Catalogue no. 18-001-X
ISBN 978-0-660-06020-0

Reports on Special Business Projects

An Overview of Selected International Business Record Linkage Programs

by Julio Miguel Rosa

Release date: October 27, 2016

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Table of contents

Acknowledgments

I would like to thank the following persons for their support, encouragement and comments in the writing of this report: George Sciadas, Frances Anderson, Jean-François Dubois, Amélie Angers and Danny Leung as well as the two anonymous referees. Finally, special thanks are extended to my colleagues in the statistical agencies who took the time to respond to the mini-survey sent out by Statistics Canada.

An Overview of Selected International Business Record Linkage Programs

by **Julio Miguel Rosa**

1. Introduction

The main objective of this report is to review selected business record linkage programs and practices commonly used by statistical agencies across the world.

> The definition of record linkage[1] that will be used in this report is the following:
>
> "**Record linkage** is defined as 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 two or more datasets that were inputs to the record linkage activity. In other words, record linkage is the integration of several sources of information in the form of independent data. Record linkage continues to be an important technique for developing data, producing information and databases, and conducting statistical analysis and evaluation of data."
>
> **Source:** *Statistics Canada (August 2011). Internal document entitled "Directive on Record Linkage."*

Statistics Canada's experience in this regard will serve as a benchmark when describing the approaches adopted by other statistical agencies. There is a reason for this choice, in that Statistics Canada has some of the most longstanding and extensive expertise in record linkage. Internationally, Statistics Canada is a leader in developing linkage methods, including the theoretical work which Fellegi and Sunter (1969) began, and which Fellegi, by then Chief Statistician of Canada, pursued in 1999 (Fellegi 1999). Today, many statistical agencies throughout the world are inspired by Statistics Canada's model and practices in record linkage.

Although record linkage is widely used for social data, especially in the area of health and epidemiology (Winkler, W. E. 1999, Newcombe et al. 1992), this report focuses mainly on record linkage programs related to business data. One of the reasons for this choice is that worldwide, there is renewed interest in the linkage of business data. The very nature of business activities is leading Statistics Canada to take special restrictive measures on issues relating to protecting the confidentiality of business databases.

To make it easier to compare different linkage methods and practices at the international level, in addition to drawing on the information publically available, a mini-questionnaire was prepared and sent to a number of statistical agencies to collect uniform information on their record linkage methods, practices and issues.

This report is not an exhaustive and detailed list of international practices, but it will give readers a general overview of what is currently done in the field of record linkage. It will be of particular interest for readers seeking to get a grasp of international practices and experience in record linkage and to understand the main issues for statistical agencies.

The report is organized as follows. In the following section, the general record linkage practices methods are presented. Section 3 is dedicated to the review of published available information of international record linkage practices and programs. Finally, Section 4 summarizes the results of a mini survey sent to different international statistics agencies.

2. General record linkage practices

Data matching process

To respond to the growing demand for increasingly complex and detailed data, statistical agencies are incorporating more information from multiple sources and combining the information to improve its quality, increase the quantity of information available, share this data and make it possible to conduct more detailed

1. Throughout the report, the terms "record linkage" and "data matching" will be considered equivalent.

analyses (Christen 2012a). Techniques for linking information on individuals with illnesses are widely used in the health field. These techniques lend themselves to better monitoring of heart disease and contagious diseases, at a lower cost than is possible in a non-integrated disease surveillance system. Beyond the field of social, business and health data, record linkage has shown strong growth in fields such as the development of research-oriented websites (Su, Wang, and Lochovsky 2009), but also in fields such as crime, fraud prevention and terrorism. This latter field of application is of great importance for national security issues and is contributing to the renewed interest in these techniques (Larsen 2006, Jonas and Harper 2006).

Record linkage is above all a powerful instrument to assist governments and other institutions in their decision-making. It increases considerably the analytical potential of micro databases as well as their quality,[2] especially when administrative data are linked with survey data. In addition to being an excellent tool for analysis and research, record linkage greatly reduces the response burden as well as cost and time involved in collecting information.

There are mainly two types of record linkage: the **deterministic linkages** and the **probabilistic linkages**.

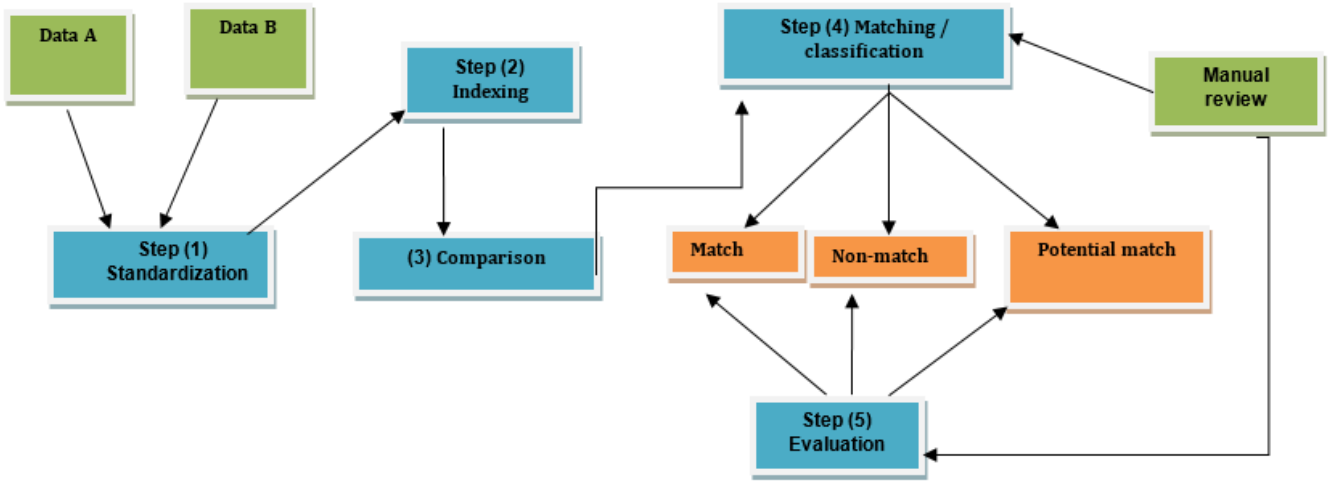
> "Probabilistic linking and deterministic linking are methods for combining information from records on different datasets to form a new linked dataset. Probabilistic linking has been described as a process that attempts to link records on different files that have the greatest probability of belonging to the same businesses/person. Whereas deterministic linking uses a unique identifier to link datasets, probabilistic linking uses a number of identifiers, in combination, to identify and evaluate links.
>
> Probabilistic linking is generally used when a unique identifier is not available or is of insufficient quality. The method derives its name from the probabilistic framework developed by Fellegi and Sunter (1969) and requires sophisticated software to perform the calculations."
>
> **Source:** *National Statistical Service. Australian Government- Data linking information series* (http://www.nss.gov.au/nss/home.nsf/NSS/D46ECD302D0D984FCA257BC700237287?opendocumant)

Probabilistic record linkage

Figure 1 below provides a schematic illustration of the record linkage process generally used by most statistical agencies. This figure is extensively inspired by the model presented in Christen, P. (2012a). This illustration provides an example of two databases (A and B) to be linked. In practice, it is are not necessarily limited to linking two databases; one could of course link several databases simultaneously. A technique must be selected at each step in this process. Although many techniques can yield equivalent results, some techniques may be more effective than others, regardless of the needs and configuration of data. Part of the application of this process depends on the know-how and judgment of the statistician applying the technique. The statistician must determine the relevance of each technique at each step in the process in order to optimize linkage results. Hence, for matching where the identifiers are known with certainty, it is not necessary to use complex approaches such as probabilistic matching, which apply when the databases have no common identifier. For example, Statistics Canada primarily uses deterministic matching techniques to link business surveys in the Statistics Canada Linkable File Environment (LFE). Probabilistic methods will only be used when it is difficult or impossible to obtain an identifier.

2. Record linkage can improve the quality of a database by exploiting the best information available in various information sources. If, for example, the variable showing the number of employees is of better quality in database A than in database B, then it is best to keep the employment variable from database A in the linked database C.

Figure 1
Classical diagram for probabilistic record linkage



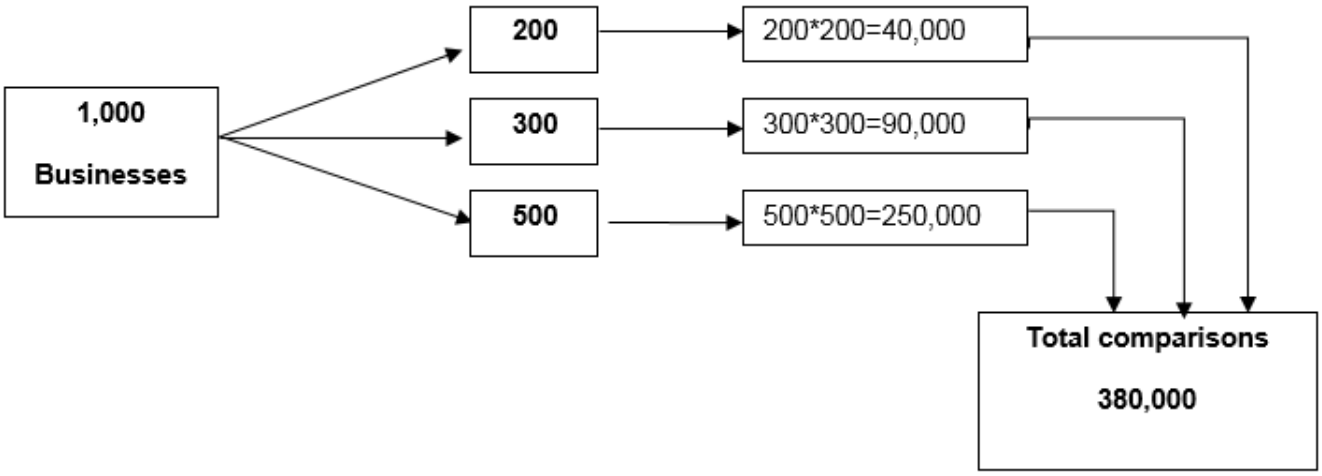
Source: Christen, P., (2012b)

Step 1 – Standardization: This record linkage process consists of standardizing syntax, distributing characters into identifiable fields in a coherent form and format to facilitate processing (parsing), transformation and editing of data. For example, standardization can involve assigning the term 'Corporation' to all variable fields containing the name of a company or editing a date to the same format. Parsing consists of separating a business address into easily identifiable fields, such as number, street, postal code or province. Transformation could include changing lowercase characters to uppercase, or vice versa, or changing a numeric character to an alphanumeric character. Editing would involve excluding an observation that was input by mistake. Therefore, this step is essential in the preparation of the database. This step is sometimes called pre-grooming.

Step 2 – Indexing: The indexing process consists of generating pairs of candidates for matching. The goal is to use a technique that can minimize the number of observation pairs to be compared. Professor Peter Christen identified six indexing methods (see articles Christen, (2012a); Baxer et al. (2003)] for more details.[3] This report will describe only the most common technique: Blocking. With this technique, only the units in a common block are compared. If a database contains 1,000 observations to be compared with each of the 1,000 other observations from the other database (assuming there is no duplication), a million pairs of observations would have to be compared to select only one pair of observations. For example, a simpler way of counting businesses would be to separate the businesses into common blocks where each block contains businesses that have the same number of employees, and then compare the businesses within the same block. Figure 2 illustrates the example where 1,000 businesses need to be compared. If blocks of businesses of the same size are created, let's say that one obtain three sizes or three blocks of 200, 300 and 500 businesses respectively. Then, only 40,000, 90,000 and 250,000 business pairs would have to be compared within each block, compared to one million if the businesses had not been separated.

3. The six techniques detailed by Christen (2012b) are: *1-Traditional Blocking; 2-Sorted Neighbourhood indexing; 3-Q-gram Based Indexing; 4-Suffix Array Based Indexing; 5-Canopy Clustering; 6-String-Map Based Indexing.*

Figure 2
Example of blocking



There are also many phonetics-based algorithms that can be used to create blocks. The two most common types are Soundex and the New York State Identification and Intelligence System (NYSIIS).[4] These algorithms use phonetic recognition to code names and alphanumeric fields. Chapter four of Peter Christen's book explains these algorithms and provides a few comparative tables on the difference in the results obtained using these various algorithms when applied to the same alphanumeric field. For more details on these phonetic coding systems, see Herzog et al. (2007) Chap. 11 and Part II Chap. 4.

Step 3 – Comparison: This process consists of comparing the pairs of observations selected in Step 2. The character strings are compared between the observation pairs, assigning them a score that is calculated using algebraic functions of character string comparison. These functions make it possible to compare the similarities between observation pairs using alphanumeric character strings. Indeed, observation pairs can contain minor differences in the names of the variables used for comparison. The algebraic functions of character string comparison take into account both the length of the character strings and the possible errors in these characters, such as a transposed letter in the names. The most commonly used algebraic string comparators are: Jaro-Winkler String Comparison; Q-gram Based String Comparison; Edit Distance String Comparison (Smith-Waterman edit distance); Monge-Elkan String Comparison; Extended Jaccard comparison; and Syllable Alignment Distance. For details on the algorithms of these algebraic functions of character string comparison, see Christen, (2012b) Part II Chap. 5.

Step 4 – Matching and classification: This process is the record classification and linkage operation. Sometimes, the two operations are separated into two different steps. The goal of classification is to be able to place observation pairs according to whether they are a *match, non-match or potential match*. In doing so, the user must determine the order of priority of the rules for defining matches between observation pairs. For example, users may decide that the concordance rule will be based on the address, name of the business and the name of the owner. For these rules, concordance level probabilities will be calculated (scores) and assigned to each pair of observations. The decision to classify the observation pair in the match, non-match or potential match category is based on the concordance similarity score. The score falls within a value interval between M (for matched observations) and U (for non-matched observations). The rule of decision could be probabilistic or deterministic (see section II-A "Objectives and types of linkages''). Although the classification step is based on a statistical method, for observation pairs classified in the "potential match" category, human intervention will be required to place these pairs in either the "match" or "non-match" category.

Step 5 – Evaluation: Once the mathematical algorithm has determined the observation pairs that will be matched, non-matched or potentially matched, match quality must be evaluated. This step consists of distinguishing true

4. Opposite is a list of other phonetic systems used for the indexing step: Phonex; Phonix; Oxford Name Compression Algorithm; Double-Metaphone; Fuzzy Soundex.

matches from false matches, and classifying them according to linkage error type or match category as indicated in Table 1 below.

Table 1
Match categories (Type I and Type II linkage errors)

	The observation is associated with the same unit	**The observation is not associated with the same unit**
The observation is matched	The observation is correctly matched (*true match*) ***(TM)***	The observation was matched by mistake - **Type I error or false negative match (FN**)
The observation is not matched	The observation was not matched by mistake - **Type II error or false positive match (FP)**	The observation was correctly not matched (true non-match) ***(TN)***

This step requires human intervention to differentiate between a false negative match (FN) = the observation was matched by mistake – Type I error, and a "false positive match (FP) = the observation was not matched by mistake – Type II error" by minimizing these two types of errors. The other two categories are true match (TM) and true non-match (TN). Table 1 illustrates the types of possible association errors. It is in this step that match quality is evaluated. At the statistician's discretion, a review may be required when matching using known algorithms cannot be done or requires a decision by the statistician [Guiver 2011]. In this step, judgment is relied on to determine the most appropriate linkage. Generally, whether for deterministic matching or probabilistic matching, systematic manual validation remains an important step for ensuring linkage quality.

3. An overview of the international record linkage programs

Contextual framework of record linkage

In Canada, record linkage programs have existed since the 1970s in the health field (http://www.statcan.gc.ca/health-sante/link-coup-eng.htm). For example, Canadian data on mortality, cancer, births and stillbirths have been combined for use in many scientific studies and the monitoring of serious diseases. For statistical agencies, matching techniques are important both for producing information that would otherwise be impossible to trace and for analysis, but also for improving the quality of the statistical system in general.

For most agencies, the use of record linkage requires an ethical and regulatory framework with respect to privacy and confidentiality (see Annex A). Most statistical agencies have adopted policies and directives covering the production and use of linked data. Statistics Canada is mainly governed by the federal ***Statistics Act*** and guided by the **Policy on Privacy Protection** [Statistics Canada (1985)] as well as the **Directive on Record Linkage** [Statistics Canada (2011)] as a framework for implementing linkage practices. Other statistical agencies have their own policies and laws. One example is Australia, whose counterpart to the Statistics Canada legislation is the *Australian Bureau of Statistics Act 1975* (http://www.comlaw.gov.au/Details/C2012C00137) and the *Census and Statistics Act 1905* (http://www.comlaw.gov.au/Details/C2006C00178). The Australian Bureau of Statistics can also find support in the Privacy Act 1988 (http://www.oaic.gov.au/privacy/privacy-act/the-privacy-act) to regulate statistical activities involving confidentiality and privacy.

Policies and directives on record linkage in Canada

Statistics Canada's mandate is to collect, compile, analyze, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and condition of the Canadian people (Section 3 of the federal *Statistics Act*). Record linkage is one of the activities included in the mandate of Statistics Canada and other agencies throughout the world. But before approving a record linkage, most national statistical agencies impose prior conditions to preserve respondents' privacy and the confidentiality of information on businesses while ensuring that the linkage is carried out within a framework of law and mutual consent. To satisfy these conditions, Statistics Canada has, since 1986, a Directive on Record Linkage as well as a Policy on Privacy and Confidentiality and a Policy on Privacy Protection [Statistics Canada (1985)].

The Directive on Record Linkage stipulates that each record linkage proposal must form a separate submission. The proposal should include an outline of the proposed research plans and briefly state the objectives of the research project or study. The purpose for undertaking the proposed record linkage must be described and analyzed in detail, including the key reasons for conducting the linkage and the intended use of the results.

The record linkage proposal is submitted to the Executive Management Board. The Board will also assess the benefits of the linkage, how the public interest is served by the project and why a record linkage is the best means to achieve this public benefit. The Board will also consider any efficiencies and/or savings in terms of costs, resources, timeliness, and reduced response burden that will result from the linkage. To reduce the privacy invasiveness of the linkage, identifiers must be stripped from the linked file as soon as the linkage is completed, and stored separately. The Board also requires that the management and retention of all record linkages be kept up to date. Authorized linkages must be posted on the Statistics Canada website. For linked data, the retention period must be stated. Divisions responsible for the linkages are responsible for determining the retention period for the linked files and for the destruction of these files. In some cases, the record linkage may be covered by an existing, approved omnibus or on-going record linkage, in which case a new submission is not required.

Statistics Canada ensures that respondents' privacy and the confidentiality of information are protected. Statistics Canada's commitment to protect the information transmitted to it by the Canadian public is guaranteed by the *Statistics Act* and the various policies and practices that frame data collection, analysis and dissemination activities.

In accordance with the requirements set out under the *Statistics Act*, Revised Statutes of Canada, 1985, Statistics Canada must give respondents notice that it plans to link their responses to the survey with data from other surveys or administrative files. The notice may be either specific or general. If a respondent to a voluntary survey states that he/she is opposed to the planned record linkage, no linkage of that respondent's responses will be allowed.

Furthermore, with respect to record linkage, the conditions listed below must be met to perform the linkage.

- The purpose of the record linkage activity is statistical/research and is consistent with the mandate of Statistics Canada as described in the *Statistics Act*.
- The products of the record linkage activity will be released only in accordance with the confidentiality provisions of the *Statistics Act* and the relevant provisions of the *Privacy Act*.
- The record linkage activity has demonstrable cost or respondent burden savings over other alternatives, or is the only feasible option.
- Record linkage activity will not be used for purposes that can be detrimental to the individuals involved, and the benefits to be derived from such a linkage are in the public interest.
- The record linkage activity is judged not to jeopardize the future conduct of Statistics Canada's programs.
- The linkage satisfies a prescribed review and approval process (this includes review by Statistics Canada's Confidentiality and Legislation Committee and Executive Management Board, as well as ministerial review for types of linkages not previously approved by the Minister; new kinds of major linkage projects are discussed with the Office of the Privacy Commissioner).
- The linked files will be destroyed when the project is completed, in accordance with the date prescribed for destruction of files.

Source: Statistics Canada, Policy Manual, 4.1, Policy on Record Linkage, Treasury Board Secretariat.

A few examples of record linkage initiatives at the international level

Statistical agencies' growing interest in record linkage is increasingly leading them to develop, define and document their methodological approaches and to develop internal policies on record linkage. In consideration of the fact that data linkage experiences and procedures vary from one statistical organization to another, the sections that follow present an overview of a number of statistical agencies' recent experiences with record linkage. Countries were selected solely on the basis of documentation available on the Internet. Insofar as possible, emphasis was placed on what distinguishes each agency's data linkage system.

Canada

To meet growing needs for economic information and with the support of federal government departments, Statistics Canada undertook an ambitious project on data integration in the business sector, with its initial phase beginning in 2008. This project, known as the Linkable File Environment (LFE), was implemented through the creation of a relational database that associates numerous information sources (surveys and administrative data) with the Business Register (BR), which constitutes the reference database. This integrated environment is comprised of several stages, including the transfer of data from various Statistics Canada sources, quality assurance of the linkage and the variables used, and production of a report on linkage quality and dictionaries of the variables. In the majority of cases, record linkage is relatively straightforward, since the identifiers are known.

Because most internal Statistics Canada business sources have unique identifiers determined by the Business Register, the match rate is close to 100% (deterministic linkage). However, external information sources do not have this unique identifier and one must use probabilistic matching techniques, as in the case for patent data. The match rate for this type of linkage is quite variable, since it depends on the quality of the information source. Without an identifier obtained from the Business Register, it is necessary to use other information about the record, such as the name or the address in the record (probabilistic linkage).

The LFE is not a longitudinal database but is rather a cross-sectional database compiled over a 15-year period (from 2000 to 2014). It currently includes some twenty sources of administrative data and survey data. The originality of the LFE lies in its wide use and its ability to extract linked databases on a custom basis for the researchers. Indeed, since its introduction, a large and growing number of research and analytic projects have been carried out based on this environment.

In recent years, Statistics Canada has made considerable efforts to facilitate access to microdata. In particular, there are ongoing efforts to extend access to academics and various government agencies. The Canadian Centre for Data Development and Economic Research (CDER) has been established to provide facilitated access to external researchers to Statistics Canada's holdings of business microdata, including linked datasets. The results produced by a researcher must adhere to strict confidentiality rules for their release to be authorized. (For more details see: http://www.statcan.gc.ca/cder-cdre/index-eng.htm).

New Zealand

New Zealand is one of the countries for which considerable record linkage documentation is available on the Internet specifically on the official website of Statistics New Zealand (www.stats.govt.nz). What distinguishes New Zealand's record linkage system is that it was developed from an **Integrated Data Infrastructure (IDI) prototype**. This infrastructure is designed to create an environment of linked longitudinal microdata including information covering individuals, households and businesses. Access to this information is possible for any researcher with a "bona fide research goal" (as assessed by the Government Statistician). Business tax data are restricted to government department employees (including those contracted by government departments) and access is also subject to approval by Inland Revenue.

New Zealand has a statistical system very similar to Canada's, especially with respect to the policies and directives on integration of data from separate sources which minimize risks to the confidentiality of information on individuals and businesses (Statistics New Zealand 2012a and 2012b). The IDI is similar to what exists at Statistics Canada with the LFE, but it goes further in that it also integrates data on individuals and households. However, at Statistics Canada one of the main source for longitudinal business data is the Longitudinal Employment Analysis Program (LEAP) that contains historical employment information on employer firms.

The Privacy Act prohibits the application of a universal identifier between agencies, but does not prevent Statistics New Zealand from using the unique identifiers of data-supplying agencies to link data within the IDI and to create new longitudinal identifiers for research purposes (which only exist within the secure data environment). For example, Statistics New Zealand uses Inland Revenue (IR) numbers to link data supplied by various agencies. Statistics New Zealand creates a new longitudinal identifier that allows researchers to track units over time, but prevents identification of the linked record via the original IR number. Without the ability to create a common identifier across data sources, the data would be neither integrated nor longitudinal.

The IDI was established to exploit the power of integrated data to aid decision makers. Statistics New Zealand was seen as a natural home for the integration project because of its prior track record as a data linker/custodian and the related level of public trust, and because of the unique role the Statistics Act assigns to the agency. In principle, any government department could have hosted the IDI.

The data integration policy of Statistics New Zealand contains four overarching principles:

> **Principle 1:** The public benefits of integration must outweigh both privacy concerns about the use of data and risks to the integrity of the Official Statistics System, the original source data collections, and/or other government activities.
>
> **Principle 2:** Integrated data will only be used for statistical or research purposes.
>
> **Principle 3:** Data integration will be conducted in an open and transparent manner.
>
> **Principle 4:** Data will not be integrated when an explicit commitment has been made to a respondent that prevents such action.
>
> **Source:** Statistics New Zealand (2012a)

New Zealand has considerable experience with record linkage. Since 1997, this has included the following projects: Linked Employer-Employee Data (LEED); the Longitudinal Business Database (LBD) prototype; the Household Labour Force Survey (HLFS); Employment Outcomes of Tertiary Education (EOTE) and the Student Loans and Allowances (SLA) integrated dataset. The IDI program integrates data from the Department of Labour, Migration and International Movements with the Longitudinal Business Database (LBD), while consolidating the existing matched data described above. This program has received authorized funding to 2020, with the goal of constantly improving the data integration program. To obtain more information on this program, see Statistics New Zealand (2012a).

Germany

Like some other countries,[5] Germany has a private research centre entirely devoted to data integration services, the German Record Linkage Center (German RLC). This centre is currently conducting some 15 record linkage projects.[6] For more information, see: http://soz-159.uni-duisburg.de/linkage/?Projects.

The matched data of the government and the German RLC are available for research purposes only. In addition to providing information services, the German RLC provides researchers with access to the Merge ToolBox (MTB) data integration software.

In Germany, respondents' consent is required in order to link their information; otherwise the agreement of the federal data protection agency must be obtained. Also, to preserve respondent confidentiality, Germany is exploring methods of encrypting identifiers (Schnell et al. 2004).

One example concerns the integration of data from the German Business Register (URS) with administrative data from institutions such as the Federal Employment Agency and the Deutsch Bundesbank. This program is called the KombiFi (Kombinierte Firmendaten für Deutschland (see Konold and Assainato (2009) for further details). In this project, some 60,000 firms are asked permission to link information concerning them that was formerly held separately by various separate institutions (statistical agencies, banks and the Federal Employment Agency). Of these 60,000 firms, 16,571 agreed (Vogel and Wagner 2012).

Another example is the German experience with linking administrative data on employment with data from the ALWA survey (Arbeiten und Lernen im Wandel—working and learning in a changing world),[7] which contains

5. Other countries included are Australia: Western Australia Data Linkage Branch; United Kingdom: Oxford Record Linkage Group; New Zealand: Department of Public Health and General Practice, University of Otago.
6. Record linkages include: 1) Linkage of apoplectic stroke documentations on different stages of therapy in Hesse, on behalf of the Regional Office for Quality Assurance Hesse (GQH); 2) Linkage of the German SAVE study with administrative data from the Federal Employment Agency; 3) Private Equity and Employment; 4) Wage structure and patterns of employment in the energy sector; 5) Innovation behaviour, regional clustering and agglomeration effects of the German biotechnology industry; 6) Occupational health research in epidemiological cohort studies (Aeko); 7) Panel Analysis of Intimate Relationships and Family Dynamics (PAIRFAM); 8) Survey data from the panel study 'Labour Market and Social Security' (PASS) linked to administrative data from the IAB; 9) Interactions between realisation opportunities in working life and in private life; 10) Geocoding of IAB administrative data; 11) Linkage of new migration study of the GSOEP with IEB data from the Federal Employment Agency; 12) Record linkage of administrative patent-inventor data with administrative data on labour market biographies; 13) Pro Kind, a randomized, controlled trial of early childhood intervention in Germany.
7. For more information, see http://fdz.iab.de/en/FDZ_Individual_Data/ALWA.aspx.

longitudinal information on education, place of residence, employment, marital status, regional mobility, etc. for 10,400 individuals born between 1956 and 1988. The interviews for these individuals were conducted between August 2007 and April 2008. The results of this linkage have been the subject of two detailed studies (Antoni 2011, Antoni and Seth 2012). The data from these two information sources were standardized by correcting typographic errors on all the variables. With a deterministic approach, 53% of the 10,400 respondents were matched. Using the probabilistic approach (Jaro-Winkler method), the percentage of matching successes was raised to 83%, and with manual editing this percentage reached 86%. This linkage exercise made it possible to overcome the information deficiencies of each of the data sources. For example, the administrative data contained little or no information on education, while the survey data lacked details on income. The development of data linkage in Germany is recent: previously, only two studies (Beste 2011, Hartmann and Krug 2009) had been conducted in Germany on similar data (employment data and administrative information).

OECD

In 2010, Eurostat, the statistical office of the European Union, funded a major project for a two-year period, designed to integrate the survey data and administrative data of a number of countries. In all, fifteen agencies[8] joined together to create a common and cooperative project called ESSlimit.[9] The purpose of the project was to establish new indicators on the characteristics of firms in the information technology field, their innovation efforts and their economic performance (OECD 2012). The project was to be less oriented toward basic academic research and more oriented toward partnership among national statistical institutions (NSIs) for data development. The integration of information from a number of surveys[10] and administrative data from various countries can be used to produce indicators at the enterprise level.

The combined information was estimated using weightings common to all the countries, which suggests that the published results could not be compared with figures from official publications (OECD 2012). Methodological efforts focused on producing indicators that were consistent from one country to another (data from different countries were processed according to the same rules), making this project a unique source of information for comparison between countries, by industry and over time. Several data integration methods were used, including probabilistic record linkage. In the majority of cases, the unique identifier at the enterprise level was sufficient to link firms on the basis of different surveys.[11]

This data integration project is interesting in several respects. Conceptually, it served to compile identical and comparable indicators at an industry aggregation level across multiple countries over several years, thus showing that characteristics of a number of countries' science and technology firms can be integrated in a consistent manner.

Eurostat recently produced a report (Eurostat 2013) on the results and status of projects initiated by Essnet (also integrates data on individuals). There are thus some 20 projects, either completed or under way. Some of these projects focus on the methodological development of data. For example, ESeG (European socio-economic classification), a project coordinated by France, is designed to construct a socioeconomic classification system covering individuals from all the countries of Europe with similar socioeconomic characteristics and a comparable lifestyle. Another project, coordinated by Denmark, has as its main objective to conceptualize measurement of the global value chains (GVCs) and strengthen the methodology for it. This project has led to analyses on the measurement of economic globalization and its impacts on the creation of new jobs and economic growth in European countries. Other projects have been more oriented toward the development of computer tools. For example, the SDMX project, coordinated by Istat (Italy), seeks to develop an informatics infrastructure whose purpose is to integrate different existing statistical tools.

These few examples show all the potential and appeal of data linkage in Europe. Not only can researchers access microdata that are comparable between countries but also, for the first time, these data provide access to an unprecedented level of quality when comparing indicators, for both social data and data on enterprises. With the

8. Including those of Austria, Denmark, Finland, France, Germany, Italy, Luxembourg, the Netherlands, Norway, Poland, Romania, Slovenia, Sweden and the United Kingdom.
9. The full name of this project is ESSNET on Linking Microdata on ICT Usage.
10. Including the Community Innovation Survey (CIS); Statistics on Foreign Trade; Community Survey on ICT usage; Business R&D Survey and Production Survey (PS); E-commerce in Enterprises (EC); Structural Business Statistics (SBS) and External Trade in Services.
11. For more information on final report see Hagsten et al. (2012); for further details on the different stages of the project, see the following Eurostat site: http://ec.europa.eu/eurostat/documents/341889/725524/2010-2012-ICT-IMPACT-2012-Final-reportépdf/90cf5094-334a-4ff1-8f60-047c2d650c60.

success of this project, there is now the prospect of comparing indicators among nations outside the European community.

Australia

The Australian statistical system is very similar to what exists at Statistics Canada. The Australian system is notable for its great decision-making transparency with respect to co-operation among organizations in the approval process for data integration (*Census and Statistics Act 1905*). Australia has a data integration decision-making committee (Cross Portfolio Statistical Integration Committee—CPSIC) that includes all government departments. Since 2009, government agencies have worked together in developing its statistical system of integrated data for research purposes. In February 2010, this collaboration led to the adoption by the CPSIC[12] (headed by the Australian Bureau of Statistics) of seven "high level principles for data integration involving Commonwealth data statistical and research purposes." These principles set out the guidelines that apply to any record linkage project. They are summarized in Table 2.

Table 2
The seven principles for data integration involving Commonwealth data

Principle 1: Strategic resource	Responsible agencies should treat data as a strategic resource and design and manage administrative data to support their wider statistical and research use.
Principle 2: Custodian's accountability	Agencies responsible for source data used in data integration are individually accountable for their security and confidentiality.
Principle 3: Integrator's accountability	A responsible "integrating authority" will be nominated for each statistical data integration proposal.
Principle 4: Public benefit	Statistical integration should only occur where it provides significant overall benefit to the public.
Principle 5: Statistical and research purposes	Statistical data integration must be used for statistical and research purposes only.
Principle 6: Preserving privacy and confidentiality	Policies and procedures used in data integration must minimize any potential impact on privacy and confidentiality.
Principle 7: Transparency	Statistical data integration will be conducted in an open and accountable way.

Source: High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes, February 3, 2010, www.nss.gov.au/nss/home.NSF.

In addition to ensuring compliance with the seven principles for data integration, the Statistical Integration Committee is responsible for assessing the risk associated with each data integration project via the "Accreditation Process." This process for assessing the risk associated with data integration is based on eight dimensions summarized in Table 3. Projects are then rated according to whether they pose a very high, moderate or low risk. Only one of the eight dimensions below need be considered high risk for the entire project to be rated high risk.

12. Among the members of this committee on the integration of Australian data are most departments of the Australian government, including the Bureau of Statistics, the Department of Defence, the Department of Foreign Affairs, the Department of Treasury and health agencies.

Table 3
The eight dimensions of risk associated with data integration involving Commonwealth data

Dimension 1: Sensitivity	For example, projects dealing with sensitive subjects such as religious belief, health of individuals or crime statistics.
Dimension 2: Size	This refers to the number of observations and variables in the database. The wider the range of information available, the higher the level of risk.
Dimension 3: Technical complexity	The level of risk rises as the project becomes more complex. For example, the matching process may become more complex if there is missing information or duplication of data.
Dimension 4: Managerial complexity	The risk level rises when the data to be integrated come from multiple sources or different organizations, sectors or jurisdictions. For example, different organizations may have overlaps with respect to needs or levels of access to linked data. Also, the movement of databases between organizations increases the level of risk.
Dimension 5: Duration of project	The risk level rises when the linked database must be updated and archived over a long period. Conversely, the risk level decreases if the linked database is destroyed at the end of the project.
Dimension 6: Nature of data collection	The risk level depends on the authorization contained in the consent agreement established between the statistical agency and the respondent. If the respondent consents and is informed of how the integration concerning him/her will be used, the risk of privacy intrusion and the risks associated with confidentiality are lessened.
Dimension 7: How the data are to be linked	The risk level rises with the precision of the matching method. The deterministic matching method is associated with a higher level of risk than the probabilistic method, which is less precise.
Dimension 8: Nature of access	The risk level rises when the record linkage requires access to the record identifier. This risk increases with the number of parties involved, and it also rises if the access is between international organizations.

Source: "Data integration involving Commonwealth data for Statistical and research purpose: Governance," October 6, 2010, www.nss.gov.au/nss/home.NSF.

The *Australian Privacy Act 1988* is being revised to include provisions requiring that the entity that provides the data for matching be informed about the institutions that will have access to the matched data and the use for which these data are to be matched.

In Australia, the legislation under which the Australian Bureau of Statistics operates prohibits publication of any information making it possible to identify individual data.

4. Summary of a mini-survey of national statistical agencies

To collect uniform information on international practices in record linkage, Statistics Canada prepared and sent out a two-page questionnaire to a number of statistical agencies. This questionnaire was designed to collect very general information on business linkage practices. The questionnaire asked whether there was a formal approval process for data linkage, and it contained questions on linkage techniques, the software used, the use of linked databases and the challenges faced by the agency with respect to record linkage. Below are the main points identified and the comments received from six respondents.

Formal approval of record linkage

Question 1: Does your organization require formal approval to link databases?

Only one agency responded that it does not need formal approval to link databases. A number of agencies referred to the laws governing their agency as being the official framework under which linkage is allowed. In one case in particular, the law provides the agency with the conditions and exceptions that apply to record linkage. In this case, linkage may be performed if the statistical information obtained as a result will make it unnecessary to conduct additional statistical surveys.

In some agencies, a person or agency is given responsibility for approving the linkage. This may be a privacy commissioner, a data administrator, an internal team in charge of linkage, or in the case of Canada, a senior management council, the Executive Management Board, which includes the Chief Statistician.

Some agencies must receive the consent of the respondents who provided the data before they can perform linkages. A number of other official processes mentioned, including: informing the directors in charge; entering into agreements on services levels with the government parties involved that accept the methodology; publishing linkage projects on a national register website; and clearly show why the user needs the data.

For more information on official processes in Australia, including the general principles for data integration and assessment of linkage risks, see item III-C-5 above.

Record linkage techniques

Question 2: What linkage techniques are used by your organization?

- Deterministic linkage
- Probabilistic linkage
- Non-statistical linkage (i.e. clerical review)
- Other

All but one of the agencies uses more than one technique. Those that use deterministic techniques mentioned the use of business identifiers and tax numbers for deterministic matching. This approach is similar to the one used by Statistics Canada, which is based on enterprise identifiers. Unique identifiers are used to identify the same enterprise in all Statistics Canada surveys and administrative datasets.

One agency largely uses probabilistic linkage based on the Fellegi-Sunter method. This method is supplemented by administrative tasks that serve to establish attribution thresholds for matches. One agency said that it is currently looking for ways to improve its clerical review process.

Software applications

Question 3: Which software applications are used by your organization to link databases?

The most often mentioned software used for linkage was SAS in SQL language, followed by applications developed internally. Other applications used were Oracle, SPSS, QualityStage, Febrl V0.3 and General Clerical Reviewer (GCR). One agency is currently exploring the use of RELAIS, FRIL and G-Link.

Use of linked databases

Question 4: How are linked databases used?

- To support research
- To support program evaluation (as one of several entities and not as the only one)
- To provide descriptive tabulation to clients
- To provide public use microdata files (PUMFs)
- To support the design, maintenance or evaluation of ongoing data collection
- For publications produced by your organization
- Other uses, please specify

All the agencies reported that linked databases are used for the following purposes: to support research and support the design, maintenance or evaluation of ongoing data collection processes. All but one of the organizations said that linked databases are used for publications produced by their organization. Two agencies responded they were using linked databases for all the purposes listed. The least common use was to provide PUMFs. One agency reported that it uses linked databases for another purpose, namely to reduce the response burden in collecting survey data.

Record linkage challenges

Question 5: What are the main record linkage challenges and issues faced by your organization?

- Confidentiality
- Data access to external researchers
- The linkage of administrative data
- The linkage of survey data
- Issues related to longitudinality
- Other, please specify

Half of the organizations responded that they faced all the issues and challenges in the above list. Linkage of administrative data is a challenge and an issue for all the agencies. Other challenges and issues identified are the linkage of anonymous data, evaluation of linkage quality, compliance with international (i.e., European) legislation, the nature of the data sources (more specifically conceptual differences, timeliness, frequency, delays in transmission of data and the processing requirements for data suppliers) and the efficient dissemination of linked datasets.

Quality of linked databases

Question 6: Please describe in few words your organization's approach to determining the quality of your linked databases.

The agencies reported different ways of determining the quality of the linked databases. The responses to this question have been grouped according to theme:

Quality evaluation by external researchers

- The quality of the linked data is evaluated by the external researchers who use them.
- In both cases, it is up to researchers who use either data or metadata to determine whether the data are suitable (of sufficient quality) for the specific context of their work.
- The agency co-operates with researchers to verify the quality of their data before the linkage is performed. Once linked, the data are regularly checked to identify changes over time.
- Linkage of supplementary data may take place within the agency in a specialized unit, namely a secured environment where authorized internal and external researchers can conduct specialized analyses on different datasets, including data linkage when necessary.

Matching rates

- The rate for cases that could not be matched is an important quality indicator.
- The error rate for probabilistic record linkage is provided.

Quality indicators for data sources

- Survey data and administrative data used as official statistics must conform to quality standards. Other administrative data are linked and provided to users on an "as is" basis.

Comparison of linkages with historical data and other data

- Consistency with other data collection processes (internal and external) and adherence to statistical frameworks;
- Results from separate linkages are compared for purposes of uniformity. The results obtained at the meso and macro levels are compared with other relevant results.

A comprehensive approach to probabilistic linkage

One agency uses a very comprehensive approach to probabilistic linkage. The details of the determination of the quality of the linkage are described below:

- The type of variables used in linkage serves as a quality indicator: a “Gold Standard linkage” uses names and addresses, a “Silver Standard linkage” uses encrypted names and addresses, and a “Bronze Standard linkage” uses neither names nor addresses but instead uses statistical geography.
- Two types of comparisons are used to evaluate linkage quality. The first approach is a comparison between historical data and current linkages. The second approach is to compare “Bronze Standard linkages” and other lower quality linkages with “Gold Standard linkage” where possible. This approach assumes that the “Gold Standard linkage” file does not contain missing or false matches (and is therefore of better quality). The information drawn from clerical review is used to determine the quality of the linkage, as follows:
    - The quality of matches, as regards agreement on the fields used for linkage
    - The number of matches examined
    - The distribution of match weight values
    - The number of rejected or false matches
- The probability of missing or false matches is evaluated using the number of other possible linkage solutions and the information contained in the *m* and *u* probabilities under the Fellegi-Sunter rule.
- The information on which the records converge or diverge, that is, the proportion of matches that are entirely in agreement on fields used for linkage such as date of birth, sex, small geographic area, Aboriginal status, etc.
- The number of matches that have contradictory or illogical values on fields used for linkage such as age, date of birth, sex, number of children born, country of birth, religion, language spoken and ancestry.
- The quality of the blocking and matching strategy is evaluated on the basis of the successive, complete and strategic record linkage opportunities that it provides and its impartiality for dissemination purposes.

5. Conclusion

This report has found that statistical agencies that were contacted are engaged in business data linkage activities using similar methodological approaches. However, the rules of access and the approval process for record linkages varies from one statistical agency to another, in particular as concerns the requirement for respondents’ consent (some countries require the consent and others don’t require the consent to proceed with the data linkage).

References

Antoni, M. and Seth, S. (2012). **"**ALWA-ADIAB- Linked Individual Survey and Administrative Data for Substantive and Methodological Research''. Schmollers Jahrbuch, *Journal of Applied Social Science Studies, vol. 132(1), pp.141-146*.

Antoni, M. (2011). "Linking survey data with administrative employment data: The case of the German ALWA survey". Working paper [http://www.cros-portal.eu/content/linking-survey-data-administrative-employment-data-case-german-alwa-survey-manfred-antoni]

Baldwin, J. R., Dupuy, R, and Penner, W. (1992). "Development of longitudinal panel data from business register: the Canadian Experience", *Statistical Journal of the United Nations, vol.9, pp.289-303.*

Baxer, R., Christen, P., and Churches, T. (2003). ''A Comparison of fast blocking methods for Record Linkage'' In ACM SIGKDD'03 workshop on Data Cleaning, Record Linkage and Object Consideration, Washington DC., pp. 25-27.

Beste, J. (2011). Selektivitütsprozesse bei der Verknüpfung von Befragungs- mit Prozessdaten. Record Linkage mit Daten des Panels "Arbeitsmarkt und soziale Sicherung" und administrative Data der Bundesagentur für Arbeit FDZ Methodenreport 09/2011 (DE).

Christen, P. (2012a). "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication". *IEEE Transactions on Knowledge and Data Engineering, vol. 24, pp.1537-1555*.

Christen, P., (2012b). "Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection", Springer.

Cibella, N., Scannapieco, M., Tosco, L., Tuoto, T. and Valentino, L. (2012). "Record Linkage with RELAIS: Experiences and Challenges." *Estadistica Espanola, vol.54 (179), pp. 311-328.*

Eurostat, (2013). "ESSnet projects, 2013 assessment report". Methodologies and Working Papers, Cat. No KS-RA-12-019-EN-N; ISBN 978-92-79-29622-2.

Fellegi, I. P. and Sunter, A. B. (1969). "A Theory for Record Linkage" *Journal of the American Statistical Association, vol.64, pp. 1183-1210.*

Fellegi, I. P. (1999). **"**Record Linkage and Public Policy: A Dynamic Evolution" in Record Linkage Techniques 1997, Washington, DC: National Academy Press, pp, 3-12.

Guiver, T., (2011), "Sampling-Based Clerical Review Methods in Probabilistic Linking", Methodology Research Papers, Cat. no.1351.0.55.034, Australian Bureau of Statistics, Canberra.

Hagsten, E., Polder, M., Bartelsman, E., Awano, G and Kotnik, P. (2012). "ESSnet on Linking of Microdata on ICT Usage". November 2012, Final report. Statistiska centralbyran. Statistics Sweden.

Hartmann, J. and Krug, G. (2009). "Verknüpfung von personenbezogenen Prozess- und Befragungsdaten-Selektivität durch fehlende Zustimmung der Belfragen?" In *Zeitschrift fur Arbeitsmarkforschung, vol. 42(2), pp. 121-139.*

Herzog, T., Scheuren, F. and Winkler, W. (2007). Data Quality and Record Linkage Techniques. Springer.

Jonas, J., and Harper, J. (2006). "Effective counterterrorism and Limited Role of Perspective Data Mining" *Policy Analysis, vol. 584, pp. 1-11.*

Konold, M. and Assainato (2009). **"**Matching Business Data from Different Sources: The case of the KombiFiD-Project in Germany", Conference 'New Techniques and Technologies for Statistics (NTTS 2009), Brussels.

Larsen, M. D. (2006). **"**Record Linkage, Nondisclosure, Counterterrorism and Statistics". SSC Annual Meeting.

Newcombe, H., Kennedy, J., Axford, S., James, A. (1959). "Automatic Linkage of Vital Records." *Science, vol. 130(3381), pp. 954-959.*

Newcombe, H. and Kennedy, J. (1962). "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information" *Communications of the ACM vol. 5(11), pp. 563-566.*

Newcombe, H., Fair, M. E. and Lalonde, P. (1992). "The Use of Names for" Linking Personal Records'' *Journal of the American Statistical Association, vol. 87, pp. 1193-1208.*

OECD, (2012). "Unleashing the potential of business microdata. The ESSlimit project and beyond: international cooperation to produce new indicators and analyses". Working Party Indicators for the Information Society, DSTI/ICCP/IIS (2012)3.

Rollin, A-M. (2013). Developing a longitudinal structure for the National Accounts Longitudinal Microdata File (NALMF). Proceedings of Statistics Canada Symposium 2013.

Schnell, R., Bachteler, T. and Bender, S. (2004). "A Toolbox for Record Linkage'', *Australian Journal of Statistics, vol. 33(1&2), pp. 125-133.*

Statistics Canada (1985). Statistics Canada Act (R.S.C., 1985, c. s-19.: Web site link: (http://laws-lois.justice.gc.ca/eng/acts/S-19/FullText.html)

Statistics Canada (2011). Directive on record linkage: Web site link: (http://www.statcan.gc.ca/eng/record/policy4-1)

Statistics New Zealand (2012a). "Data Integration Policy". Wellington: Statistics New Zealand ISBN 978-0-478-37787-3 (online)

Statistics New Zealand (2012b). "Data Integration Manual". Wellington: Statistics New Zealand. ISBN 0-478-26971-4 (online)

Statistics New Zealand (2012c). "Integrated Data Infrastructure extension: Privacy impact assessment". Wellington: Statistics New Zealand. ISBN 978-0-478-40840-9 (online)

Su, W., Wang, J. and Lochovsky, F. H. (2009). "Record Matching Over Query Results from Multiple Web Database" *IEEE Transactions on Knowledge and Data Engineering, vol. 22, pp. 579-589*.

Vogel, A. and Wagner, J. (2012). "The Quality of the KombiFiD- Sample of Business Services Enterprises: Evidence from a Replication Study". Working Paper Series in Economics N.226. University of Lüneburg.

Winkler W. E. (1999). "The State of Record Linkage and Current Research Problems". Statistical Research Division, U.S. Bureau of the Census. Washington, DC.

Appendix A

Definition of privacy and confidentiality

Privacy is the right to be left alone, to be free from interference, from surveillance and from intrusions. When choosing to "invade" a person's privacy, governments have obligations with respect to the collection, use, disclosure, and retention of personal information. Privacy generally refers to information about individual persons.

Confidentiality refers to a protection not to release identifiable information about an individual (such as a person, business or organization). It implies a "trust" relationship between the supplier of the information and the organization collecting it; this relationship is built on the assurance that the information will not be disclosed without the individual's permission or without due legal authority.