

A. Integrating survey and administrative sources

77. The subject of integrating administrative data is not new in the statistical world. But the degree and systems of integrating administrative and survey sources vary greatly across countries; some already have extensive experience integrating survey and administrative data resulting in register-based statistical systems, while others are just starting to integrate the data. In some countries administrative data may have existed for some time but not been used.

78. Practical experience shows that countries who have made the most of the data integration have a tradition in register-based statistics. Register based statistics refers to a system that is based on administrative data and in which statistical registers have been organized into a linked statistical system. A register is defined as a systematic collection of unit-level data organized in such a way that updating is possible. A requirement is that each unit in the register can always be uniquely identified.

79. A system of register-based information was developed first in Nordic countries, with other countries following. It is thoroughly explained in the publication [Register-based Statistics in the Nordic Countries - Review of Best Practices with Focus on Population and Social Statistics](#). Preconditions for such a system are a legal base, unified identification systems and cooperation among institutions while registers are the spines of the practical integration of data from multiple sources.

80. Some examples of survey and administrative data integration are: social surveys (e.g. LFS, EU SILC), various registers (e.g. employment, education registers) and there are several examples of administrative data being combined with survey data for producing indicators traditionally collected through censuses (e.g. agricultural censuses and register based population censuses).

81. The data may be integrated using record linking or statistical matching or may use modelling approaches. It may involve pooling or combining information from multiple surveys, including surveys not conducted by the statistical organisations themselves.

82. There are common challenges faced in the integration. The quality of administrative dataset may be good enough for administrative purposes but not sufficient for statistical purposes. Transforming administrative datasets into statistical datasets may require improving the quality and dealing with conceptual differences, especially when the statistical organisation wants to use administrative data in a direct way. In the case of surveys carried out with the use of data from administrative sources it is crucial to gather all data.

83. Administrative data are an already existing source and are already being collected for administrative purposes. Sample surveys require the design of sample frame and contact and response from the target population. Use of administrative data can be cost-effective and cheaper than collecting data by questionnaires. As they already exist there is no additional respondent burden on individuals and businesses. Sample surveys are generally more flexible than administrative sources as they are designed to meet a precise purpose. Administrative sources are on the other hand are the result of a legislative system and may have limitations concerning statistical purposes.

84. Administrative sources can provide full coverage of populations. The ability of administrative data to cover whole populations enables the production of local area data to a level of detail not permitted by sample surveys, which is also of advantage in implementing local policies. Administrative sources can also have the ability to produce more frequent statistics. It depends on the nature of data, but in some cases, such as administrative population registers, business registers, farmer registers and social security data, the sources can be updated daily.

85. In the statistical production process, administrative and survey data can be integrated in different ways. Considering usage of administrative data there is a [distinction between direct and indirect usage](#).

86. *Direct usage* is when administrative data supplement or replace the sample survey. When administrative data supplement sample surveys there are more possibilities. Often some variables will be based on questions in the sample survey and some variables will be supplemented from the administrative data. A good example of supplementing sample surveys for a set of variables are the LFS and EU SILC. Another way is that administrative data supplement the sample survey for a part of the population. When administrative data replace the sample survey the statistics are based entirely on administrative sources. Some examples are: employment, earnings, education, agricultural statistics, [register based population censuses](#), agricultural censuses.

87. *Indirect usage* of administrative data is when they are used for sampling frames, establishing and maintaining statistical registers, data editing and imputations, data validation and estimation (e.g. small area estimation) and weighting.

88. There are a number of challenges in integrating administrative and survey data. Since administrative data are collected for non-statistical purposes, the difference in concepts might lead to coverage problems as well as bias problems. In some cases, such as business statistics, units do not necessarily correspond directly to the definition of the required statistical units. This requires some modelling to convert the administrative units into statistical units. It is likely that there will also be differences in the definitions of variables. It is important to have a thorough understanding of the impact of these differences. Sometimes it is possible to influence the administrative definition by co-operating with the responsible authority. Administrative data are compliant with the laws in the country and in some instances, definitions and concepts are good enough to provide national statistics but cannot be used in the context of international comparisons to other countries, for example in the ESS system.

89. When administrative data are used to supplement surveys, or are integrated with other administrative data it is desirable that the two data sets contain overlapping information. The ideal situation is if data sets contain unique identifiers. If there are no unique identifiers, combinations of other available individual characteristics have to be considered instead, such as name, gender, address, and date and place of birth, to identify identical subjects in both data sets.

90. Another issue is classifications. Classifications used in administrative data may not be the same as classifications in statistical production. In cases of different classifications, the usual approach is to use correspondence tables and conversion tools based on additional variables that may be available for converting into more correct classification codes. However, even the same classifications may result in different data, especially when classifications are complex, or the rules of a classification are difficult to apply.

91. In administrative sources, coding is often done by the respondent, while a sample survey may have open questions and coding is often done by experts. Co-operation between the statistical organisation and the administrative authority is an effective way to solve a part of the classification problem. The statistical organisation can provide experience and may be the one responsible for maintaining the classification.

92. Another issue that concerns classification is a decision to use directly translated international classifications or national classifications. It depends on what national data are needed; however, the first option is usually harder to implement in case of changes and revisions compared to having national classifications. To change a classification in an administrative source is a demanding task since there can be many data providers that need to become familiar with the changes.

93. Missing data and errors also need to be considered. Missing data happen due to unit or variable non-response, but in administrative sources the causes can be different. It is important to identify if errors and missing data are systematic and apply appropriate validation and editing rules.

94. Timeliness is one more point in integrating administrative and survey data. Administrative data may not be available in time or may not coincide with the statistical reference period. Sometimes it can be resolved by analysing the impact and if necessary adjusting it using models.

95. Many international statistical offices and statistical organisations have guidelines, directives, standards, recommendations concerning administrative data. Following is an example of guidelines for dealing with administrative data that can be found on [Statistics Canada webpages](#) (summarized to some extent):

- Maintain a continuing liaison with the provider of administrative records.
- Understand the context under which the administrative organization created the administrative program (e.g. legislation, objectives, and needs).
- Keep in mind that if the information provided to the administrative source can cause gains or losses to individuals or businesses, there may be biases in the information supplied which can lead to unexpected coverage problems and biases.
- Collaborate with the designers of new or redesigned administrative systems.
- Develop an imputation or a weight-adjustment procedure to deal with this nonresponse (unless non-respondents can be followed up and responses obtained). Administrative sources are sometimes outdated. Therefore, as part of the imputation process, give special attention to the identification of active and/or inactive units.
- In the case when a common matching key for both sources is not available and record linkage techniques are used, select the type of linkage methodology (e.g. exact matching or statistical matching) in accordance with the objectives of the statistical program. When the purpose is frame creation and maintenance, or data editing, exact matching should be used. In the case of imputation or weighting, exact matching should be used, but statistical matching can be also sufficient. When the sources are linked for performing some data analyses that are impossible otherwise, consider statistical matching, e.g. matching of records with similar statistical properties.
- When record linkage is to be performed, make appropriate use of existing software.
- When data from more than one administrative source are combined, pay additional attention to reconcile potential differences in their concepts, definitions, reference dates, coverage, and the data quality standards applied at each data source.
- Some administrative data are longitudinal in nature (e.g. income tax, goods and services tax). When records from different reference periods are linked, they are very rich data mines for researchers. Remain especially vigilant when creating such longitudinal and person-oriented databases, as their use raises very serious privacy concerns.
- Use identifiers with care, as a unit may change identifiers over time. Track down such changes to ensure proper temporal data analysis. In some instances, the same unit may have two or more identifiers for the same reference period, thus introducing duplication in the administrative file. If this occurs, develop a mechanism to remove duplicates.
- Document the nature and quality of the administrative data once assessed. Documentation helps statisticians decide the uses for which the administrative data are best suited. Choose appropriate methodologies based on administrative data and inform users of the methodology and data quality.