5.1 Integrate data

This sub-process integrates data from one or more sources. It is where the results of sub-processes in the "Collect" phase are combined. The input data can be from a mixture of external or internal sources, and a variety of the collection instruments, including extracts of administrative and other non-statistical data sources. Administrative data or other non-statistical sources of data can substitute for all or some of the variables directly collected from survey. This sub-process also includes harmonising or creating new figures that agree between sources of data. The result is a set of linked data. Data integration can include:

- Combining data from multiple sources, as part of the creation of integrated statistics such as national accounts;
- Combining presentational data and statistical data or other non-statistical data;
- Data pooling, with the aim of increasing the effective number of observations of some phenomena;
- Matching or record linkage routines, with the aim of linking micro or macro data from different sources;
- Data fusion - integration followed by reduction or replacement;
- Prioritising, when two or more sources contain data for the same variable, with potentially different values.

Data integration may take place at any point in this phase, before or after any of the other sub-processes. There may also be several instances of data integration in any statistical/business process. Following integration, depending on data protection requirements, data may be (de-)identified, that is stripped of identifiers such as name and address, to (try to) protect confidentiality.

Quality Indicators:
- The proportion of units covered by both the survey and the administrative sources in relation to the total number of units in the survey. The proportion is applicable:
  - to mixed statistical processes where some variables or data for some units come:
    - from survey data and others from administrative sources;
    - to production;
    - to existence and linkage variables (unique identifier) of the register (yes/no question) linking of microdata to other microdata;
    - Degree of initality of the linkage variables (high, medium, low);
    - Proportion of duplicated records in the linked data;
    - Reliability of the linkage results;
    - False link and base nominal rates;
    - Precision and recall;
    - An indicator of the effectiveness of the cut-off weight for determining the threshold of passage in probabilistic matching;
    - Percentage of errors coming from identification and transformation of population, units or data items;
    - Rates of unit change from period to period;
    - Proportion of units that may belong to more than one composite unit. This indicator records how often a base unit (e.g. a person) doesn't have a single clear composite unit to which it can be assigned without doubt. This could be units that can't be assigned to any composite unit for some reason, or units equally likely to belong to two different composite units;
    - If record linkage is required, report linkage methodology (e.g. exact, probabilistic, etc.) used;
    - Linkage rate - Proportion of units linked from each dataset to a base dataset, or percentage link rates between parts of datasets;
    - Proportion of manually linked units;
    - Macro-level Comparisons of the distribution of linked objects with reference distributions;
    - Distribution of variables in linked data.

For details, see Quality Indicators for OSBFM (version 2.0).