eurostat

Doc. Eurostat/A4/Quality/03/item6
Available in EN

Working Group
"Assessment of quality in statistics"

Sixth meeting

Luxembourg, 2-3 October 2003 at 9 h 30

Room Ampere, Bech building

ITEM 6:

QUALITY ASSESSMENT OF ADMINISTRATIVE DATA FOR STATISTICAL PURPOSES

Purpose of the document

The purpose of the present document[1] is to improve the current framework for assessing the quality of statistics based on administrative data registers.

[1] The document has been prepared within the framework of contract 2002-22100-002

Introduction

Administrative data are produced as a result of or in connection with the administrative procedures of organizations. For example, revenue authorities record personal income data through the tax forms filed by the persons eligible to pay tax; customs offices collect imports and exports data by recording the quantity, type and value of products that enter or leave a country; etc .

Administrative data increasingly become an important input to the operations of Statistical Institutes (SIs henceforth) worldwide. Several factors dictate the gradual substitution of administrative data for data collected by surveys:

- Direct data collection incurs large financial costs to SIs which have to cope with shrinking budgets.
- SIs want to reduce as much as possible the response burden they impose on enterprises and households.
- Demand for timely, detailed statistical information (e.g. small area estimates) rises. Administrative data almost always refer to whole populations (as opposed to samples) and therefore can offer enough data for analyses concerning fine subdivisions of the populations.

To give one example of the interest of SIs about administrative data, the factors mentioned above led the US Census Bureau to establish an Administrative Records Steering Committee and an Administrative Records Research Staff and to appoint an Assistant Division Chief for Administrative Records Research. The Bureau conducted preliminary investigations in administrative data issues between April 1996 and December 1998 and has drafted a research agenda (see Prevost and Leggieri (1999)) for the period 2000-2007. Many SIs use administrative data nowadays; reference to them will be made later in this chapter.

SIs have the responsibility to report about their products' quality to users. In the case of the European Statistical System (ESS) member state SIs must also report to Eurostat; standard quality reports have been developed for this purpose. Reporting about the quality of a statistical product includes reporting about the quality of the data it was produced from. Guidelines about the quality reporting of survey data have already been produced.

One difference of administrative data from survey data is that their production is beyond an SI's control. This, first of all means that the SI itself needs information about the quality of the administrative data it uses; let us call the provision of such information internal quality reporting. Another difference is that administrative data quality is determined by factors which do not play any role when survey data are involved; for example, a perfect in any sense administrative dataset may be of low quality for a SI because it does not contain a variable identifying the population units, which would permit effective record matching with the SI's datasets (see more on this later on). This chapter is concerned with quality reporting about administrative data and mainly with internal quality reporting. A subset of the internal quality report can be put in a quality report addressed to Eurostat.

It must be kept in mind that frequently it is very difficult for a SI to assess fully the quality of administrative data. For example, the SI will not be able to assess the measurement errors in an administrative dataset if the producing organization has not studied these errors itself and does not permit the SI access to micro data either. Therefore, some of the quality report items we propose may not be always present in the report because of such difficulties.

The chapter is organized as follows: it begins with a summay of the uses of administrative data in general and with some remarks about internal quality reporting. The next three sections present the specific uses of administrative data in a SI, the operations a SI needs to perform on them and the requirements the SI has from them. We do this because the uses and operations create the requirements which in turn define what is considered as good quality of administrative data. Based on the requirements, the items that should be included in the internal quality report are identified and are presented in the subsequent section. A special section is devoted to quality reporting for combinations of datasets from different administrative sources. The concluding section of the chapter refers to quality reporting from the SI to third parties (e.g. Eurostat) about administrative data and products based on them. References to experiences of SIs with administrative data are given throughout the chapter.

Uses of administrative data

We may identify two broad uses of administrative data:

- **Administrative use**: this is their use by the organizations that produce them. They may be used in order to pursue the organizations' activities, in order to help the managers of the organizations take informed decisions or in order to monitor the performance of the organizations.
- **External use**: administrative data are acquired by the state in order to assist in policy making and for reasons of monitoring and regulating the producing organizations' activities. Moreover administrative data may be disseminated to the academia, research bodies, public bodies and the general public. Part of the external use is the provision of administrative data to national and international public SIs who need them for the production of statistics. This use may more specifically be called **statistical use**.

Administrative data are usually delivered to external users in aggregated form. Microdata are not released unless specific users make specific requests (which usually are accompanied by legal obligation of the organization to release the data) and only if the intended use complies with privacy and confidentiality protection laws. National SIs are entitled by national laws to have access to data from certain administrative sources, determined by the laws. The same laws prescribe in detail the statistical use of these data.

General remarks about the internal quality report

In the remainder of this chapter apart from section 'Reporting to third parties' we consider that *the user of administrative data is a statistician, working for a SI, whose objective is to produce statistics about a particular economic, social or other phenomenon.*

The internal quality report on the administrative data should therefore enable its addressee to assess the degree to which the data are suitable for the user's particular purpose. The report will primarily be addressed to the management of the SI (e.g. to a head of unit) and can be very detailed since the management will have the official statistics expertise required for its apprehension.

We believe that two different types of internal quality reports on administrative data are needed. One type will refer to particular administrative data sources (source-specific) and the other will refer to particular statistical products (product specific). The reasoning for this recommendation is that sometimes no single administrative data source may be associated with a single product and vice versa: Data from the same administrative source may be used in different ways, in more than one statistical products of a SI. For example, they may provide raw data for a product and may be used as a sampling frame for another one. On the other hand, one statistical product may use administrative data from more than one sources.

The source-specific quality report will assess the data quality in general without mention to specific products (unless the source contributes to just one product). It will only mention the products the data contribute to and will refer the reader to their quality reports. The product-specific report on the other hand will assess the quality of the administrative data's features which are relevant to the product. It will also refer the reader to the quality reports about the administrative data sources used. Therefore, the two types of reports will partially overlap but will basically complement each other. Since the reports can be assumed to be available electronically one can easily compile a full internal report about a particular data source by pulling together its source-specific report and the relevant parts of product-specific reports.

Each SI usually has a certain group of administrative data sources from which it collects data at regular or irregular time intervals. The data files provided by the administrative data sources usually retain the same structure (e.g. variables, variable definitions, format, file layout, etc) over long stretches of time. The operations that produce the data may also remain the same for long. Therefore a detailed source-specific or product-specific quality report needs to be prepared only when an administrative data source is first considered for use. The evaluation will be based on information provided by the producing organization and possibly on a pilot use of the data. After detailed reports are available, updates to them may be issued at regular intervals indicating any changes since their previous issuing. Major updates will be required when there are significant changes in the organization's operations, in the structure of the data files or when a new use is considered for particular data.

A shortened quality report or a summary of the internal ones could be forwarded to the providers of the administrative data. Such reports, along with direct communication, may indicate to the providers ways in which they can improve the

quality of their data. In this way SIs will be able to exert some (most likely very limited) control over the administrative data's quality. In evaluating possible improvements, the cost these will incur on the providers must also be taken into account. This concern however is a matter beyond the scope of this document.

Specific uses of administrative data within a SI

A survey of the literature about administrative data reveals their different uses by SIs. Without reference to specific statistical products one may easily identify the particular statistical purpose each use serves. We have grouped the different purposes into categories and we present them here:

Survey design: administrative data are frequently used in the design of surveys. The most common use is to provide sampling frames. Moreover, administrative data give useful information about a population under study. For example, they may give information related to a population's strata: their sizes, their composition (means, variances, proportions of interest), their geographic extent, etc. The specification of stratum boundaries may also be based on administrative data. Finally, one can perform exploratory data analysis on administrative data in order to gain insight into relationships between variables which may be exploited in the design of a survey.

Survey planning: the information obtained from an administrative register is also used for survey planning. It helps, for example, in the specification of the required number of interviewers and in the allocation of effort (e.g. geographical areas) to each interviewer.

Data collection: as is commonly recognized and already stated in this document, administrative data substitute survey data and therefore help SIs to reduce the cost and effort of data collection and the burden imposed on respondents.

Sometimes all the data required for a statistical product are obtained from administrative sources. An example of this is the population and housing census of Denmark which completely relies on registers compiled from administrative data. Borchsenius (2000) describes the transition of Statistics Denmark from a conventional to a register-based census and the registers used for this purpose. The transition took less than ten years. The first full register-based census was that of 1981 and utilized mainly the following registers:

- the Central Population Register,
- the Central Register of Buildings and Dwellings,
- the Central Business Register
- the several Tax Registers
- the Pupil Register

Of all these registers, the Pupil Register is maintained by Statistics Denmark itself; the other registers are maintained by other governmental authorities.

For another example one may look at Falorsi *et al* (2000) who demonstrate an estimation method for labor input indicators, relying entirely on administrative data and which can deal with coverage and measurement errors in them.

It is also very common to combine administrative data with survey data. In other words, some questions that would be asked to respondents during a survey are not asked but data on the corresponding variables (e.g. income) are provided by administrative sources. In this way, besides minimizing cost and respondent burden the SI also reduces the risk of getting false answers (e.g. individuals are arguably less inclined to give wrong income statements to the tax authorities than to a SI).

A different form of partly collecting data from administrative sources is where administrative sources provide all required data for a particular subset of the population while a survey provides data for the rest of it. Myers *et al* (2001) mention that the Annual Retail Trade Survey and the Service Annual Survey of the Census Bureau collect the data for a particular type of firms (those with no paid employees) from administrative sources. The same paper describes the outcome of an experiment conducted in order to show whether administrative data could be used for other kinds of firms. It was found that revenue data for firms belonging in eight particular sectors, with annual revenues below a certain level, could be obtained from administrative sources with an underestimation of the total revenue of the respective sectors by about 3%.

Finally, administrative data can obviously provide data in cases of item or unit non-response in a survey.

Enhancement of a survey's coverage: an out of date sampling frame can cause many problems in a survey. When the fieldwork is conducted, it will be found out that some sample members are not in the locations indicated by the frame. Moreover, portions of the population may have been left out of the frame and therefore, they will not have been taken into account during the survey's design and planning. This will lead to biased survey results. Administrative data may first of all give the correct locations of population units which have not been found in the locations indicated by the frame. Moreover, administrative data may reveal the existence of population units which were not included in the frame. For this reason, prior to conducting the fieldwork, additional administrative data are examined, besides the frame, in order to verify the latter's quality.

Sweet (1997) presents the results of a study to evaluate the ability of administrative data to improve the coverage of the 1996 Community Census in the United States. The study was conducted in three test sites using specific administrative registers (not exactly the same in each site). It was found that very few persons found in the registers could be added to the surveyed persons and this only after careful re-interviewing in the households they belonged to. The author attributes this small coverage gain to the relative inability to match administrative register persons with specific household addresses. It was also discovered that administrative register persons who seemed to belong to a specific household should first be verified with the household. To our opinion these findings do not contradict the potential of administrative data for coverage enhancement; they show that care must be shown in selecting appropriate administrative data sources and in using the right matching algorithms (more on this will be presented in subsequent sections).

Data verification: data collected in a survey are cross-checked and verified when data on the same variables, about the same population units can also be found in a dataset compiled for administrative purposes.

Huynh *et al* (date unknown) is a paper demonstrating this use of administrative data. The authors studied the accuracy of data collected by the Survey of Income and Program Participation (SIPP). This survey is conducted by the Social Security Administration (SSA) of the United States and collects data on the benefits received monthly by persons in the US. For the examination of the data's accuracy the authors used two administrative datasets compiled and maintained by their administration: the Master Beneficiary Record, which splits into Monthly Benefit Payable and Monthly Benefit Credited and the Supplemental Security Record. These administrative data are considered to be very close to the truth. Several discrepancies were found between them and survey data. To mention just two of these discrepancies, it was found out that SIPP overestimated the total Supplemental Security Income (SSI) payments of August 1995 by 22.9% and that only 74.71% of persons receiving both SSI and Old-Age, Survivors and Disability Insurance (OASDI) benefits in March 1996 reported this at SIPP.

Auxiliary data collection: administrative data may provide the necessary data on auxiliary variables which are used for the calculation of estimation weights and in statistical methods like post-stratification, regression estimation, ratio estimation, calibration, etc.

Dorinski (date unknown) reports the results of a study of the effectiveness of adjusting the estimation weights of SIPP data by using administrative data from the Inland Revenue Service and the Social Security Administration. The adjustment resulted in variance reduction for the estimators of most, although not all, SIPP quantities of interest.

Data editing and imputation: data editing is the checking of data for errors while imputation is the replacement of erroneous or missing data with new values. In editing we examine whether the data values obtained from each respondent satisfy certain conditions. Some of these conditions involve relationships between variables that must hold for each respondent. One can conceive situations where the exact mathematical expression of such relationships is estimated from administrative data. On the other hand there are imputation methods where missing or erroneous values are imputed with values given by a model that describes the relationship between certain variables; again such models may be estimated from administrative data. We have not found any specific example of such a use of administrative data though.

The creation of statistical registers: several SIs use administrative data in order to create statistical registers. These are registers where data about whole populations of interest are stored. The commonest examples are human population and business registers. A business register can, for example, contain the following data about all enterprises of a country: name, address, tax registration number, economic activity code, number of employees, last fiscal year's revenues, etc. Administrative data are the major input to these registers. The rest of the data are statistical data coming from surveys. Once a register is created it is updated, either at regular intervals (e.g.

annually), or continuously with a constant flow of data from the administrative sources and with regular maintenance surveys.

Statistical registers are created so that they can be used (for any of the uses outlined in this section) instead of raw administrative data. They are preferred to the latter because the latter type of data rarely satisfy completely the requirements of a SI. On the other hand, the data of the statistical registers are processed so that they conform more to the requirements Another advantage of the registers is that their maintenance is a SI responsibility and therefore the SI has greater control over their quality.

Statistical registers are most easily constructed and are most effective when they are designed at the same time or before the administrative registers from which they will draw data. This reduces the amount of data treatment required before inputting data into the registers. This was the case of Denmark, as stated in Borchsenius (2000). Usually however, administrative registers are in place before the statistical ones and the creation of the latter may be difficult.

Some countries (e.g. Denmark, Finland and The Netherlands) have advanced their human population and business registers to such an extent that they are used instead of traditional surveys for the production of demographic and economic statistics. Borchsenius (2000), already mentioned earlier, describes how the Population and Housing Census of Denmark was conducted with the use of registers only. The statistical registers system of Denmark is described in Eurostat (1995). In a similar vein Carling *et al* (1996) present the statistical registers system of Sweden. Methodological issues related to the construction, maintenance and use of statistical registers are presented in detail in Eurostat (1995) and more briefly in Wallgren and Wallgren (1999). Office for National Statistics (2001) presents a review of the inter-departmental business register it uses while Vale *et al* (2001) present the register review methodology and give guidelines about how it can be used by other SIs with their business registers.

Long (1996) gives a general presentation of the use of administrative data in demographic applications. The methods presented comprise the use of administrative data in order to fill in missing items (non-response) in Census questionnaires, their use in order to estimate relationships between variables and apply them to survey data, their use in order to substitute administrative data for questions that would be asked in a survey and their use in order to estimate longitudinal models about demographic parameters.

Operations related to administrative data

When a SI uses administrative data there is a series of operations that always need to be performed. We present them here in order of application because they will help us understand the SI's requirements from administrative data. The operations are the following:

- ***Identification of data sources***: the kind of data required by the statisticians of the SI for a particular purpose may reside in more than one administrative data sources. The statisticians have to identify all potential sources, and to examine their suitability in order to select the appropriate ones. Very often a single data

source may not be adequate to cover all their needs; a combination of sources may be deemed necessary in some situations.

- ***Record matching***: very often a SI requires administrative data about particular population units for which it already has some data (e.g. when requesting the incomes of the persons in a sample). In other cases the SI needs to combine information about population units from two or more administrative data sets (e.g. when obtaining income of persons from one dataset and the persons' marital status from another). It both cases the appropriate records in the administrative datasets must be identified and must be associated with the correct population units. This is called record matching and is not always a simple task; it becomes more difficult when the number of administrative data sources used simultaneously increases. Record matching has been the subject of considerable research in the official statistics community; for an overview of respective research see Winkler (1999).
- ***Data collection***: the extraction of the necessary data from the identified data set.
- ***Data treatment***: the data in their raw form may not be 100% conformant to the SI's needs; they will therefore require treatment. This may involve editing, imputation, transformation of values, creation of new variables, etc.
- ***Use of the data***: when the data have been collected and processed to bring them in the condition required by the SI they are used as input to the statistical production processes.

Statistical requirements on administrative data

We can identify the statistical requirements on administrative data by examining the previously mentioned purposes they serve in a SI and operations performed on them. The requirements can be stated as follows:

- Each administrative dataset must be accompanied by metadata about its contents so that users may assess their suitability for their purposes. The metadata must describe, or point to documents that describe, among other things, the administrative procedures that create the data, any important administrative events relevant to the data and definitions of concepts, variables and the population they refer to. For a concise presentation of the metadata requirements from administrative datasets see Froeschl and Grossmann (1999).
- Administrative data must be suitable for the purpose for which they will be used. For example, there is no point using a certain population register as a sampling frame if it gives no information that will help locate the population units.
- Administrative data must agree with the concepts of the survey in which they will be used. Severe differences in the definitions of variables for example, may render the administrative data unusable. Differences are more likely to appear as the number of administrative data sources used in a survey increases, since concepts must agree between all data sources. The measurement methodology used for the collection of the administrative data also affects their agreement with the concepts of the survey.

- A related requirement is that the reference times of the administrative data sources should be the same as the reference time of the survey in which they are used. Since this is rarely the case, the reference times should at least be specified so that appropriate adjustments can be made.
- Administrative data must provide adequate coverage of the population under study and must not contain, as far as this is possible, duplicate or misclassified entries. Problems in this area make more difficult their use in survey design (wrong picture of the population) and can lead to population representation problems (extensive over- or under-coverage may lead to serious biases).
- Administrative data must be accurate, so that they do not reduce significantly the final statistical product's accuracy. This does not necessarily imply that they must be accurate at the micro level. If administrative data are to be used for sample size allocation to strata, accuracy of the administrative figures for strata totals is adequate.
  Myers *et al* (2001) describe a situation where amounts of money in thousands of cents had been wrongly keyed in as thousands of dollars in a dataset of the US Internal Revenue Service. They also present an edit method devised by the Census Bureau in order to identify and correct this mistake and to make the data usable.
- Administrative data files from the same source must be stable through time in every respect (file structure, variables contained, concepts, etc.). Ideally, only data values may change.
- It must be possible for the NSI to obtain administrative data in a form that suits its needs. Confidentiality constraints for example may not permit the delivery of some variables or the delivery of the data below a certain level of aggregation. A different problem occurs when the administration and the NSI use highly incompatible database systems which makes the exchange of data between them very expensive in time and effort.
- The structure of the administrative data files must permit effective record matching with survey data files or with other administrative data files. This means that each dataset must contain a variable or combination of variables which uniquely identifies each population unit (a unique key in database terminology). The use of the same unique key in all datasets (a common identifier of population units) is the best means to facilitate record matching.
  There are computer programs which perform matching in the absence of common unique keys but they are not yet 100% reliable. The results must be manually reviewed and ambiguities resolved. We elaborate more on record matching in a subsequent section.

These requirements are not independent of each other. For example, difference in concepts between an administrative data set and a survey may lead the former to over-coverage of the population under study.

A second example, reported in Wallgren and Wallgren (1999), demonstrates how differences in concepts can lead to matching problems: The example was about the creation of an Agricultural Register for Sweden with the combination of data from three datasets. The first dataset was a census based Farm Register, created by Statistics Sweden, which listed agricultural holdings; the second dataset was an administrative register listing claims for agricultural subsidies; the third dataset was the Business Register of Statistics Sweden which, obviously, among all the Swedish

enterprises also listed the agricultural ones. When matching was attempted there were situations where one record from one of the datasets corresponded to more than one records in another (for example, many claims for subsidies by the same enterprise). Inspection of the results showed that most of the 'multiple' matches were genuine ones; this example however shows what problems could arise in other datasets where many false 'multiple' matches, difficult to be resolved, could emerge.

Netterstrom (2000) presents the requirements of NSISs have from administrative data. It has served as a source for some of the material of this section and as a stimulus for the addition of further material. Marquis *et al* (1996) present simple studies of the extent to which some administrative registers in the United States satisfy some of these requirements. The studies offer insights into the importance of the requirements. Scalia (1999) finally, presents problems in the use of administrative data for the production of criminal processing case statistics, which stem from the non satisfaction of the above mentioned requirements.

The stated requirements help one to identify the necessary quality components that must be included in the internal quality reports of the SIs about administrative data. Reporting about these components is the topic of the following sections.

Internal quality reporting on administrative data

In this section we make a first proposal for the content of the internal quality report. As we have already stated this is a report concerning an administrative data source's usefulness for a SI and is addressed to the SI's management. Some quality aspects are source specific and others are related to the data's use for a particular statistical product. We do not adhere strictly to the seven components of quality recognized by Eurostat because they relate to statistical products while we refer to the administrative data as inputs to statistical procedures.

Reporting requirements have been grouped under meaningful component headings. We mainly present source specific requirements and make also reference to product specific ones. The components of the internal quality report are:

- ***Clarity***: this component contains the evaluation of the metadata content of the administrative datasets. The reporting must describe what metadata are given and what is missing. A list of metadata categories (e.g. legal context, definitions, methods of data collection, etc) can be given, with a tick mark on those categories which have information provided about them. Instead of a plain tick mark the quality of the provided information can be represented by a grade (e.g. on a five point scale). Space for notes about each category must also be provided. The lack of specific metadata may lead to inability to evaluate a source's usefulness; such problems must also be reported.
  There are no product-specific requirements for clarity which are not part of the source specific ones.
- ***Administrative concepts***: Since potential statistical uses of an administrative dataset are not known in advance the source specific reporting should allow the report's reader to understand the administrative concepts. The presentation should include the definitions of population units and variables and other

administrative concepts and a description of the administrative procedures that are used to collect the data.

In product specific reporting a direct comparison of administrative and statistical concepts can be made and differences may be presented. An evaluation of the possible effect of differences on the final product as well as possible ways to adjust for the differences can also be presented in the report.

- ***Coverage***: this component obviously refers to the extent of the coverage of the administrative dataset. Reference to specific coverage problems (over-coverage, under-coverage, misclassification, duplication) may not be possible with no specific statistical product in mind. The source specific report must at least contain a precise definition of the population units included in the dataset (e.g. "the dataset contains data on every enterprise with more than five full time employees in 2001").

  The product specific report, where a particular statistical population that must be covered is well defined, will contain a more detailed presentation of specific coverage problems.
- ***Reference time***: the report must give the reference time of the records of the dataset. For statistical purposes the reference time should be the time of occurrence of registered events but sometimes administrative data refer to the time events are reported (although this is also important for some uses, (Borchsenius, 2000)).

  The product specific report will, in addition to the above, contain an assessment of the effect of any reference time problems on the product.
- ***Data freshness***: data in registers can become outdated as time passes. For instance, a business register that is updated annually will not include or will misclassify enterprises that started or changed their activities during the year. The internal quality report should present any such problems identified in the dataset. It will naturally present the time that has lapsed since the last update of the administrative dataset and the likely extent to which the data are outdated.

  The product specific report should present an evaluation of the effect of any lack of 'data freshness' on the product.
- ***Errors in the data***: this report component refers to all errors that exist in the data. These are measurement, processing and non-response errors. Usually the SI will not be able to assess directly the extent and magnitude of such errors. The administration producing the data on the other hand, may have such information and this should be made available to the NSI.

  If the NSI possesses data about the same phenomenon from a different, reliable source then it may be able to study these errors.

  With a specific statistical product in mind the report can present the errors' effect on the product's accuracy. As we have already mentioned, errors at the micro level when the product requires accuracy only at the macro level do not invalidate a dataset. Therefore the product specific report will place emphasis on the error aspects that affect the product.
- ***Completeness***: this item appears only in the product specific report and refers to whether the administrative source will cover all the data needs about the product. The report should therefore present the needs and the degree of their satisfaction.

In a source specific report information about completeness is covered by the presentation of the variables in the dataset ('Administrative concepts' component) and of the population units covered by it ('Coverage' component).

- ***Record matching ability***: The ability to match records between the NSI's and the administrative data files. This may include the presentation of any existing common identifiers of population units in both data files (such as the unique identification number each person has in Denmark). If no efficient common identifiers exist there should be an effort to identify other fields that can be used for record matching and an evaluation of the effectiveness of record matching.
  In general no reference to specific product is required for the assessment of record matching ability. The NSI may perform test matches between the administrative dataset and datasets that refer to similar populations.
  If record linkage is performed for the production of a statistical product then information on matching quality will also be included in the quality report about this product.
- ***Confidentiality and privacy protection***: any issues related to confidentiality or privacy protection that may impose constraints on the availability of administrative data to the NSI at a desired level of detail must be reported. This may also require a reference to the relevant legislation and even presentation of any steps (e.g. a specific legal procedure) that will allow the collection of the data in the desired level of detail
  A product specific report will naturally refer to whether the data can be made available at the specific level of detail required for the product.
- ***Compatibility between file formats***: the format in which administrative data are made available to the NSI must be reported. One can envisage extreme cases where large costs in effort, time and money may be required in order to input the data in the NSI's information system.
  The requirements in a product specific report are covered by the source specific one.
- ***Comparability of administrative datasets in time***: the report must give all the necessary information in order to assess the comparability of its data through time. Therefore any administrative events that led to changes in definition, data collection methods, file structure and format must be reported and assessed.
  Product specific reports will simply add to the above mentioned information an assessment of how any lack of comparability in time affects the product.
- ***Envisaged uses of the data***: the source specific report must state what is the envisaged use of the data. This includes reference to all statistical products which are known, at the time of the report's compilation, to require input from this administrative source. This information will allow combination of the report with the product specific reports.

Quality reporting when combinating data sources

The intention of NSIs to use efficiently their resources and to take advantage of any data available leads them very often to the use of combinations of data sources, be they survey or administrative ones, for the production of statistics. We have already mentioned this issue and we have also given some respective quality and reporting

requirements. The purpose of this section is to give a more full presentation of the topic.

The situation is as follows: data about a certain population of interest exist in at least two separate data sources. The NSI combines the separate sources and creates a new dataset (physically or virtually), each record of which corresponds to one population unit and contains all the information about this unit previously scattered in the multiple sources. The two most frequent cases where the need for such combinations of data sources arises are:

- the creation of statistical registers, and
- the supplementation of survey data with administrative data.

The quality of the combined data is not always the 'sum of parts'. If the combination is conducted effectively the resulting quality will be superior to that of each source considered alone; if on the other hand the combination is poor the resulting quality will be inferior to that of the separate sources. Besides the quality of the individual data sources, two other important defining factors of combined data quality are, (a) the combination potential of the sources and (b) the competence of the NSI in combining them.

In order to understand how these factors affect quality we present very briefly the most common way of combining data sources. For simplicity let us assume that two data sources are to be combined.

Usually the sources will not describe exactly the same population, will use different population units and will have different reference times. In the already mentioned example by Wallgren and Wallgren (1999), a census-based Farm Register referring to agricultural holdings with particular characteristics was combined with a Business Register referring to all legal units (among which agricultural ones) active one year before the register's publication. Based on this information the NSI identifies the reference population of the combined data, the population units they refer to and their reference time.

Subsequently the NSI must link –or match- the records of the two sources. It therefore needs to identify and combine the records that correspond to each population unit. In other words it needs proper identification variables and a record-linkage method.

Identification variables may be names (of persons or enterprises), addresses, telephone numbers, other identification numbers (PINs, Social Security numbers, VAT numbers, etc), activity codes, occupational codes, etc. Variables may be used in isolation or in combination. The quality of these variables' data is naturally very important. Of special concern are: missing values, lack of homogeneity in recording (e.g. of first names or addresses), and type-in errors. The data may require extensive treatment by clerical staff or by software in order to correct them and bring them in homogeneous form, suitable for record matching.

After the data have been treated they are processed by record-linkage software in order to produce the combined dataset. Record-linkage methods rely on the model of Fellegi and Sunter (1969). For each pair of records from the two sources (a potential

match) the methods identify the pattern of agreement (e.g. the two records agree in first id variable, do not agree in second id variable, agree in the third and fourth id variables, etc). Depending on how much agreement the pattern shows and on how rare it is, the record-pair is declared a certain match, a certain non-match or is deferred to clerical review for final decision. Other variants of record linkage rely on agreement patterns and also on the specific values the id variables take in the pair of records. In these variants the frequency of appearance of the specific values plays a role in the decision. For more details one may refer to Winkler (1999) or Yancey (2000) and the references therein.

It is not obligatory that there is a 1-1 matching correspondence between the two sources; depending on the population units of the two sources one record from one source may legitimately match with more than one records from the other source. Such was the case in Wallgren and Wallgren (1999) where a file of agricultural holdings was linked with a file of claims for subsidies: there were owners who had filed more than one claims for their one holding and owners of more than one holdings who had filed one claim for all of them. In such a case the use of record linkage methods that force 1-1 matching would clearly be inappropriate. The NSI should therefore apply candidate record-linkage methods to test datasets, resembling the ones it wants to link, in order to test their efficiency and to identify the extent of genuine multiple matches.

Finally, after record matches have been identified, the combined dataset is created. Data may need some final treatment before combination in order to conform to the concepts of the resulting dataset (e.g. change of units of measurement, summation of values belonging to the records of a multiple match, etc.).

The items that must therefore be included in a quality report on combined use of multiple data sources are:

- Concepts: a detailed definition of the population covered by each source, the population units it is divided into and its reference time. Also a definition of reference population, population units and reference time of the combined data. Demonstration of how the population units and the reference time of the combined data were identified. Comparison of reference population with the target population (the one the NSI intended to cover with the combined data).
- Other source metadata: a description of the data files of the sources (storage format, file structure, variable names and definitions, etc.). Moreover, a brief presentation of the survey or of the administrative procedures used for the collection of the data.
- Identification variables: presentation of the variable (or combination of variables) that is used for the identification of appropriate records in each source. Presentation of missing value rates, degree of homogeneity and extent of type-in errors of these variables, prior to and after data treatment applied in order to improve their quality.
- Record linkage methodology: a brief presentation of the record linkage algorithm and reference to relevant methodological documents / scientific publications that describe it in full.
- Matching error rates: presentation of false match rates and false non-match rates. The former is the percentage of reported matches which do not correspond to true matches, while the latter is the percentage of genuine

matches which were not reported. The calculation of the rates requires pilot application of the record linkage method to test data files for which the NSI knows which record pairs match.

- Alternative methods: if the NSI wishes so it may report the two previous items for any other record linkage methods\ it applied to the test datasets. This will help third parties to judge whether the NSI has chosen the most appropriate method. Data about the cost of the methods in money, effort and time will be useful in this respect.
- Overlap of sources: presentation of the percentage of records of each source which were matched to the other sources. Presentation of the corresponding percentage of the target population covered by the combined data. Obviously, if a small portion of each source's records is used and a small percentage of the target population is covered the combination of sources may not be beneficial.
- Quality of produced dataset: the quality of the dataset resulting from the combination of the sources is reported like that of any other dataset (see previous sections). The NSI must additionally report the treatment it applied to the data when combining them.

Reporting to third parties

This section, as opposed to the two previous ones, refers to the reporting a NSI makes to third parties (e.g. to Eurostat) about the quality of statistical products it has produced based on administrative sources. The products may be statistical figures (e.g. average income of the population, volume of exports, etc) or new datasets (i.e. statistical registers). Now the term 'user' refers not to the NSI but to the recipient of the statistical products.

Quality reporting about statistical figures is not different from the reporting about survey based statistical figures. From the point of view of such reporting the kind of data used to produce the figures is in itself irrelevant; the NSI reports about the final product. For this reason, the seven recognized data quality components can now be used. Below we mention the reporting needs of each component. Emphasis is placed on presenting the differences from quality reporting about survey-based figures.

Relevance: This attribute refers to the extent to which the figures satisfy the needs of the users. There is a fundamental difference however, between survey and administrative data based statistics. In surveys the NSI has the ability to adjust concepts and methods to user needs if it cares to identify and evaluate them. In the case of administrative data the NSI has little, if any, say into what kind of data should be collected, how often etc. This means that even if the NSI's priorities are aligned with user needs the administrative data may not be so. Differences between administrative and statistical concepts have therefore a big impact in relevance as they do in other quality components as well and should be thoroughly reported.

Accessibility and clarity: accessibility to the final statistical product has no difference as a quality aspect of administrative data from that of surveys.

Clarity however may be quite different in terms of the metadata content required for proper use of the statistics. In many cases meta-information for statistics based on administrative data is quite different from the meta-information needed for surveys. Some important examples are:

- The description of underlying laws governing data acquisition by the administration. This is important even in cases of uniform laws (as in situations covered by European Legislation) where the implementation may be different and should be described as well.
- The description of the way the reference period of administrative data was determined.

It should be kept in mind that some of the crucial metadata may not even be available to the NSI. Indeed the NSI may often receive a file extracted from an administrative register without much of the description of concepts and methods which usually accompanies statistical data released to the public. It therefore might be useful to report which metadata items, considered important by the NSI, where not made available to it.

Completeness: this attribute refers to the extent to which the released figures provide all the information sought by the users. Completeness can be expressed in terms of population domains covered, the number of statistical figures provided, the level of detail information is released in, etc. The assessment of completeness is done in much the same way as with survey-based statistics. The only difference might be in the causes of non-completeness. When dealing with administrative data the cause of missing variables is that there may be a discrepancy between what is needed for a complete data set and what is available from the administrative sources.

Timeliness and punctuality: timeliness is generally measured as the time lapsed between the end of reference period and the date the statistics become available to users. For administrative data this depends very much on whether they are available to the NSI when they are needed. It also depends to the freshness of the data. Unlike surveys, many administrative registers may contain records that, for a number of reasons, have not been updated although they do not reflect the present situation. Data freshness is a timeliness problem since it results into reporting information that is out of date. For a user it has the same effect as the lack of timeliness in surveys where the user is forced to use statistics referring to earlier dates since more current information is not available[2].

Punctuality, the difference between delivery date and target date, depends on the regularity and consistency with which the administrative data become available to the NSI. Unlike surveys, improvement in terms of punctuality may be beyond the NSI's control. The reasons for any unusual lack of punctuality however, are important and should be reported.

Coherence: coherence of statistics is their adequacy to be reliably combined in different ways and for various uses. Coherence refers, among other things, to the differences between provisional and final figures. Other types of coherence or the lack

[2] Freshness can also be seen as an accuracy problem (more precisely as a measurement error) since the data provided is erroneous. We have elected, however to classify it in the timeliness component.

of it are the discrepancies between the figures provided by a particular administrative data set and the figures given by other data sets describing the same phenomenon. The identification of possible incoherence is quite similar and relates to the consistency of different estimates. The reasons however that lead to lack of coherence may be different when one or both of the data sources under consideration are administrative sources (for instance conceptual differences between different administrations or between administrative and statistical concepts).

Usually coherence is assessed between datasets originating from the same country (and sometimes from the same agency). There is a number of cases however where administrative data from different countries may be checked for coherence. This is the case with flows (of goods, capital, travellers etc.) where the outflow from one country to another must be equal to the inflow for that other country (Carson and Laliberte, 2002).

Comparability: comparability refers to the ability of statistics to be compared in time, in space and between domains. It is reduced by conceptual and methodological differences of the statistics under consideration. Lack of comparability can be evaluated using relevant metadata that thoroughly describe the concepts and methods used. What is different between administrative and survey based figures are the factors that might affect their comparability. For example, administrative sources are usually affected by changes in the legislative environment under which the relevant administration operates. It should also be mentioned that the situation becomes quite complicated when one needs to check if statistics from administrative sources and surveys are comparable.

Accuracy: this attribute refers to the closeness of the numerical information conveyed by statistical figures and the truth the information represents. The NSI's knowledge about the administrative data's accuracy will usually be derived from the producing organization. Taxation authorities for example, check a part of the income statements they receive from the public, which should give an indication of their accuracy. Other times it might be possible for the NSI itself to conduct a study of the data's accuracy.

One of the defining factors of any figure's accuracy, the effect of sampling, is absent from figures based on administrative data if the latter are complete enumerations of the population under study. In this case no information on sampling variation can or needs to be computed. If sampling is used then sampling errors are introduced and can be assessed in much the same way as in statistical surveys. Other parts of the accuracy component should however be considered.

- ***Coverage problems***: if an administrative reference file (e.g. a population register) exists, coverage errors are conceptually the same and can largely be assessed as the coverage errors in the sampling frames of surveys. Sometimes such a file will not exist and cases will be included in an administrative file as they are recorded by the relevant administrative unit (e.g. imports and exports at a customs office). In that case the evaluation of coverage errors requires detailed knowledge of the administrative processes. The evaluation of under-coverage for example, will require knowledge of the extent to which cases avoid the attention of the administration. The reasons that cause these errors should be reported.

- ***Processing errors***: such errors are present in the case of administrative data; their assessment however is rather difficult because they are probably introduced by the administration. Indeed if the data is in the form of micro-data the administration is responsible for coding, keying and editing while if macro-data are sent to the INS then the administration may also be responsible for weighting and tabulation. The assessment of processing errors therefore should probably be based on information the NSI receives from the administrative source.
- ***Measurement errors***: measurement errors exist in administrative data in much the same way as in surveys. Enterprises and individuals use forms to provide the data to the authorities that may lead to errors just like questionnaires can. Public servants may also affect the way the respondents report their data just like an interviewer may do. Finally some of the information provided might just be erroneous. Regarding macro-data, misclassification of population units by the administration may lead to erroneous aggregate figures for subpopulations.
- ***Non-response errors***: Unit non-response is less of a problem with administrative data. Item non-response however can arise because of omissions of the reporting units when filing administrative forms. (Eurostat D3, 1996). Omissions may unfortunately be systematic.

Regarding the quality of statistical registers we view reporting about it as similar to internal quality reporting about administrative data. The statistical register in other words may be viewed as an "administrative register" which is considered for use. The reports will have the same contents as internal quality reports. Input for the reports will be provided by regular reviews of the registers' quality which may be carried out by the NSIs themselves or by independent reviewers. Since the production of statistical registers is fully under the control of the NSI, the reviewer will be able to evaluate all quality aspects, which does not happen always with administrative data.

If the NSI needs to report (for example to Eurostat) on the quality of the original administrative data it uses then it can simply present the source specific reports about these data. Product specific reports may also be presented if it is deemed necessary.

References

Borchsenius, L. (2000) From a conventional to a register-based census of population. *Presented at the INSEE-Eurostat seminar on censuses after 2001*.
Available at http://www.insee.fr/en/av_service/colloques/semie_textes.htm.

Carling, J., Wallgren, B. and Wallgren, A. (1996) The role of administrative registers in Sweden's statistical system. In *82nd DGINS conference proceedings.*

Carson, S., Laliberte, L., (2002), Assessing Accuracy and Reliability: A Note Based on Approaches Used in National Accounts and Balance of Payments Statistics, IMF Working Paper, WP/02/24, available at
http://www.imf.org/external/pubs/cat/longres.cfm?sk=15592.0

Dorinski, S. M. Continuing research on use of administrative data in SIPP longitudinal estimation. *US Census Bureau SIPP working paper no 209.*
Available at http://www.census.gov/dusd/MAB/sipp~1.html.

Eurostat (1995) Statistics on persons in Denmark, a register-based approach. Luxemburg.
The document was prepared by Statistics Denmark.

Eurostat (1996), How to measure quality of statistics based on administrative data or estimations, doc. Eurostat/D3/Quality/96/10 rev.1

Falorsi, P. D., Pallara, A., Succi, R., Russo, A. (2000) Estimating indicators of labour input from administrative records having coverage and measurement errors. *Presented at the 2nd International Conference on Establishment Surveys.*
Available at www.eia.doe.gov/ices2/missing_papers3.pdf.

Fellegi, I. P. and Sunter, A. B. (1969) A theory for record linkage. *J. Amer. Statist. Assoc.*, 64, 1183-1210.

Froeschl, K. A. and Grossmann, W. (1999) The role of metadata in use of administrative sources. In *ETK '99 proceedings* (ed. D. Murphy, P. Nanopoulos and D. Wilkinson), pp. 323-327.
Available at
http://europa.eu.int/comm/eurostat/research/index.htm?http://europa.eu.int/en/comm/eurostat/research/conferences/etk-99/&1.

Huynh, M., Rupp, K. and Sears, J. The assessment of SIPP benefit data using longitudinal administrative records. *US Census Bureau SIPP working paper no 238.*
Available at http://www.census.gov/dusd/MAB/sipp~1.html.

Long, J. (1996) Demographic applications of administrative records. *Presented at the 1996 Joint Statistical Meetings, Government Statistics Section of the American Statistical Association*.
Available at http://www.science.gmu.edu/gss/96gssprc.htm.

Marquis, K., Wetrogan, S. and Palacios, H. (1996) Towards a U.S. population database from administrative records. *US Census bureau working paper in survey methodology 96/06*.
Available at http://www.census.gov/srd/www/byyear.html.

Myers, A. L., Kinyon, D. L. and King C. S. (2001) Using administrative data in lieu of survey responses for small businesses. *Presented at the 2001 Federal Committee on Statistical Methodology conference.*
Available at http://www.fcsm.gov/events/papers2001.html.

Netterstrom, S. (2000) Metadata for statistics based on administrative data, *working paper 8: UN/ECE work session on statistical metadata*, Washington, DC, 28-30 November 2000.

Office for National Statistics (2001) National Statistics Quality Review of the Inter-Departmental Business Register.
Available as a group of documents at
http://www.statistics.gov.uk/nsbase/methods_quality/quality_review/commerce.asp.

Prevost, R. and Leggieri, C. (1999) Expansion of administrative record uses at the Census Bureau: a long-range research plan. *Presented at the 1999 Federal Committee on Statistical Methodology conference.*
Available at http://www.fcsm.gov/events/papers1999.html.

Scalia Jr, J. (1999) Using administrative records to report federal criminal case processing statistics. *Presented at the 1999 Federal Committee on Statistical Methodology conference.*
Available at http://www.fcsm.gov/events/papers1999.html.

Sweet, E. M. (1997) Using administrative record persons in the 1996 community census. *US Census bureau working paper in survey methodology 97/06*.
Available at http://www.census.gov/srd/www/byyear.html.

Vale, S., Perry, J. and Pont, M. (2001) Developing a quality strategy for business registers: a UK perspective. In *2001 NTTS and ETK conference proceedings* (ed. P. Nanopoulos and D. Wilkinson), pp. 415-423.

Wallgren, A. and Wallgren, B. (1999) How can we use multiple administrative sources? *Statistics Sweden technical report.*

Winkler, W. E. (1999) The state of record linkage and current research problems, *US Census Bureau research report 00/07*.
Available at http://www.census.gov/srd/www/byyear.html.

Yancey, W. E. (2000) Frequency-dependent probability measures for record linkage, *US Census Bureau research report 99/04*.
Available at http://www.census.gov/srd/www/byyear.html.