**UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE** Working paper
**CONFERENCE OF EUROPEAN STATISTICIANS** 27 July 2016

Workshop on Implementing Standards for Statistical Modernisation,
21 – 23 September 2016

# Record Linkage Project Process Model
# Statistics Canada
# (Draft for Consultation)

Prepared by Working Group on Record Linkage
Analysis Coordinating Committee
Statistics Canada
Contacts: Claudia Sanmartin (HAD) (co-Chair)
Richard Trudeau (SDLE) (co-Chair)

## Preamble

Over the past decade, the volume of record linkages has increased significantly at Statistics Canada serving both analytical and operational needs. Linkage projects often involve various groups within the Agency including data processors, data linkers, and subject-matter specialists. Currently, a range of terms, processes, methods and approaches are used to conduct record linkage projects, in some case, with few standards to guide the process.

In April 2015, the Working Group on Record Linkage (WG on RL) was created by the Analytical Coordinating Committee with the following objectives:

- To achieve a common understanding of the concepts and processes involved in record linkage conducted at Statistics Canada;
- Identify challenges and opportunities to improve record linkage activities;
- Identify and/or develop standard approaches and recommended practices to conduct, validate and document record linkage projects to ensure alignment with existing policies and directives where relevant.

To meet the first objective, the WG set out to map the record linkage project process to reflect the general practices and activities involved in record linkage at Statistics Canada. The Agency represents a complex record linkage environment with diverse social and economic data sets linked at the individual and enterprise level using a range of methods to meet various objectives including the development of new linked data sets for research, data replacement for ongoing surveys and the development of registries for operational purposes. Recognizing these complexities, the goal of the WG was to map a "generic" process that reflects the common practices across the Agency and processes we aspire to. The model was also developed with a view to a more general use by other national statistical organizations conducting record linkage.

This report describes the *Record Linkage Project Process Model (RLPPM).* The Model builds on the *Generic Statistic Business Process Model v5.0* developed by the Joint UNECE / Eurostat / OECD Work Session on Statistical Metadata (METIS) for survey processes. (see Appendix A) It also builds

on international models of record linkage from Australia and the United States as well record linkage methodology where they meet practices at Statistics Canada. Finally, the model was informed by the relevant legal and policy frameworks which govern all of Statistics Canada statistical activities. (see Appendix C). Note that statistical matching, another approach to providing joint statistical information based on variables collected through two or more sources, is not covered by this model. A Glossary of Terms is also provided to assist readers in the interpretation of the process model. (see Appendix B)

## Record Linkage Project Process Model

| Phase 1: Project Planning | | | Phase 2: Record Linkage | | | Phase 3: Post linkage activities | | |
|---|---|---|---|---|---|---|---|---|
| **1 Specify needs** | **2 Design** | **3 Approve** | **4 Prepare data** | **5 Link data** | **6 Assess quality** | **7 Integrate and Analyzes** | **8 Access and Disseminate** | **9 Evaluate** |
| 1.1 Identify needs | 2.1 Design linkage strategy | 3.1 Consult and confirm approval process | 4.1 Standardize linkage variables | 5.1 Indexing (Blocking) | 6.1 Internal validation | 7.1 Integrate data, review and validate | 8.1 Establish access process | 9.1 Gather evaluation inputs |
| 1.2 Consult and confirm needs | 2.2 Design quality assessment strategy | 3.2 Prepare approval documents | 4.2 Assess linkage variables | 5.2 Field and record comparison | 6.2 External validation | 7.2 Apply quality adjustments | 8.2 Establish disclosure control protocols | 9.2 Conduct evaluation |
| 1.3 Check data availability | 2.3 Plan for quality adjustments | 3.3 Submit for approval | 4.3 Identify in scope records for linkage | 5.3 Linkage rules | 6.3 Adjust record linkage strategy | 7.3 Derive variables | 8.3 Store and manage access | 9.3 Agree on action plans |
| 1.4 Determine feasibility of record linkage | 2.4 Identify access needs | 3.4 Archive approval | 4.4 Evaluate results of data preparation | 5.4 Finalize record linkage strategy | 6.4 Produce linkage keys | 7.4 Finalize linked data set and document | 8.4 Destruction of files | 9.4 Add to linkage tool box |
| 1.5 Identify sponsor and custodian | 2.5 Estimate cost | | 4.5 Initiate record linkage report | 5.5 Document record linkage strategy | 6.5 Finalize record linkage report | 7.5 Analyze, validate and feedback | | |
| OUTCOMES/OUTPUTS | | | | | | | | |
| Decision to proceed with record linkage project | Project plan and budget | Approved record linkage project | Linkage ready data sets | Preliminary linkage keys | Final linkage keys; Record linkage report | Linked set; Documentation; Analytical products | Disclosure and Access protocols | Evaluation Report; Tool box |

## Description of Phases and Sub-processes

### Phase 1: Project Planning

The first meta-phase of the record linkage project process focuses on identifying the specific statistical need or data gap triggering the record linkage project. A high level assessment is also conducted to determine if record linkage is a feasible option to meet the need and fill the data gap. If deemed feasible, an initial record linkage strategy is designed and approval to proceed is sought. This phase involves consultation with a range of actors including data custodians, record linkage specialists, subject matter specialists and agents responsible to ensure compliance with applicable legislation, policies and directives.

### 1. Specify needs

This phase is triggered by a need for new data to meet statistical requirements including a new data set to support research, data replacement for surveys or the development of a new registry for operational use. It includes all activities associated with engaging stakeholders or clients to identify their detailed data and information needs. This phase also includes gathering information to determine if record linkage is a viable option to meet those needs.

This phase is comprised of five sub-processes. These processes are generally conducted sequentially but can also occur in parallel at some stages. The sub-processes are:

| 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
|---|---|---|---|---|
| Identify needs | Consult and confirm needs | Check data availability | Determine feasibility of record linkage | Identify sponsor and custodian |

#### 1.1 Identify needs

This sub-process includes the initial investigation and identification of a statistical need or data gap. It may be triggered by a need for new data to address an emerging research question, an information request or an environmental change such as a reduced budget including the need to use record linkage to replace survey content. It also includes consideration of common practices among other (national and international) organizations producing similar data, and, in particular, the methods used by those organizations. The needs may be identified internally or by external stakeholders or clients.

#### 1.2 Consult and confirm needs

This sub-process focuses on consulting with the stakeholders or clients to confirm in detail their data or information needs. This may require identifying the specific objectives of the research or statistical output to better identify the concepts of interest and understand the data gaps. This discussion may require engagement of subject matter experts who have an in-depth understanding of both the concepts of interest and

limitations of existing data. Alternatively, this may require engagement of survey managers who are considering the use of record linkage to better use existing administrative data to replace survey content in an effort to reduce respondent burden.

A good understanding of user needs is required so that the statistical organization knows not only what it is expected to deliver, but also when, how, and, perhaps most importantly, why. The detailed understanding of the stakeholder or client needs is the critical part of this sub-process.

### 1.3 Check data availability

This sub-process checks whether existing data sets could meet the stakeholder or client requirements, and the conditions under which they would be available, including any restrictions on their use. An assessment of possible alternatives would include a review of existing data sets including surveys, administrative data bases, linked data or other non-statistical data sets, to determine whether they would be suitable to meet the stakeholder or client need.

If an existing data set or linked data set is identified and deemed appropriate to fill the data gap, then the conditions under which the data can be accessed by the stakeholder or client should be explored. If existing data sets cannot fill the data gap, the project proceeds to consider record linkage as a viable option.

### 1.4 Determine feasibility of record linkage

This sub-process focuses on determining if record linkage is a feasible option beginning with identification of the data sets that could be linked (i.e. source data sets). The linkage could involve source data sets "owned" or held by the statistical agency or linkage organization or may require access to a data set owned by an external stakeholder or client. Existing data acquisition agreements stipulating access to and use of the data should be reviewed to ensure the source data sets are available for linkage activities.

This sub-process also includes a review of the availability and quality of identifying variables common across all source data sets to facilitate the record linkage process (i.e. linkage variables). Variables commonly used for record linkage include names (e.g. personal names and business names), demographic variables (e.g. sex, date of birth), geographic variables (e.g. address and postal codes), telephone numbers and unique identifiers (e.g. social or health insurance numbers, business numbers). Access to and use of the linkage variables should be explored in the context of the regulatory and policy framework governing the source data sets. This process also involves the exploration of linking the source data sets directly or via a central population registry.

An initial assessment of the discriminatory power of the available linkage variables will also inform project feasibility. Unique identifiers have high discriminatory power while variables such as gender, date of birth and geographic identifiers are less discriminating. In some cases, access to the source data sets will be required to assess the quality and discriminatory power of the linkage variables – a likely requirement when linking new data sets.

This sub-process may require consultation with various groups. Data custodians of the source data sets, both internal and external to the statistical agency or linkage

organization, should be consulted to determine availability of the data for linkage, develop a data acquisition agreement if needed, determine the quality of linkage variables and possible limitations on the use of the data. Record linkage specialists should be consulted on technical issues to determine the feasibility of the linkage. Subject matter specialists (i.e. analysts) should be consulted to determine the appropriateness of linked data set for use in analysis and research. Finally, agents responsible to ensure compliance with applicable legislation, policies and directives should be consulted to ensure that the linkage project complies with established agreements and policies. In the case of linkage to surveys, notifications to respondents regarding data linkage should be in place and

**1.5 Identify sponsor and custodian**

This sub-process involves identifying the agents of the record linkage project namely, the sponsor and the custodian of the linked data set. The *project sponsor* will champion the record linkage project and assume related tasks and responsibilities including client relations, project management, obtaining approval and liaising with relevant groups (e.g. methodology, subject matter) on behalf of the stakeholder or client. The *custodian* will assume responsibility for the linked data set(s) produced upon completion of the record linkage project including storage, managing access, responding to questions about the file, documentation and ultimately destruction of the data set as per existing policies and directives.

**Outcome: Decision to proceed with record linkage project**

## 2. Design

This phase focuses on the development of a design strategy for all the components of the record linkage project including the linkage strategy, quality assessment and access requirements. This stage also includes establishment of a preliminary budget for the project.

This phase may require consultation with record linkage and subject matter specialists to ensure that the design meets quality standards and analytical needs.

| 2.1 | 2.2 | 2.3 | 2.4 | 2.5 |
|---|---|---|---|---|
| Design linkage strategy | Design quality assessment strategy | Plan for quality adjustments | Identify access needs | Estimate cost |

**2.1 Design linkage strategy**

This sub-process involves designing the record linkage strategy beginning with the identification of the most appropriate record linkage method to use (e.g. deterministic, hierarchical deterministic, or probabilistic). Considerations include the project objectives, the ultimate use of the linked data, the input data set(s), availability and quality of linkage variables and available resources for the record linkage project. This process should include a review of methods and processes used in previous record linkage projects involving similar source data sets to determine whether existing linkage keys and/or strategies may be relevant. For recurring record linkages, such as data

replacement in on-going surveys, the need to maintain consistent methods over time should be considered.

### 2.2 Design quality assessment strategy

This sub-process involves designing the quality assessment strategy to ensure that the linked data set is fit its intended use. The quality assessment strategy focuses on the linked data sets and assumes that the quality of the source data sets has been previously established. At this stage, measures of both internal and external validity are considered.

Measures of *internal validity* focus on the accuracy of the record linkage process, on identifying potential sources of error (e.g. rates of false positive and false negative links) and on bias. Measures of *external validity* focus on data confrontation comparing key outcomes generated from the linked data set to external data sets (e.g. mortality rates, employment rates). The quality assessment strategy should consider the availability of external sources of information and resources for the project.

### 2.3 Plan for quality adjustments

This sub-process involves planning for adjustments to improve the overall quality of the linked data set. This could include, for example, the need for adjustment to account for error in the input files and /or potential bias that is introduced as a result of the linkage strategy or resulting linkage error. Other quality adjustments could include imputation to address missing information as a result of missed links.

The resource requirements for quality adjustments should be considered and balanced against the level of precision required given the intended use of the linked data set. Final determination of the need for quality adjustments will occur following the record linkage and quality assessment of the linked data set.

### 2.4 Identify data access needs

This sub-process involves an assessment of the need for access to the linked data set and consideration of the process and requirements for various dissemination scenarios. Access to the linked data set may be restricted to the stakeholder or client requesting the record linkage or restricted to those internal to the organization if being used for operational purposes only (e.g. registers). Alternatively, subsequent uses of the linked data set may be identified requiring broader access to the linked data set. Documentation and metadata requirements as well as the potential development of disclosure protocols should also be considered. Additional resource requirement incurred as a result of the dissemination plan should be noted and planned for.

### 2.5 Estimate cost

This sub-process involves a more precise estimation of the project costs given the design decisions taken to date. Cost estimates should include the full range of activities including pre-processing of the source data sets, record linkage, quality assessment, creation of the linked data set, documentation and administration costs.

**Output: Project plan and budget**

## 3. Approve

This phase focuses on obtaining approval to proceed with the record linkage project in accordance with established policies and legal frameworks governing record linkage activities at the statistical organization. Record linkage should not be conducted without prior approval.

The sub-processes in this phase are as follows:

| 3.1 | 3.2 | 3.3 | 3.4 |
|---|---|---|---|
| Consult and confirm approval process | Prepare approval documents | Submit for approval | Archive approval |

### 3.1 Consult and confirm approval process

This sub-process begins with consulting with the relevant governing body responsible for approving record linkage proposals. This sub-process also involves consulting with relevant data custodians to provide details of, and obtain support for, the proposed record linkage project. This provides an opportunity for data custodians to raise any issues or challenges regarding the use of their data for the proposed project.

### 3.2 Prepare approval documents

This sub-process involves preparation of documents by the project sponsor to obtain approval for the record linkage project. The request for approval to link the data sets should identify the following: objectives of the linkage project, public good served by the record linkage, justification of the privacy intrusion inherent to the record linkage, listing of the source data sets, structure of the final linked data set, provisions to ensure the privacy and confidentiality of the data and dates for destruction of the linked data set.

### 3.3 Submit for approval

This sub-process involves a formal submission of the request for approval for the record linkage project to the governing body.

### 3.4 Archive Approval

Once approved, the requests for approval for record linkage are archived. Destruction dates for the linked data set are noted.

**Output: Approved record linkage project**

## Phase 2: Record Linkage

The second meta-phase of the record linkage process focuses on implementation of the record linkage strategy in the following three phases: data preparation, data linkage, and

quality assessment. This phase concludes with the production of a record linkage report which documents the decisions and activities during these phases.

## 4. Prepare Data

The data preparation phase focuses on preparing the linkage variables from the source data sets. The phase begins with sub-setting the linkage variables from the source data sets (i.e. source index data sets) for privacy protection. The linkage variables must be processed and analyzed to determine the level of accuracy, completeness and to increase their comparability and ultimately to ensure they are of sufficient quality to support the record linkage. The availability and quality of the linkage variables will inform the specifics of the record linkage strategy.

This phase is comprised of the following sub-processes:

| 4.1 | 4.2 | 4.3 | 4.4 | 4.5 |
|---|---|---|---|---|
| Standardize linkage variables | Assess linkage variables | Identify in-scope records for linkage | Evaluate results of data preparation | Initiate record linkage report |

### 4.1 Standardize linkage variables

This sub-process involves sub-setting the source data sets to create the source index data sets that only include the linkage variables. The sub-setting of these variables mitigates the privacy intrusiveness of the process as the record linkage specialists only require access to the linkage variables at this time. Variables commonly used for record linkage include names, demographic variables (e.g. sex, date of birth), geographic variables (e.g. postal codes) and unique identifiers (e.g. social or health insurance numbers, business numbers). Values contained with the linkage variables may be combined or concatenated to create linkage keys.

This sub-process involves standardizing the structure, format and code sets of the linkage variables across all source data sets to ensure comparability and stability across files.

Data custodians should be consulted to obtain relevant information regarding the collection, cleaning and standardization of the linkage variables and identify any potential issues or challenges that may impact the record linkage process.

### 4.2 Assess linkage variables

This sub-process involves assessing the quality and discriminatory power of the linkage variables expected to impact both the quality of the linkage and ultimately the use of the linked data.

As with all data variables, linkage variables are subject to errors whether in survey or administrative data. Linkage variables should be assessed for accuracy, reliability and completeness (e.g. rates of missing, incorrect or invalid values).

Methods to assess the discriminatory power of a linkage variable may vary depending on the variable type. For unique identifiers (e.g. social insurance number, business number, addresses), the process may involve confirming that the identifier is associated with a

single person or enterprise and is consistent over time. For a non-unique identifiers (e.g. name, date of birth), the process may involve using statistical approaches to measure discriminatory power (e.g. Shannon entropy).

Data custodians should be consulted to obtain relevant information regarding the quality of the linkage variables. The result of this sub-process will inform the specifics of the linkage strategy (e.g. linkage rules and weights in probabilistic linkage).

#### 4.3 Identify in-scope records for linkage

This sub-process involves the establishment of inclusion and exclusion criteria to identify records from the source index data sets that are *eligible* for record linkage. This process may be informed by the assessment of the linkage variables; for example, records with incomplete or missing values may be considered ineligible for linkage. Records of respondents who have refused their information be linked need to be excluded from linkage. Alternatively, this may be informed by the requirements of the record linkage project. The result of this sub-process is a final count of eligible and ineligible records for linkage from each source data set.

#### 4.4 Evaluate results of data preparation

In this sub-process, the results of data preparation are evaluated to determine readiness to proceed with the record linkage project. Results of the evaluation may require refinements or changes to the linkage strategy as well as inform the quality assessment process (i.e. identify potential source of bias) and use of the linked data set.

#### 4.5 Initiate record linkage report

In this sub-process, the record linkage report is initiated with the documentation of the data preparation phase. Specifically, the following information should be included for all record linkage projects: a brief description of the source data sets; results of the evaluation of the linkage variables and eligibility criteria with final record counts. Counts of eligible and ineligible records for linkage will serve to calculate linkage rates later in the process, provide evidence of potential bias and inform quality adjustment strategies (e.g. weighting) if required.

Results of the data preparation stage should be saved and available for future record linkage projects involving the source data sets. Information could also be shared with the data custodian to improve the overall quality of source data.

**Output: Linkage ready data sets**

## 5. Link Data

In this phase, the source index data sets are linked. There are several approaches to record linkage: deterministic, hierarchical deterministic and probabilistic. In general, the probabilistic process involves more techniques than the deterministic process. The following describes the general sub-processes relevant for all record linkage regardless of the specific methodology used.

These sub-processes are generally conducted sequentially but can be iterative as the linkage is evaluated at each step and may require refinements of previous processes. All linkage methods incorporate the following five components:

| 5.1 | 5.2 | 5.3 | 5.4 | 5.5 |
|---|---|---|---|---|
| Indexing (or Blocking) | Field and record comparison | Linkage rules | Finalize linkage strategy | Document record linkage strategy |

**5.1 Indexing (or Blocking)**

When dealing with large source data sets, it may not be practical to assess all possible pairs of records. As such, indexing or blocking is used to reduce the number of possible pairs generated by the cross-product of the tables being linked to a smaller subset of pairs that can be processed in a reasonable amount of time and with available resources. This subset is comprised of pairs that agree perfectly on a specific key or criterion and represent potentially matched pairs, which are also called potential pairs. Indexing reduces the required computational resources. However, it can potentially increase the rate of missed links.

In a probabilistic linkage, indexing is referred to as blocking in which blocking criteria are used to reduce the number of pairs considered further for linkage. Multiple blocking criteria are often used to minimize the number of missed links. In a hierarchical deterministic linkage, indexing is implicit and used when merging large data sets.

**5.2 Field and record comparison**

For each potential pair, the attributes of the paired records are compared. This typically involves the comparison of strings (e.g. names) and/or numerical combinations (e.g. date of birth).

The comparison of attributes is based on comparison functions or linkage rules. Elaborate functions may simultaneously compare multiple attributes at multiple levels and return more complex outcomes (e.g. matrix comparisons). Comparison functions also differ according to their treatment of missing values.

In deterministic or hierarchical deterministic linkage, comparison functions are embedded in the logical conditions that identify matched pairs. In deterministic linkage for example, only exact matches on a single function are accepted as links.

**5.3 Linkage rules**

Comparison outcomes are used to make a linkage decision to determine whether the records are matched, unmatched or possibly matched and therefore subject to further manual review.

In a deterministic or hierarchical deterministic linkage, this decision is based on a sequence of logical conditions following a rule-based approach. These logical conditions are developed iteratively using manual reviews on samples of pairs.

In probabilistic linkage, a weight is assigned to reflect the similarity of each linkage variables in a record pair with higher weights assigned when there is a higher level of agreement on variables with higher discriminatory power. A total linkage weight is then calculated for each pair. The total weight is compared to two thresholds to make a decision as to whether the pair represents a match or non-match. Pairs with a weight between the two thresholds are resolved manually. No manual resolution takes place when the two thresholds are identical. The thresholds are theoretically set according to the targeted rates of linkage error, including the false match rate and missed match rate. The linkage parameters include the linkage weights and the thresholds. They must be estimated from the set of potential pairs. This estimation may be based on manual reviews or a statistical model. The most common statistical models incorporate assumptions.

### 5.4 Finalize record linkage strategy

This sub-process recognizes that record linkage is an iterative process. The evaluation of the initial record linkage strategy can lead to adjustments before the optimal strategy is obtained. Evaluation can include clerical review of selected links and initial assessment of the overall linkage results (i.e. linkage rates). The linkage strategy may once again be refined following the validation process described in the phase 6.

### 5.5 Document record linkage strategy

In this sub-process, the methods used and decisions taken in the record linkage process are documented in the record linkage report. The goal is to document to a level of detail that would facilitate the replication of the linkage strategy. For deterministic linkage, information on the logical conditions used to identify matched pairs as well as any manual review done when developing the rules should be provided. For probabilistic linkage, the following should be documented: the blocking criteria and any subsequent modification of the potential pairs after the initial creation; the comparison functions including the treatment of missing values; and the estimation of linkage parameters including the use of any statistical model or manual review.

**Output: Preliminary linkage keys**

## 6. Assess quality

The third phase of the record linkage process focuses on assessing the quality of the linkage (*internal validation*) and the "fitness for use" of the linked data set (*external validation*). The goal of this phase is to ensure that the linked data set are of a level of quality appropriate for their intended use. The quality assessment should be conducted in accordance with the plan developed in sub-process 2.2. Any limitations of the linked data identified in this sub-process should inform the use of the linked data.

The quality assessment phase is comprised of the following components:

| 6.1 | 6.2 | 6.3 | 6.4 | 6.5 |
|---|---|---|---|---|
| Internal validation | External validation | Adjust record linkage strategy | Produce linkage keys | Finalize record linkage report |

**6.1 Internal validation**

This sub-process focuses on internal validation to assess the quality of the linkage strategy. The validation at this stage is restricted to using the linking variables. This process can begin by assessing the face validity of the linkage results including: comparison of the overall linkage rates to expected levels informed by experiences with previous linkage projects and/or external sources; analysis of linkage rates across sub-groups or populations to identify potential bias and/or confirm expected patterns where not all records are expected to link (e.g. linkage to mortality records). Analysis of unlinked records may also be conducted at this stage. Subject matter specialists may be consulted at this stage.

This sub-process should also include a more rigorous assessment of the accuracy of the linkage strategy to confirm "good" links and estimate the rate of "bad" links or errors. Post-linkage approaches such as clerical assessment, comparison to a "gold standard" or simulation can be used to generate error estimates for the linked data set (e.g. false positive and false negative rates).

**6.2 External validation**

This sub-process is the first step in assessing the "fitness for use" of the linked data set. The methods used at this stage should be aligned with the intended use of the linked data set. For linked data developed to support analysis and research, additional variables from the source data sets may be required. This step is conducted following internal validation (sub-process 6.1) in which the linkage has been demonstrated to be of acceptable quality. External validation could include data confrontation in which statistical outputs generated using the linked data set are compared with external data sets. This process should involve the input from subject matter specialists where appropriate. External validation can also occur following data integration in which the final linked data set is created and analyzed (Phase 3, 7.5).

**6.3 Adjust record linkage strategy**

This sub-process involves refinements to the record linkage strategy based on the findings of the internal and external validation if required. This may involve further refinement of the linkage rules or thresholds used to identify pairs as matches or non-matches. If refinements are required, sub-processes 6.1 and 6.2 should be repeated to evaluate the new linkage strategy. Further refinements may be required following the final stage of validation once the final linked data set is created and used for analysis (sub-process 7.5).

**6.4 Produce linkage keys**

This sub-process is the final output of the record linkage process in which the linkage keys are created and stored in a linkage key file. The linkage key file is an anonymous

file that contains the variables with an identifier component role necessary to identify the records related to the same entity in the source data sets but does not contain any identifying information that may have been used to create the links. Access to the linkage keys file is restricted to those individuals who have approval to create linked data sets.

#### 6.5 Finalize record linkage report

This sub-process involves documenting the results of the internal and external validation to complete the record linkage report. Any limitations of the data that may affect its use should be noted. The record linkage report should be in a format such that users of the linked data can understand the basics of the linkage strategy and the results of the quality assessment. It should also contain a level of detail that would allow the record linkage project to be repeated using the same methodology.

**Output: Final linkage keys; Record linkage report**

### Phase 3: Post Linkage Activities

In this final meta-phase, activities focus on using the results of the record linkage (i.e. linkage keys) to integrate the source data sets and create a linked data set ready for use. Protocols governing access to and use of the linked data are specified. Finally, an evaluation of the record linkage project is conducted.

### 7. Integrate and Analyze

This phase focuses on using the linkage keys to integrate data from the source data sets to produce a linked data set ready for use. Use of the data can include analysis for further validation, the production of statistical outputs or operational purposes.

This phase includes the following sub-processes:

| 7.1 | 7.2 | 7.3 | 7.4 | 7.5 |
|---|---|---|---|---|
| Integrate data, review and validate | Apply quality adjustments | Derive new variables | Finalize linked data set and document | Analyze, validate and feedback |

#### 7.1 Integrate data, review and validate

This sub-process integrates data from source data sets using the linkage keys resulting in a linked data set. Data integration typically includes merging routines, using linkage keys to create a linked data set and reconciling variables when two or more source data sets contain the same variables. The linked data set created for the purposes of analysis or as data replacement for a survey should not contain personal identifiers. Linked data sets used for internal purposes such as the establishment of frames may retain identifying information.

To review the merging process, frequency or record checks should be carried out on each individual source data set prior to the merging process and once again following the merge. Record counts are then compared to identify any discrepancies.

Further validation of the linked data set can include a more comprehensive assessment to identify potential errors and discrepancies such as duplicate records, outliers, item non-response and miscoding. This process may be run iteratively, validating data against predefined edit rules, usually in a set order. It may apply automatic edits, or raise alerts for manual inspection and correction of the data. Reviewing, validating and editing can apply to records both from surveys and administrative sources, before and after integration. This process is particularly critical in cases where one or more of the source data sets are new to the statistical or linking organization and evidence of data quality is lacking.

### 7.2 Apply quality adjustments

This sub-process involves applying the post-linkage quality adjustments to the linked data set. The adjustments should be implemented according to the plan elaborated in sub-process 2.2. Some additional quality adjustments may also be required as a result of the quality assessment conducted in sub-process 2.6.

Quality adjustments could include, for example, weighting to adjust for potential bias that is introduced by linkage errors or to account for design weights when linking a survey data set. This could be adjustment of survey weights for surveys used in the linkage process or the creation of new weights for administrative data.

Other quality adjustments could include imputation to address missing information as a result of missed links. This may be required in cases where record linkage is conducted for the purpose of data replacement in on-going surveys.

### 7.3 Derive new variables

This sub-process derives new variables that are not explicitly provided in the original source data sets, but are needed to deliver the required statistical outputs. It derives new variables by applying arithmetic formulae or different model assumptions to one or more of the variables that are already present in the linked data set. This may need to be iterative, as some derived variables may themselves be based on other derived variables. It is therefore important to ensure that variables are derived in the correct order. New variables may be derived by aggregating or splitting data from linked records. Derived variables in the context of linked data sets may involve information from more than one input file.

### 7.4 Finalize linked data set and document

In this sub-process, the linked data set is finalized and documented including necessary metadata and information to inform users on the use of the linked data set. The level of documentation should be appropriate to the intended use of the linked data set. This documentation can be a compilation of documentation related to the source data sets, the record linkage process, the results of the quality assessment and a record layout of the new linked data set if required. If these documents exist (e.g., user guides, data dictionary, and record layout) and are of the quality required for the intended use of the linked data set, they can be referred to in the linked data set documentation. Alternatively, if the source data set documentation is not available or incomplete for the intended purpose, more comprehensive documentation may need to be developed for the

linked data set  New or modified variables (e.g. weights and derived variables) resulting from sub-processes 7.2 and 7.3 must be fully documented.

**7.5 Analyze, validate and feedback**

In this sub-process, the first analysis of the linked data set is conducted.  This process pertains primarily to record linkage projects conducted to support the creation of new statistical outputs including indicators and research findings. Before analyzing the data, users should familiarize themselves with all documentation related to the linked data set including data dictionaries of the source data sets and the record linkage report to be aware of any limitations related to the original source data sets and/or resulting from the record linkage process.

Analysis of the linked data set is conducted to ensure that the data can be reliably used to produce the intended statistical output (e.g. new indicators, research results, new frame). This is often accomplished in the context of the original objectives of the analysis or research identified in Phase 1 for record linkage projects conducted to support the creation of new statistical outputs. This may include a further assessment of the linkage rates among the study cohort (e.g. sub-populations, industry specific enterprises). This is also the first opportunity to test associations and/or causal pathways across variables from the various source data sets to determine if they exist in the expected direction.

Results of the analysis or research derived from the linked data set should be compared to external sources to provide further validation. Analysts validate the quality of the statistical outputs produced from the linked data set in accordance with a general quality framework and with expectations. This sub-process also includes activities involved with the gathering of intelligence, with the cumulative effect of building up a body of knowledge about a specific statistical domain. This knowledge is then applied to the linked data set, in the current environment, to identify any divergence from expectations and to allow informed analyses.

Validation activities can include confronting the statistical output against other relevant data (both internal and external); confronting the statistics against expectations and domain intelligence; or investigating inconsistencies in the statistics.

This sub-process is important for providing additional information regarding the fitness-for-use of the linked data set. Any discrepancies or limitations identified regarding the use of the linked data set should be noted in the documentation to guide future users. Serious errors or limitations identified at this stage may require further refinements of the record linkage strategy, quality assessment and/or quality adjustments necessitating redoing some sub-processes in phases 5, 6 and 7.

**Outputs: Linked data set; Documentation; Analytical product**

## 8. Access and Disseminate

This phase focuses on establishing the provisional agreements under which the linked data will be accessed and/or disseminated to ensure compliance with existing regulatory and legal frameworks governing the data and record linkage process within the statistical or linking organization.

This phase is comprised of the following activities:

| 8.1 | 8.2 | 8.3 | 8.4 |
|---|---|---|---|
| Establish access process | Establish disclosure control protocols | Store and manage access | Destruction of the linked data sets |

**8.1 Establish access process**

This sub-process establishes the process under which the linked data will be accessed. This begins with a review of the access requirements identified in the planning stages (sub-process 2.4) and revised as required. Access may be restricted to the stakeholder or client whose data requirements initiated the record linkage project; alternatively provisions can be made to permit access to the linked data set to other users with similar data needs. Linked data sets created to support operational needs (e.g. registers) may be restricted to use by internal employees of the statistical organization.

The process should include consultations with the custodian of the linked data set who will ultimately be responsible for providing consent to access the data. Processes established to provide access to the linked data set must be compliant with the legal and policy frameworks that govern the source data as well as the linked data. A mechanism should also be established to track users of the data.

**8.2 Establish disclosure control protocols**

In this sub-process, protocols for the disclosure of statistical outputs generated from the linked data sets are established to ensure confidentiality is maintained. These could include the requirement for minimum cell counts or maximum proportional contributions in tabular and model outputs, restrictions on levels of geography or on groupings of categorical variables, or treatment of statistical outputs (e.g. random rounding). These are requirements in cases where the linked data set will be used for analysis and the development of statistical outputs.

**8.3 Store and manage access**

In this sub-process, the linkage keys are stored in a secure manner. In cases where the linked data will be used for analysis, the identifiers are removed; in cases where the linked data are used for operational reasons, identifiers will remain as part of the linked data. This involves managing access to the data sets, including the linked data set for subsequent use projects as per the process specified in sub-process 8.1.

**8.4 Destruction of files**

At the end of the approved retention period, the linked data set is destroyed or an extension is requested and approved prior to the end of the retention period.

**Output: Disclosure and access protocols**

## 9. Evaluate

This phase focuses on the evaluation of the record linkage project as opposed to evaluation of the data linkage which is conducted in Phase 2. The evaluation may include aspects of the record linkage process as well as general project management. During this phase, relevant aspects of the record linkage project that may inform future projects are also identified.

This phase is made up of sub-processes, which are generally sequential, but which can overlap to some extent in practice:

| 9.1 | 9.2 | 9.3 | 9.4 |
|---|---|---|---|
| Gather evaluation inputs | Conduct evaluation | Agree on action plan | Add to the record linkage "toolbox" |

### 9.1 Gather evaluation inputs

Evaluation material can be produced in any other phase or sub-process. It may take many forms, including feedback from users, process metadata (paradata), system metrics, and staff suggestions. Reports of progress against an action plan agreed to during a previous iteration may also form an input to evaluations of subsequent iterations. This sub-process gathers all of these inputs, and makes them available for the person or team producing the evaluation.

### 9.2 Conduct evaluation

This sub-process analyses the evaluation inputs and synthesizes them into an evaluation report. The resulting report should note any quality issues specific to this iteration of the statistical business process, and should make recommendations for changes if appropriate. These recommendations can cover changes to any phase or sub-process for future iterations of the process, or can suggest that the process is not repeated.

### 9.3 Agree on action plan

This sub-process brings together the necessary decision-making power to form and agree to an action plan based on the evaluation report. It should also include consideration of a mechanism for monitoring the impact of those actions, which may, in turn, provide an input to evaluations of future iterations of the process.

### 9.4 Add to record linkage "toolbox"

This sub-process involves identification of information relevant for inclusion in a record linkage toolbox. The toolbox is a depository of record linkage concepts, best practices, tools and training material. The purpose of the toolbox is to facilitate knowledge transfer between individuals involved in current or future record linkage activities with an ultimate goal of increasing efficiency and quality of record linkage at the organizational level.

**Output: Evaluation Report; Contribution to the Record Linkage Toolbox**

## Appendix A: Generic Statistical Business Process Model v5.0

### Levels 1 and 2 of the Generic Statistical Business Process Model

| Quality management / Metadata management | | | | | | | |
|---|---|---|---|---|---|---|---|
| **1 Specify needs** | **2 Design** | **3 Build** | **4 Collect** | **5 Process** | **6 Analyse** | **7 Disseminate** | **8 Evaluate** |
| **1.1** Identify needs | **2.1** Design outputs | **3.1** Build collection instrument | **4.1** Create frame and select sample | **5.1** Integrate data | **6.1** Prepare draft outputs | **7.1** Update output systems | **8.1** Gather evaluation inputs |
| **1.2** Consult and confirm needs | **2.2** Design variable descriptions | **3.2** Build or enhance process components | **4.2** Set up collection | **5.2** Classify and code | **6.2** Validate outputs | **7.2** Produce dissemination products | **8.2** Conduct evaluation |
| **1.3** Establish output objectives | **2.3** Design collection | **3.3** Build or enhance dissemination components | **4.3** Run collection | **5.3** Review and validate | **6.3** Interpret and explain outputs | **7.3** Manage release of dissemination products | **8.3** Agree on action plans |
| **1.4** Identify concepts | **2.4** Design frame and sample | **3.4** Configure workflows | **4.4** Finalise collection | **5.4** Edit and impute | **6.4** Apply disclosure control | **7.4** Promote dissemination products | |
| **1.5** Check data availability | **2.5** Design processing and analysis | **3.5** Test production system | | **5.5** Derive new variables and units | **6.5** Finalise outputs | **7.5** Manage user support | |
| **1.6** Prepare business case | **2.6** Design production systems and workflow | **3.6** Test statistical business process | | **5.6** Calculate weights | | | |
| | | **3.7** Finalise production systems | | **5.7** Calculate aggregates | | | |
| | | | | **5.8** Finalise data files | | | |

**Appendix B: Glossary of Terms**

| | |
|---|---|
| Data Custodian | The senior manager responsible for assigning a responsible manager for the linked analysis file; for ensuring that the file is managed according to the Directive on the Management of Statistical Microdata sets and the Directive on Record Linkage and all related legislation, policies and other governing documents; and for managing access to the file. |
| Direct identifiers | Means any information that could directly identify an individual person, business or organization, but that is not information that is used for statistical analysis. Examples of direct identifiers are name and address (street and postal code). |
| Entity | Refers to an individual respondent or unit of observation, such as a person, family household, dwelling, farm, company, business, establishment, etc. (Source: Directive on Record Linkage) |
| Micro-record | Information about an identifiable entity. (Source: Directive on Record Linkage) |
| Linked data set | The file containing the composite records resulting from the linkage of two or more source data sets. |
| Linking keys | The unique identifiers of the source micro-records that were linked and their association. |
| Linkage variables | Linkage variables are defined as those variables used to link the source data sets. Linkage variables contain personal identifiers that are used in the record linkage process. Linkage variables could include unique identifiers (e.g. Social Insurance Number, Health Insurance Number) or other personal identifiers such as date of birth, postal code or sex. |
| Record Linkage | The combining of two or more micro-records to form a composite record containing information about the same entity. The output of a record linkage must contain information that originated from more than one data set data set that were inputs to the record linkage activity. (Source: Directive on Record Linkage – Appendix A) |
| Source data set | The source sets are defined as the input data sets linked in the record linkage process. Source data sets can be structured in two ways: *source index files* contain personal identifiers *without analysis variables* and *source data set* contain analysis variables *without personal identifiers*. |

| | |
|---|---|
| Sponsor | The senior manager responsible for assigning a responsible manager for the record linkage project; for ensuring that the project is managed according to project existing management policies, directives and guidelines. |
| Statistical Matching | Statistical matching (also known as data fusion, data merging or synthetic matching) is a model-based approach for providing joint statistical information based on variables and indicators collected through two or more sources. The potential benefits of this approach lie in the possibility to enhance the complementary use and analysis of existing data sets (e.g. cross-cutting statistical information that encompasses a broad range of socio economic aspects), without further increasing costs and response burden.<br><br>An essential feature of statistical matching is that, although the units in the concerned data sets should come from the same population, they are usually not overlapping. You identify and link records from different sources that correspond to similar units. This is the basic difference compared with record linkage, where units included in the data sets overlap that allows to link records from the different data sets that correspond to the same unit. Therefore, **record linkage** deals with **identical** units, while **statistical matching**, or synthetic linkage, deals with '**similar**' units.<br><br>Statistical matching: a model based approach. Eurostat Methodologies and Working Papers, European Union 2013. |
| Unique Identifier | A numeric or alphanumeric string that is associated with a single entity (individual) within a given service delivery program or system. Examples of unique identifiers include a Social Insurance Number, a Personal Health Information Number, a student number, a driver's license number, etc. |

## Appendix C: Source Documents

### References:


Agency For Healthcare Research and Quality. Linking Data for Health Services Research: A Framework and Instructional Guide. AHRQ Publication No 14-EHC033-EF, September 2014.

Australian Institute of Health and Welfare and Australian Bureau of Statistics 2012. National Best Practice Guidelines for Data Linkage Activities Relating to Aboriginal and Torres Strait Islander People AIHW Cat. no. IHW 74. Canberra: AIHW. (http://www.aihw.gov.au/publication-detail/?id=10737422216)

Centre for Health Record Linkage. Edit checks to perform before analysing your linked data. 2013(http://www.cherel.org.au/media/28218/dataset_edit_checks_-_july_2013.pdf)

Joint UNECE / Eurostat / OECD Work Session on Statistical Metadata (METIS). Generic Statistical Business Process Model (GSBPM) v5.0, 2014. (http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0)

Fellegi, I.P., and Sunter, A.B. (1969), "A Theory of Record Linkage", *JASA*, 64, pp. 1183-1210.

Christen, P. Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin: Springer, 2012, section 6.5, p. 139.


### Directives and Policies:

- Directive on Record Linkage
- Directive on Informing Survey Respondents
- Policy on Privacy Protection (Treasury Board Secretariat (TBS))
- Directive on Privacy Impact Assessment (TBS)
- Directive on Privacy Practices (TBS)
- Directive on Social Insurance Number (TBS)
- Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes (United Nations)
- Policy on Privacy and Confidentiality (Statistics Canada)
- Policy on Information Management (Statistics Canada)
- Policy on Informing Users of Data Quality and Methodology
- Directive on the Management of Statistical Microdata sets (Statistics Canada)
- Directive on Obtaining Administrative Data under the Statistics Act
- Directive for the Validation of Statistical Outputs
- Directive on the Use of Deemed Employees
- Directive on Discretionary Disclosure
- Directive on the Security of Sensitive Statistical Information

## Appendix D: Circulation process for approval of record linkage applications

### Formal requests are circulated via routing docket to the following:

Director to program area submitting the request
Director General of the Program Area
Record Linkage Section of IMD
Director of IMD
Director of ADD (when TAX files are linked)
Director of COD (when Census files are linked)
Assistant Chief Statistician of Program Area
Chief of Staff
Chief Statistician

### Omnibus requests are approved electronically by:

Director to program area submitting the request
Director of IMD
Director of ADD (when TAX files are linked)
Director of COD is informed (when Census files are linked)

### Linkage requests submitted as part of the Survey Prescription

#### Approved electronically by:

Director of IMD
Director of ADD (when TAX files are linked)
Director of COD (when Census files are linked)

#### Approved via routing docket by:

Assistant Chief Statistician of Program Area
Departmental Secretary
Chief Statistician

### Secondary Use of linked files are approved electronically by:

Director of custodian division
IMD