Statistics Netherlands |
| 0.2.1 | 04-10-2013 | preliminary release | | |
| 1.0 | 26-03-2014 | final version within the Memobust project | | |
| | | | | |
| | | | | |
| | | | | |

## 16. Template version and print date

| | |
|---|---|
| Template version used | 1.0 p 4 d.d. 22-11-2012 |
| Print date | 21-3-2014 17:51 |