Statistics New Zealand
TATAURANGA AOTEAROA

Data integration manual: 2nd edition

New Zealand Government

 **Crown copyright ©**
This work is licensed under the Creative Commons Attribution 3.0 New Zealand licence. You are free to copy, distribute, and adapt the work, as long as you attribute the work to Statistics NZ and abide by the other licence terms. Please note you may not use any departmental or governmental emblem, logo, or coat of arms in any way that infringes any provision of the Flags, Emblems, and Names Protection Act 1981. Use the wording 'Statistics New Zealand' in your attribution, not the Statistics NZ logo.

Liability
While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, Statistics New Zealand gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

Citation
Statistics New Zealand (2013). *Data integration manual: 2nd edition.* Available from www.stats.govt.nz.

ISBN 978-0-478-42948-0 (online)

Published in March 2015 by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact
Statistics New Zealand Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4610
www.stats.govt.nz

Contents

List of tables and figures

List of tables

List of figures

1 Purpose and summary

The *Data integration manual* 2nd edition provides a guide to data integration at Statistics New Zealand. The manual was written by Statistics NZ staff, following our involvement in several large inter-agency data integration projects.

The manual's purpose is to guide best practice and share the insights gained from our experience. The manual will assist agencies collaborating with Statistics NZ, and others interested in data integration, to understand the basic concepts, theory, and processes involved in data integration, as well as providing practical advice.

1.1 Manual's structure

The manual begins with an introduction that describes what data integration is and why it is carried out, and outlines the key steps involved.

Chapter 3 introduces the legal environment and Statistics NZ policy on data integration.

Chapter 4 describes operational aspects of Statistics NZ data integration projects.

The remaining chapters focus on technical aspects of the linkage itself: the data preparation needed before record linkage can be undertaken, the statistical theory of record linkage, and how to practically implement record linkage techniques.

2 Introduction to data integration

In this chapter we introduce data integration, describe why it is carried out, and present a brief history of data integration at Statistics NZ.

2.1 Key points about data integration

2.1.1 What data integration is

Data integration is defined broadly as combining data from different sources about the same or a similar individual or unit. This definition includes linkages between survey and administrative data, as well as between data from two or more administrative sources. Another application of data integration theory is in identifying records on a single file that belong to the same individual or unit.

Other terms used to describe the data integration process include ‘record linkage’ and ‘data matching’.

2.1.2 Data integration levels

When integration occurs at the micro level, information on one individual (unit) is linked to:

1. a different set of information on the same person (unit)
2. information on an individual (unit) with the same characteristics.

At the macro level, collective statistics on a group of people or a region can be compared and used together.

The main focus of this manual is micro-level data integration of the first type – that is, linking records that are likely to belong to the same individual or unit.

2.1.3 Data integration’s role

Data integration’s role in helping to produce an effective official statistical system is becoming increasingly apparent. By bringing together information from different sources we can answer a broader range of questions. Through integration it becomes possible to examine underlying relationships between various aspects of society, thus improving our knowledge and understanding about a particular subject.

2.1.4 Why we integrate data

Linking administrative data from different sectors creates a valuable source of information for statistical and research purposes because researchers can examine relationships that previously could not have been considered. Sometimes there are other methods of investigating relationships of particular interest, for example conducting a survey. However, data integration offers a less time-consuming and less costly alternative, although it still requires a significant level of time and resource. Data integration also has the advantage of reducing respondent burden by making more effective use of existing data sources.

2.1.5 Legal and policy considerations

Data integration raises a range of legal and policy issues, some of which are complex to resolve. Chapter 3 has more detail.

2.2 Key data integration concepts

2.2.1 Integration for statistical and administrative purposes

When we link data for statistical purposes, individuals (or units) are identified only to enable the link to be made. When the linkage is complete, the individual's (or unit's) identity is no longer of any statistical interest. We use the linked dataset to report statistical findings about the population or sub-populations.

In contrast, when data is linked for administrative purposes, individuals are identified not only to enable the link to be made, but also for administrative use subsequent to the linkage. This may sometimes result in adverse action (eg prosecution) being taken against individuals.

Statistics NZ undertakes data integration only for statistical purposes.

2.2.2 Exact linkage and probabilistic linkage

There are two key methods for linking records. Exact linkage involves using a unique identifier (eg a tax number, passport number, or driver's licence number) that is present on both files to link records. It is the most efficient way to link datasets, and is easy to carry out using general-purpose statistical software such as SAS.

Where a unique identifier is not available, or is not of sufficient quality or coverage to be relied on alone, we employ probabilistic linkage. This involves using other variables that are common to both files (eg names, addresses, date of birth, and sex) to find records that are likely to belong to the same person. Probabilistic linking is more complex than exact linkage – sophisticated data integration software is required to achieve high-quality results.

Note: Different terms are used to describe types of linkage, including 'probabilistic', 'statistical', 'stochastic', and 'demographic'. In the literature, different authors use these to communicate different concepts. Sometimes they are used interchangeably within a paper. We use the term 'probabilistic' as defined above throughout this manual.

2.2.3 Quality assessment

Either linking method can result in two types of errors: false positive links and false negative links. A false positive link is where two records are linked together, when in reality they are not the same person or unit. A false negative link is where two records are not linked together, when they do in fact belong to the same person or unit. Generally there is a trade-off between the two types of errors since, for example, reducing the rate of false positives may increase the rate of false negatives. Thus it is important to consider the consequences of each type of error and to determine whether one is more critical than the other.

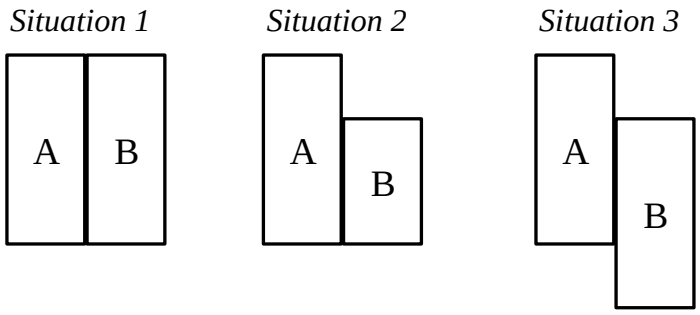
Researchers should assess the size of each of these sources of linkage error as part of the integration, and make the results available. Analysis of an integrated dataset should take into account the possible effects of the linkage error.

Further details on quality assessment are in section 6.4.

2.2.4 Data integration scenarios

There are different ways in which two datasets being integrated relate to each other.

Figure 1
Data integration scenarios



Situation 1

Every individual on dataset A is also on dataset B and vice versa. For example, dataset A might consist of addresses while dataset B contains rates information for each address.

Situation 2

Every individual on dataset B is on dataset A but some individuals who appear on dataset A are not on dataset B. For example, dataset A could be student enrolments and dataset B could be information for the students who have student loans.

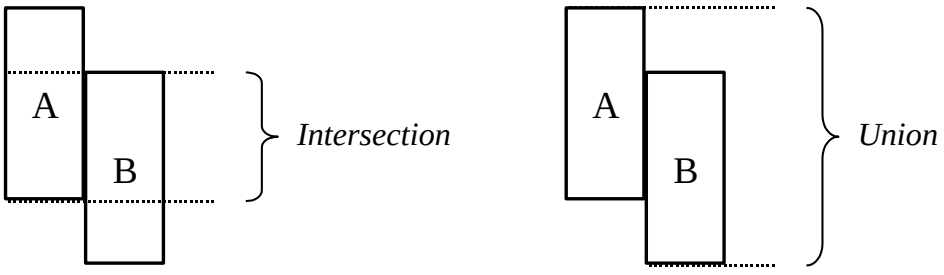
Situation 3

Some individuals appear on both dataset A and dataset B. However, other individuals will appear on only one dataset or the other. For example, dataset A might be Accident Compensation Corporation (ACC) clients, while dataset B could be people admitted to hospital.

Note: these are 'theoretical' relationships between pairs of files. Real life is rarely that perfect. For example, in situation 1 there could be duplication and omissions within the files, and timing differences between the two files, which mean they do not have 100 percent overlap.

There are also different desired results from a pair of integrated datasets: the union or the intersection.

Figure 2a
Capturing different target populations using an intersection or union to integrate two datasets

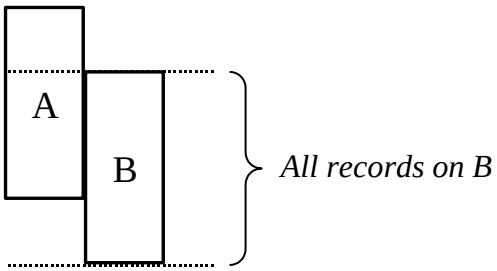


For example, if dataset A was ACC claims and dataset B was hospitalisations for injury, the **intersection** would be of interest if statistics were wanted on the number of ACC claimants admitted to hospital as a result of their injury. The **union** would be of interest if

statistics were wanted on the total number of injuries, without double-counting injuries represented in both datasets.

Sometimes a different combination of records may be required – for example, “all records on B, with information added from A if it is available”.

Figure 2b
Integration based on the population in one dataset



Continuing the previous example, this combination may be of interest if hospitalisation costs were to be combined with costs to ACC, for the population of ACC claims.

2.3 Data integration at Statistics NZ and elsewhere

2.3.1 The emergence of data integration

The most significant early contributions to record linkage came in the 1950s, in the field of medical research (Gill, 2001). Two of the most influential early papers were by Newcombe et al (1959) and Fellegi and Sunter (1969).

2.3.2 Data integration at Statistics NZ

Data integration has been used in a variety of ways at Statistics NZ, beginning in the 1990s, and gaining momentum in recent years. The earliest uses were mostly Statistics NZ driven, but more recent advances have come from external interest in integrating datasets collected from different agencies for unrelated purposes.

In 1997, the Government directed that “where datasets are integrated across agencies from information collected for unrelated purposes, Statistics NZ should be custodian of these datasets in order to ensure public confidence in the protection of individual records” (Cabinet minutes,1997). In the same meeting, it was agreed that Statistics NZ should “carry out a feasibility study into the costs and benefits of integrating cross-sectoral administrative data to produce new social statistics”.

This feasibility study was subsequently carried out, including a trial integration of Inland Revenue income information with beneficiary data from the Department of Social Welfare (now the Ministry of Social Development). The feasibility, costs, barriers, and benefits of that integration were assessed. The resulting final report (Statistics NZ, 1998) laid the foundation for future data integration projects.

The Statistics New Zealand Statement of Intent: 2012–17 (2013) states:

> We will promote the effective use of administrative records as a key data source. This, and a coordinated approach to reducing respondent burden across the system, will ensure that the Official Statistics System functions effectively and efficiently to deliver better value for money.

We have used data integration at Statistics NZ to create survey frames, supplement survey data, and produce new datasets.

See data integration for detail on our use of data integration and a list of data integration projects.

2.4 Key steps in a data integration project

Through experience gained from data integration projects, we have identified a number of necessary steps for a successful outcome. The list below is a high-level summary of what must be addressed. Each step is fundamental, and none is trivial. The importance of clear and well-defined objectives cannot be overemphasised.

The objectives will inform your decisions at every other step of the project, from gaining initial approval to carry out the project (under Statistics NZ policy requirements), to assessing whether the integrated data can support outputs that are fit for purpose.

Most data integration projects go through a feasibility or development stage, where the steps are investigated and carried out to a greater or lesser extent. The results of the feasibility study determine whether to develop full production systems.

Many of the steps have parallels in producing statistics from survey data. Differences occur where the nature of the data collection is different, and because there is an additional step of linking two or more data sources.

Key steps:

- establish need
- develop clearly defined objectives
- address legal, policy, privacy, and security issues
- consider provision of access to microdata and confidentiality of published outputs
- define governance structures and establish relationships with data providers and data users
- complete business case, privacy impact assessment, Government Statistician and agency chief executive sign-off, support from agencies supplying data
- obtain source data and gain a thorough understanding of data sources
- decide how to do the linking
- define and build information technology data storage and processing requirements
- carry out the linking
- validate the linking and provide quality measures
- open up access to linked data (if required)
- carry out the analysis and disseminate results.

It is important to engage with the data supplier (whether an internal unit or external agency) during the entire process to ensure you are making the correct decisions at each step. Data suppliers generally have a great deal of useful knowledge about their data, which may not all be easily available in formal documentation.

These aspects are discussed in more detail in the remainder of the manual.

3 Legal and policy considerations

All data integration projects are subject to legislation, codes of practice, protocols and policies. This chapter gives an overview of the relevant guidelines and offers a practical guide to how we apply them at Statistics NZ. Analysts from other agencies will find the information helpful for their own data integration projects.

Staff working on a data integration project should be aware of the various policies and legislative provisions that affect their project. Some of these, such as the Statistics Act 1975, apply to much of Statistics NZ's business.

Others, such as our data integration policy, are more specific to data integration projects. Sometimes it can be difficult to interpret legislation and apply it in a practical way to a particular situation. The first project teams to work on data integration projects debated and worked through issues, which often established precedents.

Questions can often be resolved by discussion with experienced colleagues, the project manager, and stakeholders. Sometimes you can seek advice from external parties such as a reference group or the Privacy Commissioner (see section 3.6.2).

The following sections outline relevant documents and processes.

3.1 The Statistics Act 1975

Statistics NZ operates under the authority of the Statistics Act 1975. The Act provides the framework for producing official statistics in New Zealand. It covers statistics collected in surveys of households and businesses, as well as those derived from administrative records of central and local government agencies. It covers official statistics produced by Statistics NZ and by other government agencies.

The Statistics Act 1975 does not specifically refer to data integration, so we need to interpret its provisions in the data integration context. The following points are relevant to data integration.

- Section 3[1] from the Act
    - Official statistics shall be collected to provide information required by the Executive Government of New Zealand, Government Departments, local authorities, and businesses for the purpose of making policy decisions, and to facilitate the appreciation of economic, social, demographic, and other matters of interest to the said Government, Government Departments, local authorities, businesses, and to the general public
- Official statistics means statistics derived by government departments from:
    - statistical surveys as defined in this section
    - administrative and registration records and other forms and papers the statistical analysis of which are published regularly, or are planned to be published regularly, or could reasonably be published regularly [section 2]
- Information may be required of any person in a position to provide it to enable the production of official statistics of any or all of the kinds specified [section 4]
- Independence of the Government Statistician in respect of deciding:
    - the procedures and methods employed in the provision of statistics produced by the Statistician
    - the extent, form, and timing of publication of those statistics [section 15(1)]
- Furnishing of information required [section 32]

- Security of information provided [section 37]
- Information furnished under the Act to be used only for statistical purposes [section 37(1)]
- Only employees of the department may view individual schedules [section 37(2)]
- No information from an individual schedule is to be separately published or disclosed [section 37(3)], except as authorised by the Statistics Act 1975 (the Act permits others to see information from an individual schedule, but only when it is in a form that prevents identification of the respondent concerned, and then only under strict security conditions)
- All statistical information published is to be arranged in such a manner as to prevent any particulars published from being identifiable by any person as particulars relating to any particular person or undertaking [section 37(4)].

Data obtained by Statistics NZ for integration, and all integrated datasets, are considered to be furnished under the Statistics Act 1975, and are therefore subject to the Act's provisions.

3.2 The Privacy Act 1993

The Privacy Act 1993 aims to promote and protect individual privacy. It relates to personal information (not information about businesses). Section 6, outlines 12 principles relating to the collection, storage, security, access, retention, use, and disclosure of personal information. An unpublished Statistics NZ corporate document (1999b) outlines how the Privacy Act 1993 relates to statistics.

Several of the principles allow exemption on the grounds that the information is used for statistical or research purposes and will not be published in a form that could reasonably be expected to identify the individual concerned. Regardless of these exemptions, it is important for researchers to consider the ideals expressed by the principles. Each situation must be evaluated in the light of the other privacy principles, and likely public perception of the proposed use.

The Act has no guidelines about linkage for statistical purposes. In lieu of this, we have produced a set of data integration principles and guidelines (see section 3.3).

The Privacy Act 1993 contains a chapter governing 'information matching'. This does not relate to data integration as carried out by Statistics NZ; it relates to comparing two files for the purpose of producing or verifying information that may be used for to take adverse action against an identifiable individual.

Our data integration policy states that "All use of integrated data must comply with the requirements of the Statistics Act 1975. All information collected under the authority of the Statistics Act 1975 (including data supplied for integration) is to be used for statistical or research purposes only." This means data cannot be used for regulatory purposes.

Researchers can communicate with the Office of the Privacy Commissioner on privacy issues in data integration projects (see section 3.5.2).

3.2.1 Using unique identifiers

Principle 12 of section 6 the Privacy Act 1993 regarding unique identifiers (UIDs) has particular relevance for data integration projects; it was discussed at length with the Office of the Privacy Commissioner.

Principle 12 states (under the heading “Unique identifiers”):

1. An agency shall not assign a unique identifier to an individual unless the assignment of that identifier is necessary to enable the agency to carry out any one or more of its functions efficiently.
2. An agency shall not assign to an individual a unique identifier that, to that agency’s knowledge, has been assigned to that individual by another agency, unless those 2 agencies are associated persons within the meaning of subpart YB of the Income Tax Act 2007.
3. An agency that assigns unique identifiers to individuals shall take all reasonable steps to ensure that unique identifiers are assigned only to individuals whose identity is clearly established.
4. An agency shall not require an individual to disclose any unique identifier assigned to that individual unless the disclosure is for one of the purposes in connection with which that unique identifier was assigned or for a purpose that is directly related to one of those purposes.

A UID is defined in section 2 of the Privacy Act 1993 as follows:

> Unique identifier means an identifier –
>
> 1. That is assigned to an individual by an agency for the purposes of the operations of the agency; and
> 2. That uniquely identifies that individual in relation to that agency – but, for the avoidance of doubt, does not include an individual’s name used to identify that individual.

Not all identifiers used by other departments are UIDs in terms of the above. However, in practical terms, this privacy principle affects Statistics NZ’s ability to use them in data integration projects, particularly in retaining the UIDs over time.

Our current practice is to remove the original UIDs from the dataset when it first arrives and replace them with Statistics NZ’s UIDs. This new identifier holds the same properties as the original identifier – two different datasets containing the same original UID variable have the same Statistics NZ UID. The final datasets contain (anonymised) unit records identified only through a Statistics NZ unique reference (Statistics NZ, 2012b).

3.3 Statistics NZ’s data integration policy

Our data integration policy’s purpose is to ensure we minimise risks to personal privacy when integrating personal data.

In the past, we relied on surveys and censuses to produce statistics. These are costly to run, and burdensome for respondents. Therefore, we aim to increasingly use administrative data and integrate existing datasets to produce our statistics and research. This is part of our organisation-wide programme of change, Statistics 2020 Te Kāpehu Whetū, detailed in our Statement of Intent 2012–17 (Budget 2013).

While data integration can lessen the burden on respondents and lower production costs, it also raises real and perceived privacy risks. We need to consider these when determining whether to integrate datasets.

The policy aims to assure government, respondents, and other interested parties that we take privacy concerns seriously, especially when integrating data that uses information originally provided for different purposes. Staff must use the policy (and supporting guidelines) when deciding whether to integrate datasets.

The policy states:

> Statistics New Zealand will integrate data from separate sources when necessary to ensure we can efficiently produce the information New Zealand needs to grow and prosper.

Principles for integrating data

We consider integrating data to produce official statistics and related research only when all four of the following principles are met.

- Principle 1: The public benefits of integration 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 respondents that prevents such action.

All data integration activities must also comply with Statistics NZ's Security Policies Framework and our methodological standards for confidentiality, as well as the Statistics Act 1975, the Privacy Act 1993, the Public Records Act 2005, and any other legislation relevant to the source datasets.

Anyone embarking on a data integration project must read the data integration policy.

3.4 Codes of practice

The Privacy Commissioner writes the following about codes of practice:

> The Privacy Act 1993 gives the Privacy Commissioner the power to issue codes of practice that become part of the law. These codes may modify the operation of the Act for specific industries, agencies, activities or types of personal information. Codes often modify one or more of the information privacy principles to take account of special circumstances which affect a class of agencies (eg credit reporters) or a class of information (eg health information). The rules established by a code may be more stringent or less stringent than the principles they replace.

Although there is no specific code of practice for data integration, the Privacy Commissioner has issued other codes of practice.

Examples are:

- Health Information Privacy Code 1994
- Post-Compulsory Education Unique Identifier Code 2001 (revoked 2009).

Data integration project researchers should be aware of the codes of practice that relate to their project and, in particular, how these affect the agency supplying the data.

3.5 Data integration business case

An approved data integration business case ensures the aims and boundaries of the integration project are defined clearly. This section sets out how we ensure policies and legislation have been complied with and describes the approvals that need to be obtained before a project begins.

It is important that before data is actually acquired and linked, researchers have appropriate discussions and gain approvals. Our policy requires that you complete either

a privacy impact assessment or a privacy benefit/risk analysis and submit it to the Government Statistician, in the former case, or a deputy Government Statistician in the latter case.

Both processes require researchers to analyse and document the benefits and risks of data integration, and to consult stakeholders. Ideally, these analyses are completed as part of an overall proposal for data integration, such as a business case.

The policy outlines the key steps to get a data integration project approved. See section 1.4 of this manual. In summary, these are:

- establish need
- prepare the privacy impact assessment or privacy benefit/risk analysis
- consult stakeholders
- obtain approval.

The policy also lists stakeholders who must be consulted, both internal and external.

These approval processes also apply to pilot or feasibility studies.

Once approved, the project must be included in the list of data integration projects on Statistics NZ's website.

3.5.1 Privacy impact assessment

A privacy impact assessment (PIA) or a privacy benefit/risk analysis must be completed before the integration project is approved. PIAs are not unique to Statistics NZ – they are used in many situations where risks to privacy arise. Those for our data integration projects systematically evaluate the privacy risks associated with the project and state how these risks will be mitigated.

PIAs for Statistics NZ data integration projects are available online. A recent example is the Privacy impact assessment for the serious injury outcome indicators (Statistics NZ, 2014).

A privacy benefit/risk analysis is required when the potential for privacy concerns is **low**. The potential is low in the following circumstances:

- when we integrate two or more datasets collected under the Statistics Act 1975 (eg the General Social Survey probabilistically linked to the Household Labour Force Survey)
- when we integrate samples of data from the Census of Population and Dwellings with other Statistics NZ data (eg a random sample of census data probabilistically linked to the General Social Survey).

A PIA is required when the potential for privacy concerns is **moderate or high**. It is moderate or high in the following circumstances:

- when one or more of the datasets we integrate was not collected under the Statistics Act 1975 – that is, external datasets such as administrative, survey, or commercial datasets from other agencies.
- when we integrate data that comes from all or most individuals and/or dwellings in the census – this includes integrating different census years.

PIAs are usually compiled by the project manager, with contribution from team members and policy staff. However, it is important for project staff to be aware of privacy issues relating to their project and how they are to be managed. PIAs for established projects are a useful way for you to get to grips with privacy issues

While not a formal part of a PIA, any issues regarding appropriate protection for business information should also be considered.

3.5.2 Consultation with other agencies

Office of the Privacy Commissioner

It is sometimes necessary to consult with the Office of the Privacy Commissioner (OPC) regarding proposed data integration projects. The office works to develop and promote a culture in which personal information is protected and respected. Staff provide advice to agencies around privacy matters and applying the Privacy Act 1993.

Statistics NZ seeks the commissioner's advice as to whether the proposed approach raises any concerns, then works closely with him or her to determine the most appropriate solutions.

If a Statistics NZ project manager has to decide between continuing with an approach that precedent and best current understanding can accept, or arguing for change in what the OPC is comfortable with – to get a better outcome – the first option is easier and more certain. The OPC is not resourced for quick responses to approaches, because it is not their core business. It is up to Statistics NZ to comply with the Privacy Act 1993 and, if necessary, to seek professional advice on issues.

Other agencies

Integration projects must recognise the direct interests of stakeholders and take their concerns into account in deciding to integrate. Researchers must consult providers of source datasets, and groups that represent the interests of people whose information is being integrated.

3.6 Statistics NZ's confidentiality protocol

Statistics NZ's confidentiality protocol (1999a) also affects data integration projects. The protocol includes sections on restricting information use to statistical purposes, protecting confidential information, and rules to avoid disclosing confidential information in outputs and microdata.

Although this protocol is not specific to integrated data, its provisions apply to integrated data and researchers should be aware of its content. There is greater risk of disclosure from integrated datasets and therefore extra care is required to protect the data.

3.7 Statistics NZ's microdata access protocols

Integrated datasets potentially pose more disclosure risk so any access to microdata needs careful consideration. The final decision about providing microdata access is made by the Government Statistician with the guidance of the relevant subject matter area and the Chief Methodologist.

In some cases this decision is only made after consulting with data providers. This consultation is needed as the data is not collected under the Statistics Act. Tax data is an example. Consider the following when deciding if microdata access should be given.

- the legislation under which the data was collected
- Statistics NZ's microdata access protocols
- the data integration business case (Statistics NZ, 2005a) (eg allowed uses of integrated data)
- any agreements made with data suppliers.

The protocols are based on six principles, which provide guidance on the purpose, methods, and conditions of access.

4 Operational aspects of a Statistics NZ data integration project

Individual data integration projects vary greatly in the data sources and methods used. This chapter outlines operational aspects that are common to most data integration projects we carry out.

4.1 Early stages of a project

A data integration project begins with the approval processes outlined in section 3.5. We recommend that a pilot study be undertaken to assess the feasibility of the project.

Once approved, a successful data integration proposal moves into the project's usual phases – developing project initiation documents, and bringing together a project team. Inter-agency relationships and support from interest groups should be established as early as possible.

4.2 Other relationships

Statistics NZ's data integration projects usually involve us receiving data from external agencies, and producing outputs of wide interest outside the organisation. As well as a project team and internal Statistics NZ governance, each data integration project is likely to have critical relationships with external groups.

The nature of these relationships differs, depending on how the project is structured.

The quality of the relationship with the data providers and users is an important contributor to the success and efficiency of any data integration project.

4.3 Obtaining and keeping external data safe

4.3.1 Data extraction

A business case submitted for Statistics NZ's approval needs to detail the data it proposes to integrate, including a list of the variables from each data source. If researchers cannot determine these details, then a business case for a pilot study should be produced instead. A pilot study will aim to determine which variables are needed for integration and whether the integrated dataset is suitable for achieving the statistical objectives of the project.

The specifications for the data extract should use the source data's table and variable names. Specific inclusions and exclusions must be clear and unambiguous.

The data received should be examined to ensure it complies with expectations (see section 5.3).

4.3.2 Data transfer, storage, security, and internal access controls

In keeping with Statistics NZ's core security value, data integration projects must provide adequate protection to the data.

Our data integration policy provides guidelines to ensure protection of the linked data. Information about individual records cannot be sent to data providers. Names and addresses can only be retained in an integrated dataset for a limited period if this was

approved in the data integration business case. Unique identifiers assigned by an external agency must be removed immediately. Moreover, all data integration projects must have exclusive use of their own physical server(s) for processing and exclusive use of their own physical disk(s) for storage, and be accessible only to the smallest practical number of Statistics NZ employees.

A privacy impact assessment needs to consider security issues. Our framework for security policies covers the requirements for keeping data secure. The project team should actively contribute to maintaining high standards of security.

4.4 Documentation and quality assurance

See the appendix for detailed guidelines about writing and peer reviewing technical descriptions for record linkage projects.

4.4.1 Documenting record linkage methodology

Any linking exercise should be accompanied by full documentation of the method used. This is the ‘technical description’ of the linking methodology and has two main uses:

- to allow peer review of the methodology
- to provide a record of what has been done for the future.

You need to record full details of the linking method and results. They provide the formal documentation of what was done, both for future linking with the same data sources, and as examples for other linking projects.

4.4.2 Reviewing record linkage methodology

A peer review is needed to confirm that a sound job has been done. The peer review should be a review of the process, not a repeat of the linking.

The review should be done before the linked data is handed to clients so that any methodological improvements suggested by the reviewer can be carried out. This might mean a two-stage process, where the first results are essentially a trial. The linking method and results are reported and reviewed, then any modifications you carried out, and lastly the final linked file that goes to clients. In practice, this may not be possible, in which case improvements can be noted for future.

Documenting and peer reviewing the linking methodology should be part of project planning; enough time and resources needs to be allowed for them.

4.4.3 Supporting documentation

All output data should be supported by adequate metadata – information that enables a user to come to an appropriate understanding of the data. For example, information about:

- the context of the source data
- data processing at Statistics NZ
- quality indicators, including link rates
- the format of the output data, including detailed information about individual variables (a ‘data dictionary’)
- and advice on how the technical description of the linking can be obtained.

4.5 Information technology considerations

The size of an integrated dataset, its complexity, and the differing needs of official statistics and researchers all place considerable demands on the IT solution for a data integration project.

Using administrative data in data integration can result in much larger data files than for data based on sample surveys. Some Statistics NZ data integration projects (eg the Integrated Data Infrastructure) have datasets that are orders of magnitude larger than that usually processed for sample surveys. This presents challenges for data storage capacity and efficiency of updates, general processing, and retrieving information.

Integrated data is often conceptually complex in structure. A link that may be longitudinal and/or cross-sectional needs to be maintained. There may be complex relationships between the original unit-record structure and the statistical units used for analysis. For example, the student loans data includes units for loans, enrolments, individual students, and tertiary institutions.

Integrated data often has two types of uses, with possibly conflicting requirements. Official statistics outputs are aggregated and summarised data, usually tabular and standardised. In contrast, researchers often require access to microdata. The range of variables and scope of the investigation are likely to be different for each research project and unpredictable. A general indication of how the integrated data will be reported, the type of statistical outputs, and their breakdowns is needed at an early stage to ensure that appropriate IT systems can be developed.

Researchers need to carefully consider each of these features when designing IT storage and processing systems.

5 Preparing for record linkage

The actual process of doing the record linkage is only a small fraction of the overall data integration project. This chapter describes the tasks that should be carried out before starting the actual linkage.

Gill (2001) estimated that in implementing record linkage:

- 75 percent of the effort is in preparing the input files
- 5 percent of the effort is in carrying out the linkage
- 20 percent of the effort is in checking the linkage results.

This is consistent with the balance of work we experience at Statistics NZ. Adequate preparation for linkage is important. This includes investigating, obtaining, assessing, and transforming the input data.

Also, at implementation, everything collected and used for record linkage should be well documented – for ongoing maintenance and for future data users.

5.1 Gathering information about source data

Developing a thorough understanding of the source data is fundamental to obtaining meaningful results from analysis of the integrated data. Understanding new data sources is time consuming and resource intensive so the investigation can be carried out in several phases.

An initial investigation may focus on what you need to determine the data's relevance to a particular research programme – mainly the population and variable concepts and values – and be restricted to useful variables only. Use available documentation and contact with the agency to do this.

You can use information from this early stage to determine whether a data source is likely to be of sufficient quality to meet the data integration's objectives. The outcome of the initial investigation should be a 'stop/go' decision about whether the project is likely to be worthwhile.

Obtaining sample/test data to investigate before getting the entire dataset can be useful. This allows you to gain more extensive knowledge about the data structure and variables included.

More detail will be required to specify data to be transferred, to prepare files for the linkage, and to analyse the integrated data.

5.1.1 Preliminary investigation of source data

Once a data integration project is initiated, it is usually clear where the data to be linked will come from. It may be internal to Statistics NZ – such as the Business Register, the population census, or survey datasets – or from an external agency.

Using external data

The following discussion assumes data from an external agency is involved (most principles apply equally to internal datasets). It also assumes there is a well-managed relationship, with a service agreement, such as a memorandum of understanding (see section 5.2.1), either being developed or already in place.

Before requesting the datasets from the source agency or agencies, you should carry out a preliminary investigation to clarify dataset specifications for the integration. When Statistics NZ does this, the investigation involves some (or all) of the following tasks:

Collating existing departmental knowledge

If the data source has been used previously within Statistics NZ, other staff may be able to help by briefing on the data, providing written documentation, or explaining how the data was used before. Over time, it is likely we will use more administrative data. Sharing knowledge is important, both for internal efficiency and to convey professionalism to data providers.

Reviewing hard- and soft-copy information

Organisations' websites generally provide an excellent overview of the context of the data source, the motivation for collecting the data, the data collection environment, and how the data is used. More detailed information is often available – fact sheets and frequently asked questions, downloadable data collection forms, and data dictionaries.

Meeting with data providers

Meeting with the data providers helps shift knowledge from the people who work with the data on a daily basis. These meetings may involve: general briefings, the opportunity to question, viewing the electronic data storage/query system, and gathering further documentation (eg data dictionaries and data models).

5.1.2 Target population and units

Identifying a population

Understanding how the data sources relate to one another and defining the target population is important in preparing for linking. The nature of an individual dataset itself is also important when thinking about population coverage. Administrative data records are often compiled into a dataset from multiple locations (eg the National Minimum Dataset is a collection of discharge information from all public and private hospitals).

Each source data file has its own target population and actual population, which can differ. Deciding where the populations overlap is an important step in preparing for record linkage.

Target set

The target set is the theoretical population that the administrative data covers. For example, a legal requirement may define the target set. In other cases, a transaction undertaken, or a voluntary application may mean the target set is the collection of reporting units on the dataset.

Accessible set

The accessible set is the population that the administrative data includes in practice – that is, the population actually on the administrative dataset. It is only this population that results can be reported for, although it is sometimes possible to have an estimate of the undercount and bias (Statistics NZ, 2005b). The accessible set may be more than, less than, or the same as the target set. 'Coverage' is the difference between the target and accessible sets.

An integration project's purpose is to produce a new set of data, for which a target set must also be defined. This is an ideal population to make inference about by using the integrated dataset. Depending on which accessible sets are available in the source datasets, it may not be possible to produce a combined dataset that exactly covers the target set. Therefore some compromises need to be made, and a dataset utilised that does not exactly correspond to the 'accessible integrated set'.

Some data providers have unpublished research to verify the coverage of their target set. This information could be useful to estimate the link rate.

Identifying units

Another critical step in preparing for integration is to identify the reporting units on the source data files, assess whether they are consistent across the files, and to carry out any necessary transformations. However, reporting units in the source data may differ from the units of interest in the target set of the integrated dataset. Also, there may be multiple units of interest in the integrated dataset as well as differences between reporting units in the different data sources. For example, Linked Employee-Employer Data within the Integrated Data Infrastructure is used to produce statistics for businesses, jobs, and workers.

Developing the methodology to transform reporting units into the units of interest is a time-consuming and complex process. For example, in the Injury Statistics Project (see Injury information portal) the unit of integration and analysis is the 'injury'. However, the reporting unit on the New Zealand Health Information Service input file is at a lower level: 'health event'. A particular injury can have several health events (discharges).

5.1.3 Understanding the source data – metadata

Statistics NZ's Methodological standard for metadata content and the associated guidelines (Statistics NZ, 2012b) are a helpful tool for a full understanding of a data source. The metadata template is based around a quality framework for statistical outputs, where quality has six dimensions: relevance, accuracy, timeliness, accessibility, interpretability, and coherence. Brackstone (1999) discusses what these terms mean and how they interrelate.

Statistical outputs produced from administrative data, including integrated data, also need to meet data quality standards. The six dimensions are useful to assess and report on the quality of the source administrative data, and the final integrated data. The metadata template provides headings as prompts to ensure adequate information is collected in each area.

Using the template provides:

- a focus and direction for people needing to understand an administrative data source
- a structured format for recording information about the administrative data
- the information needed to assess the statistical integrity of an administrative data source
- the information needed to determine the usefulness of the administrative data source in any given context.

Field visits to data collection agencies or identifying data entry points can also help you understand the collection system.

5.1.4 Implications from the metadata

Once the source data is understood, your next task is to assess the implications for the integration project, and determine the usefulness of the variables or fields in the data. This assessment should include the following points.

- Comparison with similar fields in other datasets (eg format, ranges, classifications) – are the formats, ranges, and classifications compatible? If not, how much work is required to make them compatible? For example, sex is coded as 1 for male and 2 for female in one dataset and the other way around in another. Comparing the

metadata may highlight the need to standardise these codes before you attempt linking.

- Understand how the field came about – for example, finding many records with sex recorded as male but with undoubtedly female first names can be due to a system that has default sex as male.
- What fields could be useful for linking the files? Fields common to both files are candidates for linking. The quality of these variables needs to be assessed. See section 5.3.1 and sections 7.3.1 and 7.3.2 for more information.
- Are there any implications for the expected main outputs from the integrated dataset? That is, are there any changes in collection or maintenance practices in the source data that have implications for the quality of the integrated dataset? For example, if one output was to be a breakdown by region, was the address collected once and then never checked again? If so, and the integrated dataset is a time series, the regional breakdown may not be accurate.

5.2 Procedure for obtaining data

5.2.1 Request for supply of data

Following approval of a pilot and/or an integration project, Statistics NZ formally requests a dataset from the source agency or agencies. All data in datasets that we obtain for integration are considered to have been collected under the Statistics Act 1975 and all relevant provisions of that Act apply to the data.

Before requesting the data, an agreement document such as a service level agreement (SLA) or a memorandum of understanding (MOU) should be in place. An SLA is a formal agreement between two or more parties that seeks to achieve a mutually agreed level of services through the efforts of all the parties involved. An MOU is a formal voluntary agreement between two or more parties that seeks to achieve mutually agreed outcomes through the efforts of the parties. The only difference between the two is that an SLA is a contract, while an MOU is a voluntary agreement.

In our data integration projects, an MOU is commonly used and stipulates:

- a specification on which variables to request or the formats of the data
- security and confidentiality measures
- the frequency of data supply.

We prepare a request to obtain a dataset once the source agency and Statistics NZ have agreed on the appropriate dataset specifications to allow us to proceed with integration in the most cost-effective way. These specifications include the points above, as well as the transport mechanism, how missing data is handled, information about data quality, and responsibility for cleaning the data.

The source agency:

- is responsible for ensuring that all required variables are identified, specified, and supplied as agreed
- must provide relevant documentation about the dataset, as well as access to data experts who can assist with queries about, for example, structure and format
- must also provide information on who currently, or potentially, has access to the dataset – to help us establish confidentiality requirements
- must advise us of any changes in its collection mode or classifications.

All requests to the source agencies should be clearly documented in the MOU, and information collected about administrative data should be well documented in the metadata. It is helpful to obtain the data model of the provider's database so data can be requested in terms that the source agency uses.

Once you have sufficient understanding of the data, a request for a data extract can be prepared. The request includes the following items:

- content of the file: population, time period, fields required
- how the file will be formatted: file type, field separators, provision of separate look-up tables
- checklist for the extract before it is sent (eg no special characters, valid data values, range checks)
- how the data will be delivered: media, transport, encryption (see section 5.2.2).

Good communication with the data providers, and unambiguous specification of data requirements, reduces the likelihood of the data extract failing to meet the needs of the project. Allow sufficient time for the source agency to extract and supply the data – this can be days to months, and should be discussed with the provider before you submit the request.

As well as the data fields and format, the time period over which the request is made is an important decision. Data integration projects may integrate data over a monthly, quarterly, or annual cycle. This requires precise definition of which records should be supplied for each period, without overlap or gap. Again, using the field names on the data model helps avoid misunderstandings. For example, "all records collected in December quarter 2002" and "all records with creation date field within 1 October to 31 December 2002" could result in different datasets. In the first instance records updated in that period could be included, while in the second instance that would not happen.

The agency receiving the data can detect records that should not have been received. However, there is usually no way for it to check which records have not been received. Deciding on the time period can be complex, because it can involve decisions around, for example, frequency of supply, and what to do about late returns/claims filed much later.

The data request and supply process may need to be iterative, with modifications or corrections being made to the data supplied as needed. We recommend the specification be tested first by transferring a small version of the full dataset.

5.2.2 Data transfer

Statistics NZ corporate standards and policies on data transfer are being developed; at present we have no standard approach.

Various transmission modes, storage, and encryption have been employed in data integration projects, and these methods change as technology improves. If you are undertaking an integration project you should contact Information Management and the Security Office.

5.2.3 Data verification

When a data extract is received, checks can verify:

- the number of records extracted is equal to the number received
- there are no duplicate unique identifiers
- numeric fields contain numbers, and text fields are predominantly text

- all the variables requested are present, and whether any extra variables are provided by mistake
- the range of values in each field is appropriate, and there are no unusual or surprising values
- the distribution of values in each field is as expected
- there is consistency with other fields in the data
- the relationship between files is as expected (only relevant if more than one file is supplied).

5.2.4 Feedback to provider

When the data is successfully validated, the data custodian informs the provider of the successful data transfer and validation. If the data transfer or data validation fails, the provider has to be informed about why it failed; a new set of data is requested. Due to privacy reasons (Privacy Act 1993, see section 3.2), on contacting the provider about a failure in transfer or validation, the data custodian should never disclose personal information, such as 'the record with IRD number XXXX has the payment field missing'.

5.3 Preparing data for record linkage

A number of issues need to be addressed when linking data. Often, data is recorded or captured in different formats and classifications, and data items may be missing or contain errors. A pre-processing phase that aims to edit and standardise the data is therefore an essential first step in every linkage process. Datasets may also contain duplicate entries, in which case you may need to apply linkage within a dataset to de-duplicate it before attempting linkage with other files.

5.3.1 Typical errors in linking variables

Errors in linking variables may occur when capturing and processing these variables. Sources of errors in the linking variables include: variation in spelling, data coding, and preparation; using nicknames or anglicising foreign names; using initials; truncating or abbreviating names and addresses; using compound names; or missing or extra words (Gill, 2001).

Errors occurring in the commonly used linking variables are illustrated below.

Unique numeric identifiers

Unique numeric identifiers, when available, are excellent linking variables. However, very strict control over issuing new identifiers, and recording in the data file, is necessary to produce high-quality linkage with the numeric identifier alone. Typical errors include: missing identifiers (particularly important where links are longitudinal); transcription errors such as transposing digits when recording data; the same identifier may be used for more than one unit; the same unit may have more than one identifier assigned to it (duplicates); the units may refer to different identities in different files. Numeric identifiers that include a 'check digit' are much less likely to be incorrectly recorded.

Surname

The main difficulty with surnames is name changes due to marriage or divorce. People in some ethnic groups have many surnames and the order of their use varies. Linking the birth surname and the marriage or partnership name into a compound (or hyphenated) name is common; both parts are required for linking purposes. Spelling variation is also quite common, due to the effects of transcribing the names through various systems. Some cultures have no exact equivalent of a surname (Gill, 2001).

First names

There are wide variations in the spelling of first names, due to recording and transcription errors. Problems include the use of nicknames and contractions. Some are readily identifiable (eg Jim for James, Will for William, Liz for Elizabeth), but others are not (eg Ginger for Paul, Blondie for Jane). Some records may just record the fact that the person is a baby, or a twin, and until the birth is registered, the record may contain 'baby of ...' or 'twin of…'.

Address

Address is an excellent variable for confirming otherwise questionable links. However, disagreements are hard to interpret because of address changes, address variation, and difference in mailing and physical addresses (Gill, 2001).

Sex

Sex is generally well reported and, except for transcription and recording errors, it is a very reliable variable. The main difficulty is that sex may not always be available in some administrative records when it is not required for operational purposes. Where a dataset does not include sex, it can be generated using the first name, although this cannot be done with complete accuracy (Gill, 2001). Some datasets collect titles such as 'Miss' or 'Mr', which can be used for sex imputation.

Date of birth

Date of birth is generally well reported. Problems may occur when the date is filled in by others (ie by proxy); for example for children and the elderly, when an approximation may be provided. Typical transcription errors arise when day and month are transposed, or when two digits for year are transposed (eg a correct birth date of 11 March 1975 may be recorded as 03/11/75 or 11/03/57). Error also occurs when the current date is entered in the date of birth field, or the current year in the birth year field.

Other problems encountered in using linking variables are:

- first name swapped with surname – happens occasionally
- embedded titles in the name – surname and first name fields may contain titles (eg 'Mr', 'Mrs', 'Dr', 'Jr'). Before the names are used for linking they should be parsed and the various components identified and separated (Gill, 2001).

Sections 7.3.1 and 7.3.2 give more detail on choosing linking variables.

5.3.2 Standardisation: editing, parsing, formatting, concordance

The success of a data integration exercise depends on having standardised data fields. Because of potential quality problems, some variables may be not suitable for linking. Rigorous editing, parsing, and formatting of the linking variables, and creating concordances minimises errors.

- **Editing** is the process of detecting and dealing with erroneous or suspicious data.
- **Parsing** a field separates the entities within that field to make comparison easier (eg a field containing both first name and surname could be split into two new fields).
- **Format** standardisation is necessary when a field is stored in different formats, (eg date of birth is "01Jan2002" in one file and "010102" in another).
- **Concordance** involves creating consistent coding across files and is very important for variables that require classification (eg sex coded as 1 and 2 in one file and as M and F in another).

Editing

While probabilistic linking takes data errors into account, basic data cleaning may be needed before the linking to remove definite errors. Use edit checks to identify invalid responses, such as character strings in a numeric variable, or non-alpha characters (eg # or ^) in a character text response. Other edits may check for 'out of range' or impossible responses (eg birth dates in the future). Often the best approach is to treat these invalid responses as missing values.

Parsing and standardising linking variables

The process of parsing and standardising linking variables involves identifying the constituent parts of the linking variables and representing them in a common, standard way by using look-up tables, lexicons, and phonetic coding systems (Gill, 2001). These standardised individual elements are then rearranged in a standard order.

An example of parsing a name that has title, first name, and surname:

	title:	Mr
Mr John Peter Smith	first name 1:	John
	first name 2:	Peter
	surname:	Smith

Ways to parse and standardise commonly used linking variables are detailed below.

Standardising surnames and first names

The basic uses of name standardisation are: first, to replace many spelling and abbreviation variations of the commonly occurring names and addresses with standard spelling and fixed abbreviations; and, secondly, to use key words generated during the standardisation process as hints for developing parsing subroutines.

Name standardisation in data integration allows the data integration software to work more efficiently, by presenting names in a consistent fashion and by separating out parts of the name of little or no value when making comparisons (Gill, 2001).

In the standardisation process, first name spelling variations such as LIZ and BETTY might be replaced with the original or formal spelling, such as ELIZABETH. Identifying stem words such as FRED can also be converted, although these could equally be associated with ALFRED or FREDERIC. However, these short forms could actually be the real name, so take care when applying this kind of standardisation. Other standardisation procedures sometimes used in formatting names include removing punctuation or blanks. For example, O'BRIEN becomes OBRIEN, TE AROHA becomes TEAROHA, and VAN DAMM becomes VANDAMM.

Dictionaries and lexicons exist that relate commonly used nicknames and name contractions to formal names (eg BOB to ROBERT, LIZA to ELIZABETH) and link common variations in spelling (SMITH, SMYTH, SMYTHE) (Gill, 2001).

Phonetic coding

Phonetic coding involves writing a string of characters based on the way the string is pronounced. It is a useful tool to summarise names and allow for some spelling variations. Used in data linkage, its aim is to dampen the effects of coding errors, which could otherwise result in spurious disagreement between two variables. Two traditional phonetic coding methods used are the Russell SOUNDEX, initially developed for the

1890 United States Census, and the New York State Identification and Intelligence Algorithm (NYSIIS) (see Taft, 1970).

An example: a surname is listed as Camden in one dataset and misspelled in another as Comden. If a character comparison is done on these character strings, the surname variables would disagree. Using SOUNDEX, both Camden and Comden would be coded as ‘C535’. In NYSIIS, both would be coded as ‘CANDAN’. Thus, even if a typographical error occurred in encoding the surname, the two field entries would still agree. (Unless the error was in the first letter, when reverse SOUNDEX /NYSIIS may be employed.)

The NYSIIS has a higher accuracy (discriminating power) than SOUNDEX, but a lower selectivity factor (ie bringing together alternative forms of the same name). For example, assume ‘Days’ and ‘Dais’ are the same surname – NYSIIS would perform relatively poorly, as it codes these surnames differently (non-match). Conversely, assume ‘William’ and ‘Williams’ are the same – SOUNDEX performs poorly this time, as it would code these two surnames as distinct from one another. The choice between NYSIIS and SOUNDEX comes down to the level of trade-off the analyst is willing to accept between these two measures of accuracy and selectivity.

Standardising business names

The main difficulty with business names is that even when they are properly parsed the identifying information may be indeterminate. Sometimes two quite different names can be used to refer to the same business. For example, the burger shops Habib Burger and Smith Burger might merge and change their name to Karori Burger. Or, the names can be quite similar, but the businesses very different. For example, Habib Burger and Habib’s are a burger shop and Arabian food restaurant, respectively. Because the name alone may be insufficient to accurately determine the status of the link, we obtain address information and other identifying characteristics for integration (Gill, 2001).

Standardising addresses

Standardising addresses operates as for names. Abbreviations like Rd or Cres should be replaced by appropriate expansions to Road or Crescent, or to a set of standard abbreviations commonly used by the organisations. For example, when a rural address variation (eg R.D. 1 Tauranga, or Tokoroa Farm Kapiro Road) is encountered, the software should use a set of parsing routines different from those associated with home-number/street-name address.

Parsing divides the free-form address variable into a common set of components that can be compared (eg by street number, suburb, and town). Parsing algorithms often use standardised words. For example, ‘Street’ or ‘Road’ would cause parsing algorithms to apply different procedures than words such as ‘R.D.’ or ‘Auckland’. While exact character-by-character comparison of standardised but unparsed names could result in no links, using the components in the address might help designate some pairs as links. Commercial software and purpose-built Statistics NZ methods are available for parsing and standardising addresses.

Once addresses have been standardised further geocoding may be carried out. Geocoding is the process of finding associated geographic coordinates from other geographic data. For example, standardised addresses might be matched to a list of known dwellings that includes accurate X-Y location details. Geocoding is a complex process that often makes use of specialised datasets and tools.

Concordances

Often there is interest in a variable that is collected using different classifications. One classification may be a simplified version of the other, or one part of the classification may agree across files, while the rest does not. Correctly comparing the variable requires a consistent classification to be used for the linking variable across all datasets.

A concordance can be thought of as a pre-defined conversion table that translates between classifications. Usually the concordance will be between a standard classification and a non-standard one used on a particular dataset. In a simple case, a table could be created that lists each category in the non-standard classification next to its best match in the standard classification. The correspondence is typically not exact, so several concordances are possible. Choosing the best solution requires understanding the concepts behind the data collection and the needs of the people who will use the data.

Ethnicity is a common example. Suppose one dataset uses the categories European New Zealander, Māori, and Other, while another source has European New Zealander, Māori New Zealander, Cook Island Maori, and Other. European New Zealander translates directly, but Cook Island Maori could be included in either Māori or Other in the first file. Knowledge of the collections and expert guidance will lead to your decision.

5.3.3 Deduplication

Duplicate records are common in administrative datasets. They are usually created by mistake, either from the form-filling process or at the input stage in the data source agency. Examples are: typos in filling out a form, or forms being filled out many times when processing a case. Although most agencies have systems to deal with duplication, often some duplicates are left. Deduplication allows a data integration analyst to eliminate duplicates – a file is made to link to itself using the same techniques as for integration. Another option is to ignore the duplicates but to estimate how many the dataset contains, which is helpful in understanding their effect on the final integrated data.

The effect duplicates have on integration depends on their frequency, how they are generated, and what type of integrated dataset is being created. False positives occur if the duplicates are linked to the wrong unit. If the resulting integrated dataset is the intersection of the source files, then unlinked duplicates will appear as false negative links.

Researchers need to take most care where the integrated data is the union of the source files, as the unlinked duplicates will inflate the number of cases in the final integrated file.

5.3.4 Anonymising unique identifiers

The use of unique identifiers (UIDs) assigned by other agencies must meet the requirements of the Privacy Act 1993. See chapter 3 about using them in data integration, in particular section 3.2.1.

Statistics NZ converts any UID assigned by another agency and passed to us as part of an integration project to an internal unique identifier (IUID) as soon as practicably possible. We create the IUID using a common encryption process and a key unique to each project.

External UIDs are retained within our systems (servers, databases, and applications) only as long as is necessary to perform validation, editing, and integration. They are then replaced and removed completely.

An externally assigned UID is not used for longitudinal linking. The IUID provides the capacity to create a consistent longitudinal link to the same unit without the need for us to store the original UID in any production database.

6 Statistical theory of record linkage

To understand the process of data integration, it is important to understand the statistical theory behind it. This chapter looks at the data integration process, building up from a simple integration situation (that of exact linking), to how human beings process the integration, to the mathematical/computer process.

Figure 3
Schematic representation of the record linkage process

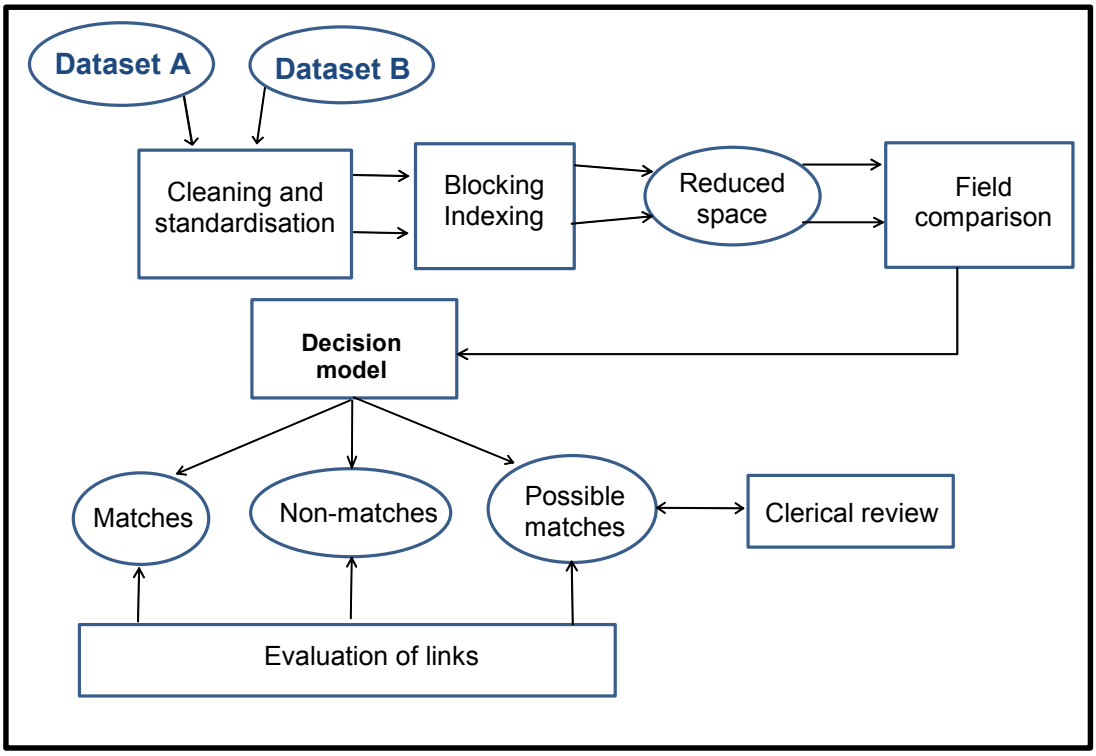


Figure 3 is taken from CULT Project (2014). It shows the record linkage process from a statistical point of view, beginning with two datasets A and B and finishing with evaluation of the matches produced. At each stage, you should consider a range of theoretical and practical options, and the solutions chosen will depend on the nature of the data and the project.

Cleaning and standardisation includes all the pre-linkage data manipulation that is necessary before actual record linkage can begin. It is described in section 5.3.

Blocking and indexing are methods for sorting records into manageable subgroups, to reduce the number of required field comparisons to a practical number and improve the efficiency of the linking process. Blocking is explained in section 6.5.

Reduced space refers to the subgroups of records from the two datasets that are identified in the blocking and indexing stage. The search for links between the datasets occurs within these reduced spaces.

Field comparison is where the values of matching variables on a pair of records are evaluated to determine how close they are to each other. This requires comparison functions, which have many forms. Section 7.3.3 has an overview of comparison functions.

The **Decision model** is a set of rules that categorise pairs of records into matches, non-matches, or possible matches based on the results of the field comparisons and other information about the datasets. The widely used Fellegi-Sunter rule, cut-off weights, and thresholds are described in this chapter, along with intuitive rules that would be applied in clerical checking or manual linking.

Clerical review and evaluation are the final steps of the linkage process, where the output of the rules and comparisons is measured and important quality measures such as the rate of linkage errors are estimated. These are explained in section 7.4.

This chapter explains the theory behind some of the major components of this process, such as blocking and field comparisons. We also introduce the mathematical foundations of record linkage theory at a basic level. It is not intended to be a complete view of the mathematics of data integration.

Chapter 7 examines the practical aspects of the decision model in more detail.

6.1 Exact linking

When two files contain the same unique identifier (UID), they can be linked via that UID – 'exact linking'. A UID might be either a single variable, or a combination of variables (eg name, date of birth, and sex) as long as they are of sufficient quality to be used in combination to uniquely define a record.

Exact linking has no uncertainty. Either a pair of records agrees on the unique identifier or they do not. The problem is when the quality of the variables is not good enough to guarantee that the value of the UID is available, correct, and unique. Where exact linking alone will not result in a sufficiently robust integrated dataset, probabilistic linking may be used.

6.2 Matches and links

Two records are a 'match' when they relate to the same unit (person/ business/ entity/ event). Data integration's role is to determine which records are a match. To differentiate the term match from other uses of the word, we use the term 'true match'.

Two records are a 'link' if, by some process, we determine the two records refer to the same unit. Creating links is what the data integration process does. Ideally all true matches in the datasets would be linked, and all links created would be true matches. In practice, our process may link records that are not a true match, or fail to link two records that are a true match.

Table 1
Possible linking results

	True match	True non-match
Link	Correct outcome	False positive link
Non-link	False negative link	Correct outcome

See section 7.4.2 for more details on false negative and false positive links.

Matching, record linkage, or simply 'linking' is the process of comparing records and deciding which ones are links. The variables used in the linking process are called blocking variables and linking variables, linking fields, linking variables, or comparison variables.

6.3 Linking files

This discussion assumes the following scenario. There are two files, file A and file B. The task is to compare one record from each file, and decide whether or not the records should be linked.

Table 2
Two records to be compared

File A	Recorded values
Name	John Black
Date of birth	23-11-63
Sex	M
Address	112 Hiropi Street

File B	Recorded values
Name	Jon Block
Date of birth	23-11-65
Sex	M
Address	89 Molesworth St

6.3.1 The human approach

When looking at two records and comparing them, human beings judge how likely they think it is that the two records refer to the same unit.

Consider table 2. The aim is to determine if the two records refer to the same person, so each field must be compared to judge how likely this is. An initial impression might be that these records are for the same person, so we seek further evidence to see if this is true.

For the **name** field, the difference can be caused by: spelling mistake; error created when one record was entered; someone trying to read scrawled handwriting; or a scanning error. For the **date of birth** field, the day and month are the same, although the year is two years out, which could have a similar explanation as for name. The **sex** field agrees. If one record had ‘F’, people would likely view that as an obvious error, as neither John nor Jon are feminine names (although Jon may be a misspelling of Joan).

For **address**, the fields do not agree at all. John may have moved between file A and file B being recorded. Or one file may be the home address and one the work address. Information about how these files were collected, for what purpose, and what the information means, would influence the quality and how much reliance to place on the address.

Assuming the address field is unreliable and the difference is explained as a data quality issue, we would declare these records refer to the same person. However, there is the possibility that another record might be an even better match for one of these records.

Other information can be a factor in decision making. For example, there might be 100,000 records with Black as a surname and 50,000 with Block as a surname. In such a situation, the chance of getting the two records in table 2 is likely. With this knowledge, we may decide the differences are very important, and therefore these two records do not refer to the same person.

A central problem in record linkage is the number of comparisons that must be done: if file A contains 1,000 records, and file B contains 10,000 records, the number of possible record pairs is 1,000 x 10,000, or 10 million record pairs. While humans can decide whether to link records or not, they cannot do thousands of records a minute, which is the degree of speed needed to make a data integration project feasible. Computers can handle such speeds, but they need to be told how to apply judgement under uncertainty and in a consistent manner.

6.3.2 The mathematical approach

Here, we introduce the theory of probabilistic record linkage as formalised by Fellegi and Sunter (1969) in simple terms. Other useful and accessible references include Jaro (1995) and Winkler (1995). Much of the record linkage software available, including the software used at Statistics NZ, is based on this approach.

When comparing two records, the computer compares each field and assigns a measure that reflects how similar they are. This measure is called the 'field weight'. It is calculated from two pieces of information: the reliability and commonness of the data.

The m probability

The **reliability** of the data is described by the 'm probability' or 'm prob', a measure of the trustworthiness of the data. It can be expressed as the 'probability of two values agreeing given that they refer to the same unit'.

m = Pr (two values agree | the records are a match)

Another way of thinking about it is: given that two records are a match, how likely is it that an issue with the data makes the values different (eg an error, inconsistent definition, timing difference)? This is how the m prob is related to the data quality.

m = 1 – Pr (two values disagree | the records are a match)

For example, sex might be very well collected and monitored, so we give it an m prob of 0.98. Address might be collected as a matter of course with no checks made on it at all, so it gets an m prob of 0.70.

Chapter 7 discusses how to determine the m prob for a particular variable.

The u probability

The **commonness** of the value is described by the 'u probability' or 'u prob', a measure of how likely it is that two values will agree by chance. It is expressed as the probability of two values agreeing given that the records do not relate to the same unit.

u = Pr (two values agree | the records are not a match)

Simply, this is a measure of relative frequency. The more common a value is, the more likely two unrelated records are going to contain that value. Hence, the u prob is often defined as:

u = 1 / (number of different values)

For example, sex is either male or female with about equal probability in the general population, so the u prob is 0.50. There are 12 months in a year, so a month of birth variable would have a u prob of 0.08. Address is usually unique, so the u prob is 0.01 or lower.

These u probability examples assume that each value has the same probability of occurring and is known as the 'global u probability'. Specific u probabilities can be created for each value that a field can take, allowing for non-uniform distributions. For

example, surnames such as ‘Smith’ and ‘Quimby’ would have different relative frequencies, and could usefully be assigned a specific u probability.

The field weight

From these component probabilities, we calculate a weight for the field. The calculation used depends on whether or not the two values in the field agree. If they do agree, a positive weight is generated, and if they disagree a negative weight is generated. The weight’s size measures the evidence the values provide about the record pair being a match. The two calculations are:

$$\text{agreement field weight} = \log_2 \frac{m}{u}$$

and

$$\text{disagreement field weight} = \log_2 \left[\frac{1-m}{1-u}\right]$$

This is the log of the likelihood ratio and is related to the probability of agreement (or disagreement). Logs are used to simplify calculations because field weights, assuming independence, become additive. The use of base 2 in the logarithm is a convention that is used in the information theory work that these formulae are based on.

The composite weight

Once field weights are calculated, we calculate a composite weight for the entire record, based on the variables examined. The composite weight for the record pair is simply the sum of the field weights, and is referred to as ‘weight’ when talking about record pair weights.

Note: adding the weights is equivalent to multiplying the likelihood ratios, just as one would multiply independent probabilities. We assume the fields are independent of each other, and that errors in the fields occur independently, although this is not necessarily true in the real world.

Example

Using the data from table 2, first the m and u probs are set up, and then we calculate the field weights.

Table 3
Calculating field weights

File	m prob	u prob	Agreement field weight	Disagreement field weight
Name	0.95	0.01	6.57	-4.31
Date of birth	0.9	0.01	6.49	-3.31
Sex	0.95	0.5	0.93	-3.32
Address	0.7	0.01	6.13	-1.72

Next, the fields are compared.

Table 4
Comparing fields for agreement

File	Agreement?	Field weight
Name	No	-4.31
Date of birth	No	-3.31
Sex	Yes	0.93
Address	No	-1.72

This gives the final composite weight for this pair of records as:

$$(-4.31) + (-3.31) + (0.93) + (-1.72) = -8.41$$

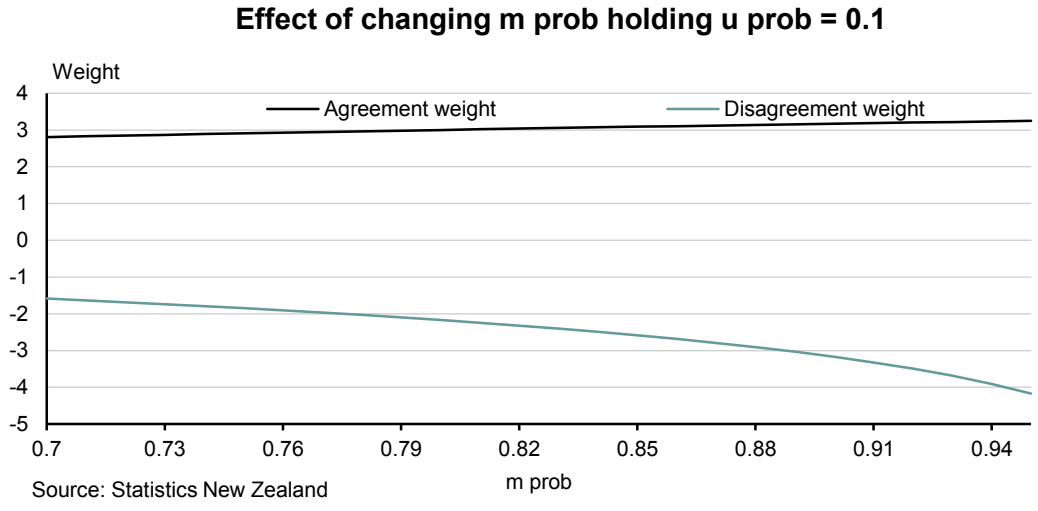
As this final weight is negative, the linking process would reject the link between the two record pairs. In practice, we allow for partial agreements; for example, a minor difference in spelling could generate a lower, but still positive, agreement weight.

Changing the m and u probabilities

The best way to understand m and u probs and how they affect weight calculations is to take an example and make changes to one value at a time, holding the other constant. The following examples look at agreements and disagreements between record pairs on a single comparison variable.

Figure 4 demonstrates the effect of a changing m probability on the agreement and disagreement weights of a given field when the u probability is fixed. Fixing the u probability is like assuming that the chance of two records agreeing when they are not a true match is constant. A higher m probability means there is a higher chance that two records that are a true match have the same value.

Figure 4

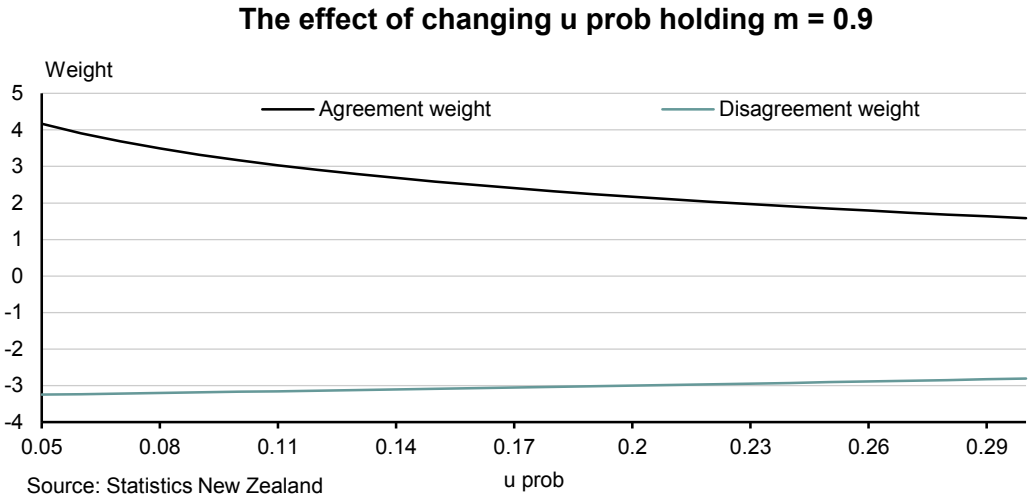


The main feature of figure 4 is that the increasing m probability causes a disagreement to weigh more heavily against the possibility that a given comparison pair is a true match. A high m probability implies a low probability of error in a field, so there is a small chance that disagreeing records are a true match.

On the other hand, the agreement weight changes very little in this example graph because we have assumed that the rate of false agreements between records which are not true matches (the u probability) is unchanged.

In figure 5, the opposite situation is pictured: the m probability is held constant while the u probability increases. Here it is the agreement weight which is strongly affected. This is because a higher u probability implies that there are more agreements which occur purely by chance.

Figure 5



To summarise:

- As the m probability changes, the disagreement weight changes more than the agreement weight.
- As the u probability changes, the agreement weight changes more than the disagreement weight.

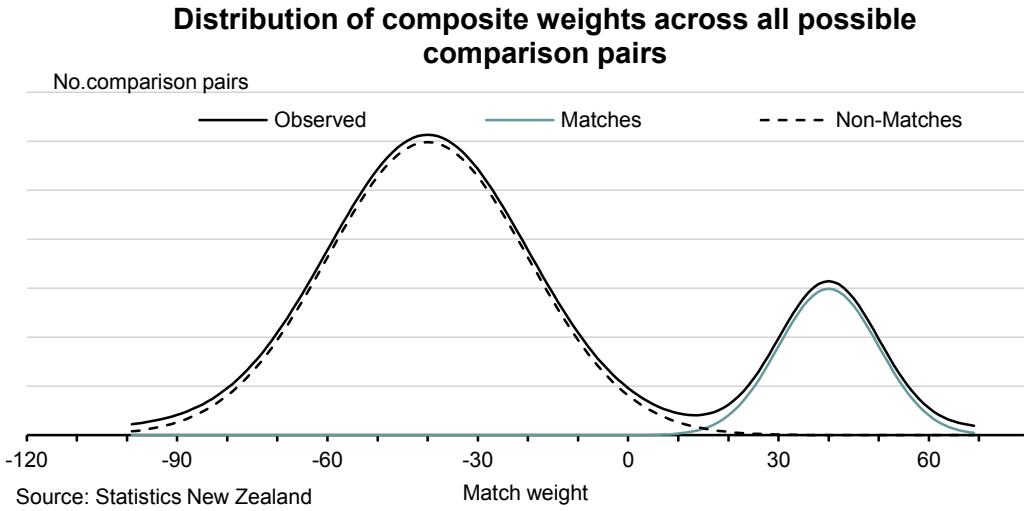
6.4 Weights

Using the calculated weights as evidence, the next step is to decide which records are links and which are non-links.

6.4.1 Distribution of composite weights

In a typical integration project there are hundreds of thousands of records, and millions of possible pairings. Most record pairs do not refer to the same unit, and thus more non-links than links are created. The distribution of these weights is therefore bimodal.

Figure 6



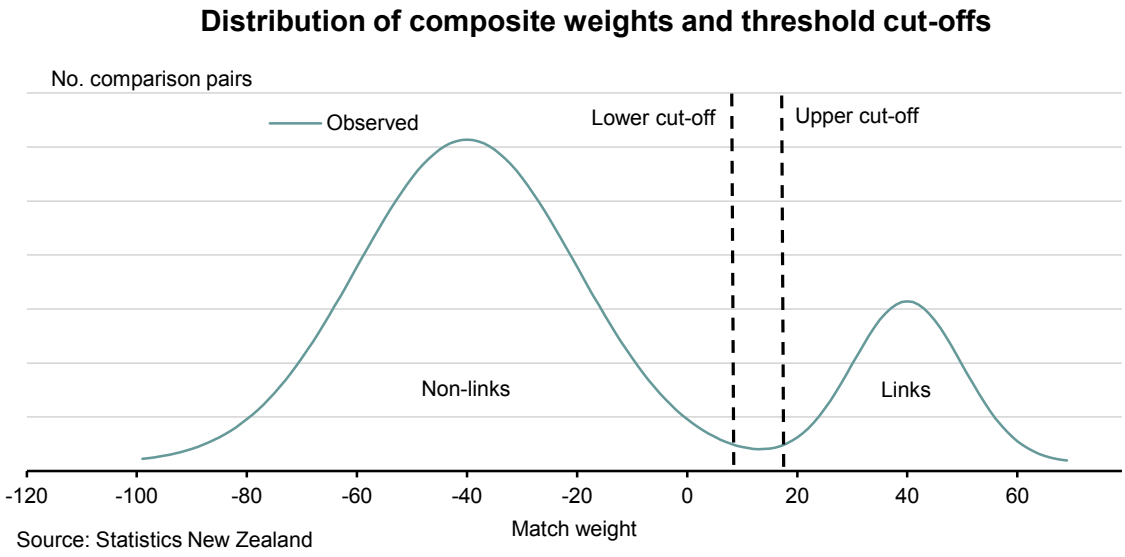
The ‘observed’ line in figure 6 is the distribution actually observed, which is offset slightly to differentiate it from the other distributions.

6.4.2 Cut-off thresholds

Once the weights are calculated, upper and lower thresholds are established. The upper threshold is the weight above which every record pair is determined to be a link. There is usually only one ‘match’ per record (this depends on the datasets being linked; see section 7.1 for different linking types). Other possible pairings can either be ignored or considered duplicate records.

The lower threshold is the weight below which every record pair is determined to be a non-link.

Figure 7



Although record pairs are in reality either true matches or true non-matches, in the world of data integration, with imperfect/insufficient data, the picture isn't so clear. Some true matches have low weights because of data errors or similar problems; some true non-matches have high weights for the same reasons.

Data integration theory does provide optimal ways to determine the threshold levels (Fellegi and Sunter, 1969), but these are often not feasible in practice. Generally it is the person working on the data integration project who decides where the cut-off thresholds go. This is usually done by reviewing record pairs near a likely cut-off point and making judgements about how the computer differentiated the pairings.

See chapter 7 for more detail about the effect of setting particular threshold levels.

6.4.3 Clerical review

If the upper and lower thresholds are equal, this divides the set of record pairs cleanly into two sets. However, if they are not, then the record pairs with weights in between the two limits are in the 'clerical review' area. In this area, the analyst decides which record pairs are links and which are non-links.

With some statistical integration software, it is not possible to assess record pairs in a clerical review area. In this case, the link and non-link thresholds need to be the same.

6.5 Blocking

Comparing 1,000 records with 1,000 records means that 1,000,000 comparisons are made. With only 1,000 record pairs being a match, this gives 999,000 records pairs that are non-matches – these are determined to be non-links.

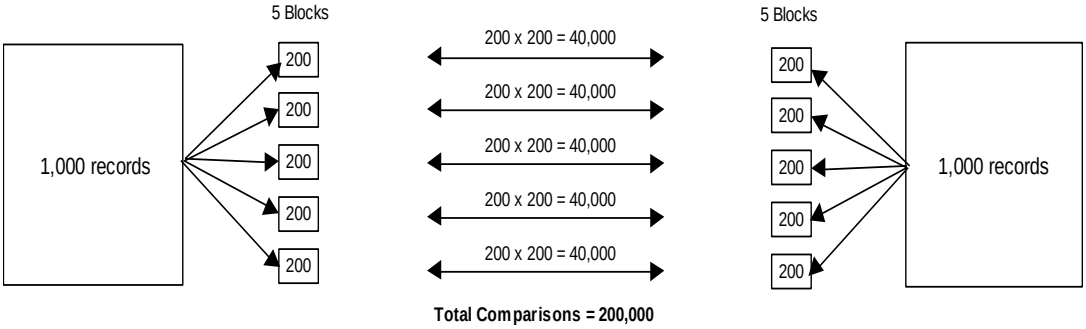
To reduce the number of comparisons made and focus on the records that are more likely to be matches, records can be filtered first so only certain records are considered when making comparisons.

This filtering, called 'blocking', is done by selecting variables to 'block on'. Only records that agree on the values in those variables are compared with each other. For example, if we choose sex as a blocking variable, only records with the same value for sex are compared. This cuts out about half the comparisons required. If month of birth was

chosen, the number of comparisons is reduced by a factor of 12. Choosing both sex and month of birth means that 1/24th of the comparisons are made.

Figure 8 illustrates the reduction in comparisons where there are five equally sized blocks on each file.

Figure 8
Record comparison process with blocking



In table 2, the John Black and Jon Block record pair, if sex is used as a blocking variable, the two records are still compared. If year of birth is used as a blocking variable, they aren't compared.

6.6 Passes

In table 2, if we chose year to block on, the two records would not be compared. If this was the only comparison done, then these records would never be compared. However, we can run more than one comparison.

A 'pass' is an iteration of record linkage that uses a combination of blocking variables and linking variables. In a data integration project more than one pass is used to block the file in different ways. This allows for different variable comparisons and for errors in the blocking variables. For example, one pass might block on year of birth, and match on name and sex. Another pass might block on sex, and match on name and address.

The proper ordering of passes helps to improve both processing speed and the quality of the final links. Usually passes should be arranged so that links identified using the most specific blocking come first, and the blocks with fewer restrictions come last. For example, in two datasets with a unique identifier (missing on some records), first and last names, and dates of birth, you could use passes in this order:

1. Blocking on unique identifier
2. Blocking on date of birth
3. Blocking on first and last name
4. Blocking on Soundex codes for first and last name.

In the first pass, large numbers of very high-quality links with the same unique identifier will be found quickly and will not interfere with subsequent passes.

The number of passes used should reflect how well the record linkage process is working and the quality and number of linking variables. In our projects, we have found that a few (between three and six) well-chosen passes will capture nearly all the reasonable-quality links that can be made. In some cases extra passes may be able to find a small number of high-quality links, but generally you should try to minimise the number of passes to avoid poor-quality linking. Past a certain point, there often simply isn't enough information to give high-quality links and more passes won't help.

7 Record linkage in practice

This chapter focuses on the practical application of linking. It covers the types of linking, the flow of the integration process, what can go wrong, and discusses what makes the output of a record linkage exercise 'fit for use'.

7.1 Types of linking

A data linkage exercise may take several forms – for instance, a many-to-one link (eg geocoding), a deduplication exercise, or a one-to-one link.

7.1.1 Many-to-one linking

In a many-to-one link, a record from one dataset (file) is allowed to link to more than one record in another. When linking data from Inland Revenue with education data, a person would have one tax record, but might be enrolled in more than one institution. Geocoding is a common application that involves many-to-one linking. (Geocoding is the process of linking addresses to a geographic location) For example:

File A (three location records):

- Rossmore House
- Aorangi House
- Backbencher

File B (one location record):

- Molesworth St

Outcome: Each record in file A links with the file B record, to produce a file with three records for location:

Rossmore House ⟷ Molesworth St

Aorangi House ⟷ Molesworth St

Backbencher ⟷ Molesworth St

Deduplicating – cleaning up lists

Deduplication removes the duplicates on a file. The process consolidates multiple occurrences of a single unit within one dataset into a single record. This might be done on an address list or client database to ensure the file is as clean as possible for its purpose. Theoretically, it is the same process as linking two files, although here the second file is the same as the first – the practice can be slightly different, depending on the software.

Table 5
Example of deduplication

Input file (5 records)	
Name	**Birth date**
Amy	15 June 1985
Bill	22 March 1987
Chris	1 September 1980
Cris	1 September 1980
Dave	12 August 1990

Output file after deduplication (4 records)	
Name	**Birth date**
Amy	15 June 1985
Bill	22 March 1987
Chris	1 September 1980
Dave	12 August 1990

In the example above, the choice to retain the spelling Chris or Cris is left to the analyst.

7.1.2 Many-to-many linking

Many-to-many linking is similar to many-to-one, in that it allows multiple records to link. But it is also possible for records on either file to link to more than one record on the other file. To date, no integration project at Statistics NZ has involved creating many-to-many links.

7.1.3 One-to-one linking

In a data integration project with one-to-one linking, one record on file A links to one record on file B. This is the situation in the New Zealand Census Mortality Study, where one death record should link to one census record.

Table 6
Example of one-to-one linking

File A (7 records)	File B (6 records)	Link or no link	Outcome 1: union of file A and file B (8 records)	Outcome 2: intersection of file A and file B (5 records)
Nicolla	Nicola	link	Nicolla or Nicola	Nicolla or Nicola
Mike	Mick	link	Mike or Mick	Mike or Mick
John	Jon	link	John or Jon	John or Jon
Sharon	Sharon	link	Sharon	Sharon
Jamas	James	link	Jamas or James	Jamas or James
Rissa	Andy		Rissa	
Allyson			Allyson	
			Andy	

Our focus in this document is one-to-one linking, but much of this content applies to the other data linkage forms.

7.2 Pre-linking process

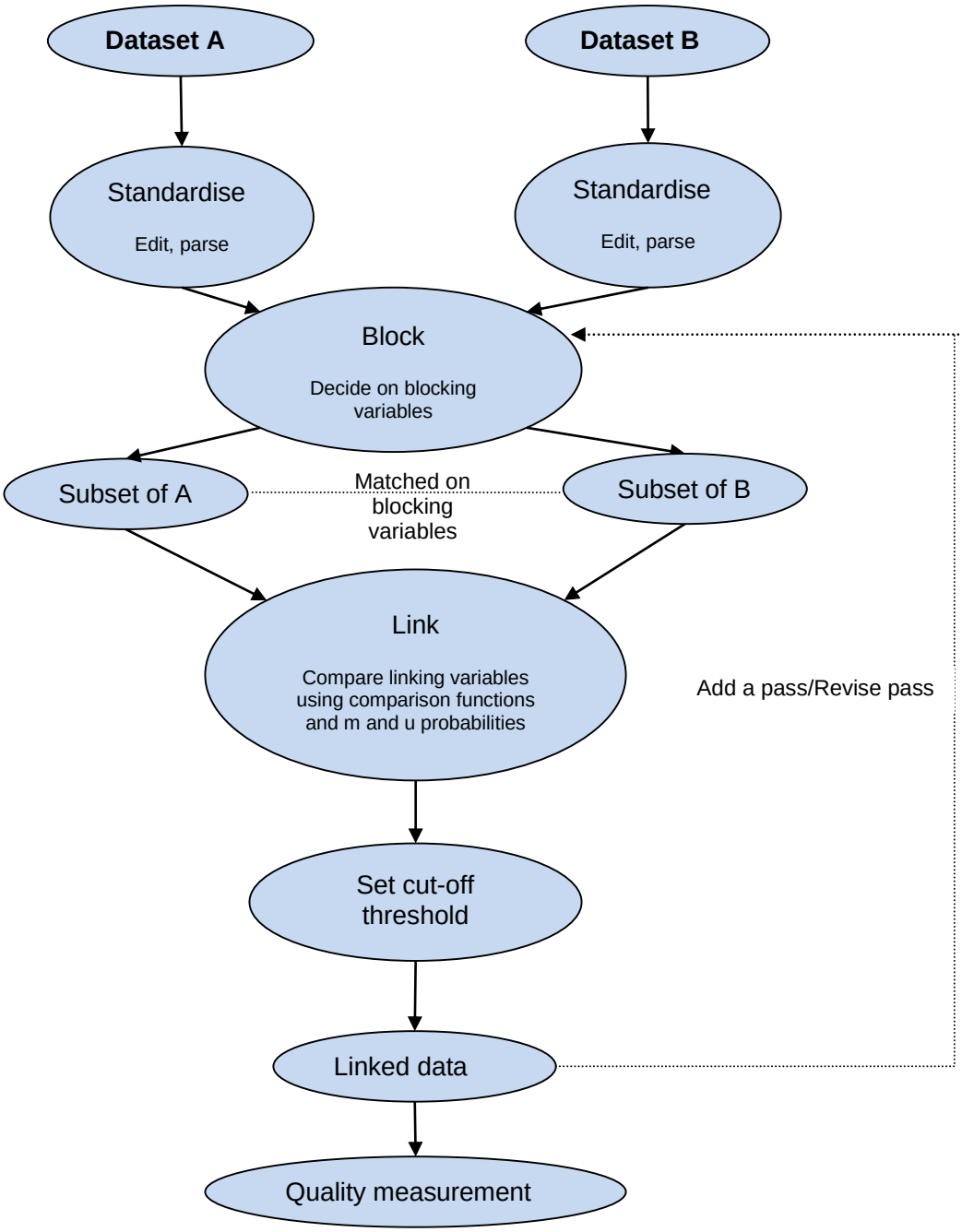
7.2.1 Deduplication

Deduplication removes duplicate records belonging to the same entity within the same file. Even if a certain level of duplication is acceptable for planning or research purposes (Community Services Ministers' Advisory Council, 2004), at times it is best to remove duplicates from the files before you begin integration. Retaining duplicates complicates the integration, especially when integrating multiple datasets.

7.2.2 A data integration process flow

Figure 9 shows a typical process flow for data integration. This chapter goes on to describe each step in figure 9 in more detail.

Figure 9
Data integration process flow



Dataset standardisation involves data editing and parsing to a prescribed format that allows you to compare field entries. Gill (2001) notes that 75 percent of the effort in a data integration exercise is devoted to preparing the input files. Chapter 5 has a detailed discussion of the data standardisation process.

Blocking, a filtering process that reduces the number of record comparison pairs, is discussed in chapter 6. Section 7.3.1 looks at how blocking is done in practice.

The linking step is closely examined in section 7.3.2 through pointers on choosing linking variables. Section 7.3.3 covers commonly used comparison functions for linking variables. These functions provide a way to decide whether two variables being

compared fully agree, partially agree, or fully disagree. In section 7.3.4, we cover the iterative approach to determining the m and u probability values.

We discuss the considerations necessary for determining a cut-off value in section 7.4.1. A composite weight cut-off value is important, as it dissociates record pairs that are considered to be linked from those that aren't. Estimating false positives, false negatives, and link rates, and measurement error in data integration, are discussed in sections 7.4.2 and 7.4.3.

Chapter 7 concludes by discussing issues to consider when the integration is not a one-off situation, but data is added over time.

7.2.3 Standardised datasets

If no standardisation is carried out, records that are true matches may not be linked – because the common variables might appear to be so different that the composite weight could be low or negative. Standardisation is discussed in detail in section 5.3.2.

Field standardisation may be carried out using rule sets. Rule sets are groups of files that contain the rules to parse free-form fields. They are based on recognising common patterns and elements, such as street numbers, street names, suburbs, and postcodes in an address field. Linking software may have built-in rule sets that can be modified. Alternatively, you can apply new rule sets or reformatting outside the linking software environment.

For example, a surname field called SURNAME_TEXT might be standardised as follows:

Surnames – File A

Take SURNAME_TEXT
Capitalise
Remove spaces
Remove any characters other than alphabetic characters
Name the resulting field SURNAME1
Define new variable INITIAL_SURNAME = first character of SURNAME1
Define new variable SOUNDEX_SURNAME = SOUNDEX code of SURNAME1

Surnames – File B

Take SURNAME_TEXT
Capitalise
Remove spaces
Set to missing if surname contains "UNKNOWN"
Remove any characters other than alphabetic characters
Name the resulting field SURNAME1
Define new variable INITIAL_SURNAME= first character of SURNAME1
Define new variable SOUNDEX_SURNAME = SOUNDEX code of SURNAME1

In this example, standardising file B requires the extra step of removing the word "UNKNOWN" if it exists in the surname text; the end result is two datasets that produce equivalent standardised variables for surname (SURNAME1, INITIAL_SURNAME and SOUNDEX_SURNAME).

7.3 Linking method

7.3.1 Choosing blocking variables

Blocking is employed to efficiently compare two datasets by reducing the number of records to compare. For example, to link a dataset having 100,000 records with another containing 1,000,000 records, the total number of comparisons would be 100,000 x

1,000,000. Blocking cuts down the total number of records to compare by comparing only the records that exactly match on the specified blocking variable. In effect, comparison is cut down to only those records with a potential to match, as specified by the blocking variables.

In choosing the blocking variable, the analyst aims to keep the size of the block small, to efficiently reduce the number of comparison pairs, yet big enough to avoid missing true matching record pairs (Baxter et all, 2003). For example, if the analyst blocks on sex, two huge blocks are created, which results in an inefficiently large number of comparisons to perform. On the other hand, blocking performed on a numeric identifier produces numerous mini-blocks – perhaps as many records as there are in the datasets. A problem arises when there is an error or missing value for the blocking variable. Two matching records will not be compared and the match will be missed.

We have successfully used the following methods to design blocks of good quality and size.

Choose a variable that has a good number of values (eg overcomes the sex variable problem in the example above), with a fairly uniform distribution to create blocks of uniform size. Uniform size is desired because the number of comparison pairs "generated by any blocking method depends on the number of blocks (the method) generated and (the resulting) blocks' sizes. Very large blocks have therefore dominant effects on the efficiency of the blocking methods" (Gu & Baxter, 2004).

Use highly reliable variables to block to avoid having two matching records failing to be in the same block, with no chance of being linked.

"Keep the block sizes as small as possible and compensate for errors in blocking by running multiple passes" (Ascential Software, 2002). This technique uses multiple blocking variables in the different passes to overcome block size problems (very large blocks may greatly slow the linkage software or even cause it to crash) and data errors. Essentially, each time a pass is run the links are kept; another pass with new blocks and new comparison pairs is performed on the remaining unlinked records. New blocks and new comparison pairs mean more chance of making true matches.

Truncated fields can mitigate the effects of erroneous encoding when blocking, using phonetic coding and variables that are reliable. For instance, because the SOUNDEX for surname 'William' and 'Williams' return different codes, consider using a new variable containing a truncated form of the surname. This new field, together with other fields, could produce new matches that might otherwise be missed.

Event or birth dates separated into day, month, and year; first names; and surnames (or their corresponding phonetic codes) are good blocking variables. UID numbers, although they may be erroneous or missing, partition the files into a large number of sets. Unless there is rigorous control of issuing and recording identifiers, we recommend using UIDs as blocking variables in the first pass, using other linking variables to verify the link. Other variables can be used to block in subsequent passes.

For datasets relating to units other than individuals, similar principles apply but the variables will be different. For example, when trying to link businesses good blocking variables could be location, legal or trading name, and industry.

Sparsely populated fields are not good for blocking purposes, since records with missing values remain unblocked and are ineligible for potential linking.

7.3.2 Choosing linking variables

You can use almost any variables common to the two datasets undergoing integration for linking. In linking, redundancies in the information from the related variables may help to

reduce linking errors, provided the errors are not highly correlated or functionally dependent (Gu et al, 2003).

Note: linkage software does not necessarily compute correlations. Moreover, we do not advise having highly correlated linking variables in the same pass, as they increase the composite weight without providing additional discrimination between record pairs that should link and those that should not. However, usually only a subset of the variables common to the datasets is used for linking.

Gill (2001) suggests six groups of variables, and using a combination of variables from the different groups for linking records about individuals. The six groups are:

- Group 1: proper names, which rarely change over a person's lifetime (except possibly for a woman's surname) (eg first names, initials, surnames)
- Group 2: non-name personal characteristics, which rarely change over a lifetime (eg date of birth, sex)
- Group 3: socio-demographic variables that may have several changes over a lifetime (eg address, marital status)
- Group 4: variables collected for special registers (eg occupation, date of injury, diagnosis)
- Group 5: variables used for family record linkage (eg surnames in group 1 plus other surnames, birth weight)
- Group 6: arbitrarily allocated numbers that identify the record (eg IRD number).

Gill notes it is common practice to combine linking variables from groups 1, 2, 3, and 6. For businesses similar principles can again be applied, but the variables will be different and will depend on what is included in the datasets.

Possible problems in choosing linking variables

When choosing linking variables, spelling errors, phonetic coding choice, and the like may affect classification of the variables as either 'agreeing' or 'disagreeing'. A quick run-through of the problems with some common variables and what has been done in practice to increase their reliability will help the analyst select the best linking (and blocking) variables.

Surnames

Surnames can change as a result of marriage and divorce and order varies in some ethnic groups. Spelling variations can result from erroneous transcription. A phonetically coded surname may reduce transcription/spelling errors. You can use a surname array (different surname fields merged into one) to handle multiple surnames. The arrays can be compared using a comparison function (see section 7.3) to allow for misspellings.

First names

First names have many of the same problems as surnames. Modernised versions and nicknames may be used in some documents, while the formal first name is used in others. Transcription or spelling errors can occur. Sometimes only first name initials, not the full name, are available from the dataset. You can create an array of the initials as a new variable.

Sex

Sex has a low discriminatory power in distinguishing between a match and a non-match. In some datasets a default value or a guess is assigned for sex when the sex is missing. Sex can sometimes be derived from first names, but this will result in some errors as well.

Birth date

Birth date format can vary (eg European v American format, although this should be handled during standardisation). Birth month and birth day are usually more reliable than the full birth date. Gill suggests some tolerance when using the birth year, as this is more prone to error than the month or day of birth. Transcription errors can occur when numbers sound similar (eg 7 and 11).

Age

Age can be used with some tolerance, as for year of birth. When age and birth date are available, perform a data check to see if these two agree.

Address

Format problems can occur with address, but the field can be standardised – although this process can be laborious without a good rule set or specialised software. Address is a good field for confirming matches, since it's unlikely that, for example, two different people with the same name would live at the same address. However, addresses are often not very useful for disagreements because people change address fairly frequently. When address is not used as a linking variable, you can sort the unlinked records from each dataset being integrated according to address, and then compare the sorted files to check if any matching records have failed to link.

We have found that the standardised forms of the above variables are reliable across most government datasets we integrate, but this may not be the case for other sources. Understanding the collection, processing, and updating of your specific datasets is crucial in determining the reliability of variables.

7.3.3 Commonly used comparison functions for linking variables

Each field that is used for linking has an agreement and disagreement weight that depend on the m and u probabilities for that field (see section 6.3.2). When two records are compared, the full agreement weight is added to the overall match weight if the fields are exactly the same. If the fields are different, the simplest approach is to assign the full disagreement weight to the comparison. However, in many situations we want to be able to find records that are matches despite not having exactly the same values across all fields. Specialised comparison functions allow for partial agreements and the computation of partial weights. This allows us to formalise intuitive ideas about how 'close' two values are so that automated software can search for the most likely match for a given record.

Some simple and commonly used comparison functions are outlined below. Many other comparison functions and encodings have been developed to assist record linkage tasks. For a detailed survey, see Koudas et al (2006).

Comparisons for numeric variables

Absolute difference comparison

The simplest comparison is to compute the arithmetic difference between two numeric values, allowing a certain fixed level of error.

Suppose we are considering an age field, and we have decided that, given the properties of the data, two ages within five years of each other will be considered to match. If the age in the first dataset is 24 and the age in the other is 28, the difference is within the allowed tolerance so the full agreement weight is assigned to the field age for this particular comparison pair. However, if the age in the second dataset is 30, the field weight for age would be the full disagreement weight, since the difference is beyond the tolerance.

Comparing numeric fields with tolerance for error

To take the closeness of two numeric values into account, you can use a prorated comparison. This assigns a partial weight, depending on how close the values of two numeric fields are. Given a tolerance parameter, a weight of zero is assigned if the difference is greater than the parameter. If the difference is zero then the full agreement weight is assigned. Any difference between zero and the tolerance value receives a weight that decreases as the difference increases. For example, given a tolerance of 15, a difference of 8 would receive a weight exactly half-way between the agreement and disagreement weights.

A generalisation of this comparison can be used in cases where the sign of the difference is important. It can be implemented by having two different tolerance values – one for cases where the value on file A is greater than the value on file B, and one for cases where the value on file B is greater than the value on file A.

Other numeric comparisons

Many mathematical functions could be applied to measure the similarity between two numeric variables, depending on the known properties and likely errors in the data. For example, for some numeric variables a percentage-based tolerance might make more sense than an absolute value tolerance. In other situations, the square of the difference between two numbers may be a better penalty function to apply.

Comparisons for character variables

There are many possible comparison functions for character variables (also referred to as ‘strings’). Different comparisons will often work better for different types of character variables. For instance, specialised name comparisons can take into account common causes of difference, such as missing middle names or likely misspelling patterns.

A simple exact comparison is easy to implement for character variables. The difficult aspect of string comparisons versus numeric comparisons is that there is no obvious best measure of the ‘distance’ between two different strings. Because of this, there is a wide range of string comparison functions, and they can be very complex. In this manual we describe only the main families and some common examples.

Koudas et al (2006) describe three main types of comparisons (or ‘similarity measures’) for strings.

1. **Coding-based comparisons**

These use a transformation designed to encode similar strings into the same code. A simple example is truncation, where only the first, say, four characters of a name field are compared, to allow for mistakes at the end of a name. A more complex example is the SOUNDEX system (mentioned earlier), which is designed so that strings with similar pronunciation turn into the same four-character code. An exact match on the encoded variables effectively becomes a comparison function on the original variables, with a certain level of tolerance allowed.

2. **Edit distance comparisons**

The basis for these comparisons is to determine the amount of alteration needed to make two strings the same. By computing a numerical ‘edit distance’ between two strings, these methods reduce the comparison to a numeric score, which can be used to weight and rank possible matches. Two of the most common comparison functions in this category are the Jaro-Winkler distance and the Damerau-Levenshtein distance.

Edit distance comparisons typically define an allowed set of string operations – such as deleting one character, inserting one character, replacing one character by another, or swapping two adjacent characters. Any string can be transformed into any other string by a sequence of these operations, so the difference between two strings can be measured

by determining the minimum number of basic operations required to complete this transformation. Algorithms to compute edit distances can be complicated but are implemented in many common software packages.

The edit distance can be generalised by assigning different 'costs' to each operation, and finding the minimum cost rather than the minimum number of operations. This generalisation can allow a more realistic measure that recognises certain mistakes are more likely in a given data set or field. For a field that relies on somebody manually typing in data, for example, swapping two characters might be given a relatively low cost compared with inserting a new character. For a field produced by optical character recognition of a handwritten form, certain character replacement errors (eg confusing 'v' for 'u') might be given lower costs.

As these examples demonstrate, determining the best edit distance comparison can be very involved and be highly dependent on the data source and its processing. Ideally, the comparison would be based on empirical testing of the frequency of different types of data entry or scanning errors. In practice, we usually use a standard edit distance function or the built-in specialised comparisons of a data integration package.

3. Term or token-based comparisons

In these methods strings are broken down into 'tokens', which are then compared. The most obvious breakdown would be by words, so that strings with many similar words are ranked as more similar than ones with only one or two words in common.

Another breakdown is 'Qgrams', which consist of all sequences of Q characters in a string. For example, the set of 3-grams for "Statistics" is {"Sta", "tat", "ati", "tis", "ist", "sti", "tic", "ics"}. The similarity of two strings can be computed by determining how many of these 3-grams are in common between them. A relatively simple comparison score is the Jaccard coefficient. Given two sets of Qgrams *A* and *B*, the Jaccard coefficient *J(A,B)* is the ratio of the number of terms in common between the two sets divided by the number of terms which appear in only one set. Mathematically,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

This coefficient can be generalised by giving each term a weight. For example, since many last names end in 'son' the fact that two names have this term in common would not provide strong evidence of similarity.

There are many possible ways to compute term weights rigorously. A standard method is to compute inverse document frequencies (idf) for each term *T* for a certain variable in a given dataset as:

$$\text{idf} = \frac{\text{Number of units in the dataset}}{\text{total number of times } T \text{ occurs across all values of the variable}}.$$

Given a particular variable value *V*, the score for a comparison is then calculated by summing up:

$$\log(\text{tf} + 1) \times \log(\text{idf})$$

across all terms in *V*, where the term frequency (tf) is defined as the number of times term *T* occurs in *V*. This is a similar idea to the Fellegi-Sunter m and u probabilities.

Comparisons for date variables

Treating dates as numeric

A date can be thought of as a numeric value where the difference between two dates is simply measured in a common time unit – such as days or months. A tolerance can be set as with numeric comparisons and the same types of comparison can be used.

In some situations it may be useful to set penalties that are not symmetric. For instance, an analyst may decide there should be zero tolerance for the date on file B, being after the date on file A, but that a few days difference is acceptable if the date on file B occurs before that on file A. This may be useful, for example, where an event such as an address change may be subject to reporting delay and will always show up later on one dataset than the other.

Treating dates as character strings

In many datasets, date variables may be subject to similar errors as names or other character variables. A common example is to transpose years in dates, such as a birth date of '10-10-1968' being incorrectly recorded as '10-10-1986'. In datasets where these types of errors are known to occur, we would probably rank, '10-10-1986' as a more likely match to '10-10-1968' than '10-10-1985', even though numerically the 1985 date is closer. The decision about whether to compare dates using numeric or character comparisons will depend on the details of the date variables in the datasets being linked and their quality.

7.3.4 The m and u probabilities

The m and u probabilities can be defined in two different ways. Global m and u probabilities assume the probability is constant through all variable values. Value-specific m and u probabilities are probabilities that may contain variable value differences.

Use global u probabilities if you can assume the distribution of possible values within the field is (nearly) uniform. In practice, the linkage software may automatically estimate value-specific u probabilities to reflect the actual distribution of variable values in the dataset.

Use value-specific m probabilities for fields where some values are more reliable than others. However, global m probabilities are generally used, because it is expected that the values in a field are affected in the same way by what makes a field reliable or unreliable (eg mode of collection, maintenance practices).

The m probability is the probability that the fields agree given that the record pair is a match. It reflects how reliable the field is and is computed as 1 minus the error rate of the field. Because all fields are not equally reliable, you can expect m probabilities for different fields will vary. In practice, the error rates are generally not accurately known. Setting a high m probability value for a field forces a high penalty for disagreement in that field. Initially, when no estimates of the m probabilities are available, use the following:

- for most fields, 0.9
- for very important fields, 0.999
- for moderately important fields, 0.95
- for fields with poor reliability, 0.8 or less.

Experience shows that variables that are collected and maintained carefully by the source agencies have good m values (are reliable), whereas variables of less importance to them – that is, those not needed to support their core operational requirements – tend to

be less reliable. Where the law requires an event to be reported within a prescribed short period of time, event dates are reliable fields.

While there are theoretical approaches to modelling m values (eg Winkler, 1988), Statistics NZ uses an iterative approach. We do the first linking using an estimate for m that is based on what is broadly known of the variable's importance, or from previous experience.

We then estimate a new m value from the values for data that has been linked. The m probability is estimated by dividing the number of times the field values agree in a comparison, by the number of times the value participated in a comparison (excluding records with missing entries for the field of interest). This should be done when the analyst has confidence that most of the good links have been captured.

The u probability is the probability that the fields agree given that the record pair is not a match. This reflects the likelihood of a chance agreement. Assuming a uniform distribution for the values a field may take, u is estimated by 1/n, where n is the number of field values. For example, u probability for sex is estimated as 1/2, as sex takes two possible values. Similarly, for the variable month of birth, the u probability is estimated as 1/12.

7.4 Quality assessment of linked data

7.4.1 Setting the cut-off threshold

A trade-off exists between the level of false positive and false negative links. The objectives of the linking exercise are important when you determine cut-off thresholds. For example, if it is critical to avoid false links, then set the cut-off threshold higher, being mindful that some true matches will be missed.

The (non-negative) cut-off threshold is the composite weight value that separates links the analyst considers to be matches and those they don't. Record pairs with composite weights greater than or equal to the cut-off are regarded as links. Deciding on the cut-off value is a difficult task as the boundary is not clear-cut – even experienced analysts can produce significantly different linked outputs (Gomatam et al, 2002).

In practice, the cut-off is initially set at zero for a given pass, and is iteratively changed before proceeding to the next pass. After running the pass, examining a histogram of the weights will help you decide the cut-off score for the pass. Ideally, the frequencies of matched records trail off as the weights become lower, while the frequencies of unmatched records trail off as the weights become higher. This ideal situation produces a 'bimodal' distribution. The farther apart from each other the modes are, the better the discrimination between the matched and unmatched records. Figure 5 shows this scenario.

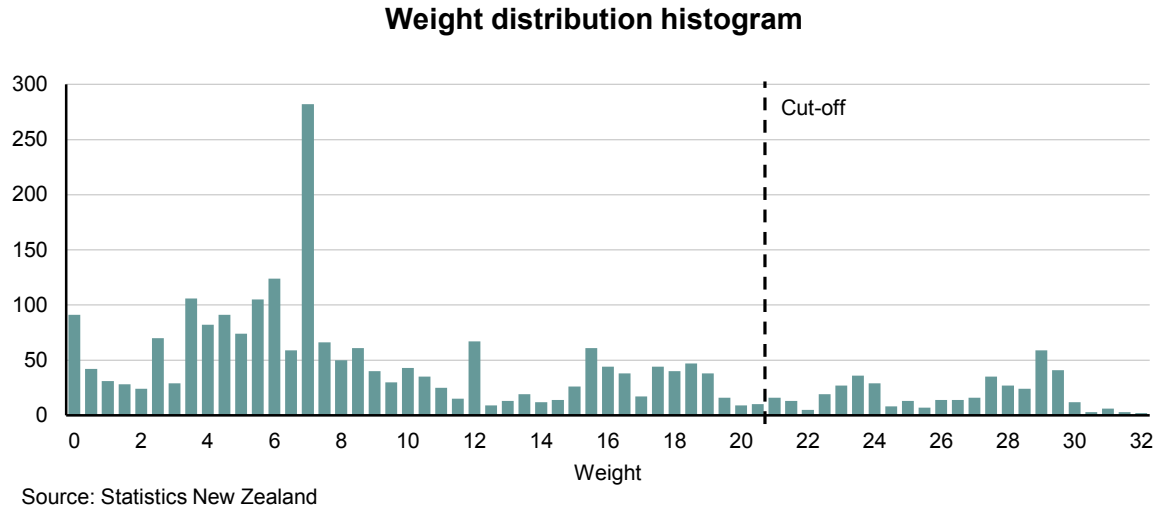
The distribution of weights is usually far more complex in real linking projects. Depending on the details of the comparison functions being used, the m and u probabilities chosen, and the contents of the datasets being linked, the distribution will often contain many apparent peaks and gaps rather than being an ideal smooth curve. Also, because some software does not make comparisons for records that have no chance of linking, comparisons with negative weights might not be included in the output.

To deal with these practical problems, to determine the cut-off weight you should produce a file of linked records to examine in detail. You can sort the file by weight in descending order. The record pairs with high composite weights represent good links. As the weight value lowers, the links become dubious. Examine the sorted record pairs for increasing patterns of field disagreements as the weights decrease, to determine an appropriate cut-off level for the pass. As you gain experience and familiarity with the data undergoing integration, you gain confidence and greater ease in setting cut-off scores.

Figure 10 shows a sample actual histogram of weights under non-ideal conditions. After visually assessing the file of linked records, the analyst set the cut-off score of 21 for this pass.

Note the multiple peaks and the not-so-distinct trailing frequencies near the chosen cut-off.

Figure 10



Source: Statistics New Zealand

A side-effect of adjusting the cut-off threshold is the possibility of creating duplicate pairs. One record in file A may form a pair with a weight above the chosen threshold for more than one record in file B.

Depending on the nature of your integration exercise, these may be treated as genuine duplicates, possibly for further review, and not be available in other passes. If you treat the linking as strictly one-to-one, then the record pair with the highest weight is taken as the link, and the rest become eligible for linking in the next pass. Where record pairs have the same weight, one can be randomly chosen as the link. Data integration software may have options for handling cases where duplicates with the same or different weights exist. Carrying out deduplication simplifies subsequent linking by generating confidence that no genuine duplicates exist.

7.4.2 False positives, false negatives, and link rates

False positives are record pairs erroneously deemed to be links despite being true non-matches. False negatives are true matches that remain unlinked.

Generally, there is no good method for automatically estimating error rates, so false positive rates are estimated by manually checking samples of linked records. In large datasets, analysing false positives can be time-consuming and it is often useful to group the linked data before selecting a sample.

For example, the passes constitute groups from which to draw samples for false positive analysis. Alternatively, new and different groups can be constructed for sampling purposes. Select and analyse samples from each group for false positives.

Clerical review is done by visually comparing the sample records. Although this method draws on subject-matter knowledge and other information, it still involves the reviewer's subjective view. By understanding where errors are most likely to occur in the datasets, you can target the sample to these areas, to improve the quality of the match. Several iterations of clerical review and adjusting match criteria may be necessary before a linked dataset is confirmed and you can calculate final false positive error rates.

If at least one of the files is expected to match completely and the false positive rate is low, then calculate the false negative rate as one minus the link rate (where the link rate for a given file is the number of linked records over total records). However, in other situations (eg when the integrated dataset is the union of two files) expected matches are unknown and the false negative rate is difficult to estimate.

7.4.3 Precision and recall

There are several ways to present the rates of erroneous links. Different measures can be easier to explain or more relevant in different situations. The most commonly used measures in data integration literature are precision and recall, which are defined as:

Precision is the proportion of all links made that are correct links.

Recall is the proportion of correct links made out of the total number of potential correct links existing in the data.

Precision and recall are related to the false positive rate and false negative rate, respectively. Like the false positive rate, precision is found by sampling and clerical review of the links – to decide if they are true links or not. Recall, like the false negative rate, is harder to estimate as the number of expected matches in the union of two datasets is often unknown.

7.4.4 Measurement error in integration

Measurement error affects inference – it can lead to bias in estimation, which can be severe. Best-practice procedures in data analysis examine the data being used for measurement errors; known measurement error properties are incorporated into the analysis (Chesher & Nesheim, 2004).

The measurement error processes arising from probabilistic record linkage are complex and non-standard. Chesher and Nesheim list these causes of measurement error in data linking:

- units incorrectly linked, so that data from one unit is incorrectly associated with another unit (false positive links)
- in many-to-one linking, statistics computed using only a few sub-units are used to measure characteristics of all sub-units
- in many-to-one linking, characteristics of sub-units are inferred from features of major units (and vice versa).

They go on to say that, practically, measurement error is inevitable and since the potential effects are so damaging, an analyst should avoid using data-linking procedures that are likely to generate large measurement errors.

The first step in estimating the quality of linked datasets is often to estimate rates of false positives and false negatives. In record linkage projects we've done previously, quality measurement has often focused on these two dimensions – to minimise false positive links.

7.5 Adding data over time

Adding data over time may impose additional difficulties for the data integration exercise. You need policies that account for changes or updates in the data for a given time period. Agencies providing administrative data commonly add, delete, modify, or update their records over time. For example, an ACC claim can be made at any time by a claimant after an accident has occurred, resulting in late claims, or an agency may update its records to account for new information such as a change in address or family name.

As a consequence, data received in one time period may not be the complete dataset for that period. A policy that imposes time cut-offs on data that arrive late is needed, to ensure lateness doesn't affect the integration exercise, and also doesn't result in major adjustments to the outputs over time.

Integrating data from multiple data sources can affect data integration. Difficulties exist when the definitions of reference periods vary between data sources – data received from different data providers must refer to the same time period. Agencies may use different dates to refer to different parts of the process they use to gather records. For example, a date for when a record is received by the agency, perhaps initially in paper form; a date for when a record is entered into a computer system; a third date for when a record is registered or accepted by the agency; and a fourth for the time period the record actually refers to.

You need to understand the nature of the data and have discussions with the data providers to ensure you receive the correct data for your specified time period.

Another potential issue is the carry-over effect of false positives in an ongoing production environment. These cases require extra caution – to keep the number of errors (false positives) down in each step of the integration period, for any given time period, and minimise the errors being propagated. Repeating the linking for every time period (including links already made) will also resolve the carry-over issue. However, this means links may change from time to time.

Glossary

The glossary lists some commonly used terms in data integration, and defines how they are used by Statistics NZ.

Glossary of common data integration terms

Term	Definition
Agreement weight	A numeric value assigned when there is agreement on a particular field for a pair of records being compared. See Disagreement weight.
Array	A number of single fields may be combined into a single array using record-linkage software. If there are several fields that contain similar information (eg alternative name fields), using arrays can reduce the number of cross-comparisons that must be made.
Bias	A data linkage method may be biased if there are systematic errors in the links created. If the linkage method is biased, then results from analysis using the linked data may differ systematically from the true results.
Block	Blocks are groups ('pockets') of files to be linked that have some information in common. Records are only compared with others in the same block. Using blocks reduces the number of comparisons that must be made.
Blocking variables	Variables used to divide a file into blocks. See Block.
Cut-off weight	The composite weight at or above which all record pairs are linked and below which all record pairs are not linked.
Comparison function	A means by which a decision can be made on whether two variables being compared fully agree, partially agree, or fully disagree.
Composite weight	The sum of the agreement weights for all linking variables that agree (positive values) and the disagreement weights for all linking variables that disagree (negative values). The composite weight measures the relative likelihood that the two records are in fact a true match. See Total weight, weight.
Data integration	The combination of data from different sources about the same or a similar individual or unit. Data integration at the micro level is synonymous with record linkage.
Deduplication	Process of identifying records belonging to the same unit (eg person). Once identified, duplicate records may be removed or combined with a record nominated as the master record.

Deterministic (exact) record linkage	Linking records belonging to the same unit through a unique identifier.
Disagreement weight	A numeric value assigned when there is disagreement on a particular field for a pair of records being compared. See Agreement weight.
False negative link	Two records that should have been linked because they correspond to the same unit (ie they are a true match) but were not linked.
False negative rate	The proportion of true matches on a file that have not been linked.
False positive link	Two records that were linked in error and do not correspond to the same unit (ie they are a not a true match).
False positive rate	The proportion of links that are false positives.
Global m and u probability	These two probabilities assume the probability is constant for all values of the variable. For example, if u probability is set to 0.03 for the year of birth variable, this applies for all years. See m probability, u probability.
Integrated dataset	The dataset resulting after record linkage has taken place.
Integration input dataset	A dataset containing data that has been edited, parsed, and standardised in readiness for integration.
Integration unit	Level at which the data is integrated. May not be the same as reporting unit.
Link	A decision that two records belong to the same unit. See non-link, match, non-match.
Link file	The file output from record linkage that lists all the linked pairs. See link, linked.
Linkage	See record linkage.
Linked	The status of a record that passed through the integration process and was linked to a record from the other file.
Linking variables	Variables used to compare two records – includes blocking variables and matching variables.
m probability	The probability that a field (in record linkage) has the same value on both files, given that the records being compared truly belong to the same individual/unit. It measures how reliable the field is. See u probability.
Matching variables	Variables used to compare two records that fall within the same block, to see how likely it is that both belong to the same unit. See block, blocking variables, linking variables.
Match	Two records are a ‘match’ when they relate to the same unit. Also called a true match.

Memorandum of understanding (MOU)	A formal voluntary agreement between two or more parties that seeks to achieve mutually agreed outcomes through the parties' efforts.
Microdata	A file that has a record for each unit (unit record data). The lowest level of data available.
Non-link	A decision that two records being linked do not correspond to the same unit. See Link, True Match, True non-match.
Parsing	Process of splitting a text string into a series of variables (eg full name splits into first names and surnames).
Pass	One iteration of a record linkage process, using a particular set of blocking and linking variables. See block, blocking variables, linking variables.
Probabilistic record linkage	Methodology based on the relative likelihood that two records belong to the same unit, given a set of similarities/differences between the values of the linking variables (eg name, date of birth, sex) on the two records. See deterministic record linkage.
Record linkage	Combining data from different sources about the same individual or unit, or a similar individual or unit, at the level of individual unit records. Synonymous with data integration at the micro level.
Reporting unit	Level at which the source data is provided. May not be the same as integration unit.
Sensitivity	The proportion of all records on one file with a match in the other file that was correctly accepted as a link.
Specificity	The proportion of all records on one file with no match in the other file that was correctly not accepted as a link.
Service level agreement (SLA)	A formal voluntary agreement between two or more parties that seeks to achieve a mutually agreed level of services through the parties' efforts.
Source data	The original dataset as received from the data provider.
Standardisation	Process of changing the formats of variables to make them comparable across different datasets.
Statistical matching	Process that occurs at the unit-record level but does not necessarily link records of the same person. In statistical matching, a unit record for one individual is linked to a record or records for similar individuals in other datasets on a probabilistic basis. See stochastic matching.
Stochastic matching	Matching groups from two different datasets based on similar characteristics, with the assumption that such people will act in the same way. Useful for creating synthetic datasets. See statistical matching.
Total weight	The sum of the agreement weights for all linking variables that agree (positive values) and the disagreement weights for all linking variables that disagree (negative values). Synonymous with composite weight.

True match	Two records that truly do correspond to the same unit. See link, match, non-link, true non-match.
True non-match	Two records that truly do not correspond to the same unit (eg two different people). See link, non-link, true match.
u probability	The probability that a field (in record linkage) has the same value on both files, given that the records being compared do not belong to the same individual/unit. It is a measure of how likely the field is to agree by chance on non-linking record pairs. See m probability.
Unique identifier (UI or UID)	A variable that uniquely identifies a person, place, event, or other unit.
Unlinked	A record that passed through the integration process and was not linked to a record from the other file.
Weight	A numeric value assigned to a pair of records compared during integration on the basis of the similarity of the linking variables. See composite weight.

References

Ascential Software (2002). *Integrity SuperMATCH Concepts and Reference Guide Version 4.0*, 5–17.

Baxter, R, Christen, P, & Churches, T (2003). *A comparison of fast blocking methods for record linkage.* CMIS technical report 03/139, First Workshop on Data Cleaning, Record Linkage and Object Consolidation, KDD 2003, Washington DC.

Brackstone (1999). Managing data quality in a statistical agency. *Survey Methodology, Dec 1999, Vol 25*(2), 139–149.

Cabinet meeting minutes CAB (1997) M 31/14 [electronic copy unavailable].

Chesher, A, & Nesheim, L (2004). *Review of the literature on the statistical properties of linked datasets*, Report to the Department of Trade and Industry, United Kingdom.

Community Services Ministers' Advisory Council (2004). *Statistical data linkage in Community Services data collections*, Australian Institute of Health and Welfare, Canberra.

CULT project (2014). *Review of Record Linkage Software*. ABS, Istat, and Statistics NZ.

Fellegi, I, & Sunter, A (1969). A theory of record linkage. *Journal of the American Statistical Association, 64*, 1,183–1,210.

Gill, L (2001). *Methods for automatic record matching and inkage and their use in national statistics.* National Statistics Methodological Series No 25, Office of National Statistics, United Kingdom.

Gomatam, S, Carter, R, Ariet, M, & Mitchell, G (2002). An empirical comparison of record linkage procedures, *Statistics in Medicine, 21*, 1,485–1,496.

Gu, L, & Baxter, R (2004). *Adaptive filtering for efficient record linkage*. 2004 SIAM International Conference on Data Mining Conference Proceedings, Florida.

Gu, L, Baxter, R, Vickers, D, & Rainsford, C (2003). *Record linkage: current practice and future directions*. CMIS technical report no. 03/83, CSIRO Mathematical and Information Sciences, Canberra.

Jaro, M (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*,*14*, 491–498.

Koudas, N, Sarawagi, S, and Srivastava, D (2006). Record linkage: Similarity measures and algorithms [PDF, 130p]. Proceedings of the 2006 ACM SIGMOD/PODS international conference on management of data.

Newcombe, H, Kennedy, J, Axford, S, & James, A (1959). Automatic linkage of vital records. *Science, 130,* 954–959.

Parliamentary Counsel Office (1993). Privacy Act 1993. Available from www.legislation.govt.nz.

Privacy Commissioner (nd). Codes of practice. Available from www.privacy.org.nz

QualityStage (2003). *Match concepts and reference guide.* Version 7.0, chapter 5, 1–34.

Statistics New Zealand (1998). *Final report on the feasibility study into the costs and benefits of integrating cross-sectoral administrative data to produce new social statistics*. (unpublished report available on request.)

Statistics New Zealand (1999a). *Confidentiality protocol*. (unpublished report available on request.)

Statistics New Zealand (1999b). *Statistics and the Privacy Act 1993*. (unpublished report available on request.)

Statistics New Zealand (2002a). *Guidelines for writing a technical description of a record linkage project*. (unpublished document.)

Statistics New Zealand (2002b). *Meta information template for description and assessment of administrative data sources.* (unpublished report available on request.)

Statistics New Zealand (2002c). *Pro forma privacy impact assessment report – data integration projects (draft).* (unpublished report available on request.)

Statistics New Zealand (2003a). *Guidelines for peer review of the technical description of a record linkage project*. (unpublished document.)

Statistics New Zealand (2003b). *Injury statistics project pilot: quality report part two – assessment of bias.* (unpublished document available on request.)

Statistics New Zealand (2005a). *Data integration policy guidelines.* (unpublished report available on request.)

Statistics New Zealand (2005b). Proposed methodology for estimating undercounting of vehicle-related injuries in NZ". (unpublished report available on request.)

Statistics New Zealand (2012a). Data integration policy. Available from www.stats.govt.nz

Statistics New Zealand (2012b). *Methodological standard for metadata content and associated guidelines*. (unpublished documents available on request)

Statistics New Zealand (2013). Statement of Intent 2012–17 (Budget 2013), Wellington: Statistics New Zealand. Available from www.stats.govt.nz.

Statistics New Zealand (2014). Privacy impact assessment for the serious injury outcome indicators. Available from www.stats.govt.nz.

Taft, R (1970). Name search techniques. New York State Identification and Intelligence System. See http://www.dropby.com/NYSIIS.html for an explanation and online implementation of the NYSIIS encoder.

Winkler, WE (1988). *Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage*. Proceedings of Survey Research Methods Section, American Statistical Association, 667–671.

Winkler, WE (1995). *Matching and record linkage*. Business Survey Statistics, 355–384.

Appendix: Guidelines for writing a technical description of a record linkage project

Note: This appendix is a combination of internal documents, Statistics New Zealand (2002a) and (2003a). Redundancies and details that are only relevant to internal processes were removed and out-of-date information was updated.

Contents

1. Purpose of the technical description

Any matching exercise should be accompanied by full documentation of the method used. This can be thought of as a 'technical description' of the matching methodology. It has two main uses:

- to allow a peer review of the methodology
- to provide a record of what has been done for the future.

A peer review is needed in order to provide assurance that the project has been carried out according to best practice. The peer review should not be a complete recreation of the matching, but a review of the process. See below for guidelines for reviewers.

The review should ideally be done before the linked data is handed over to clients or used for analysis, so that any improvements in methodology suggested by the reviewer can be carried out. This might mean a two-stage process, where the first results are essentially a trial. The match method and results are reported and reviewed, any modifications carried out, and the final linked file handed over. This may not be possible in practice, and in that case improvements can be noted for future. Whatever is decided, the documentation and peer review of the matching methodology should be included as tasks in planning the project, and enough time and resources allowed for them.

It is vital that full details of the matching method and results are written down and available for the future. These reports provide the formal documentation of what has been done, both for future matching with the same data sources, and as examples for other matching projects.

If methods are well documented, then capability can be greatly enhanced. Good documentation reduces the likelihood of making mistakes, and means that future work can build on past work, rather than rebuilding it from scratch.

2. Contents of the report

The report should be self-contained and understandable on its own, without the need to also read other papers. It should include:

1. A stand-alone summary for the 'interested lay person'
2. The reason for the integration
3. A description of the input datasets
4. A description of the matching methods used

5. Details of the results achieved.
6. A description of the final linked file

and if appropriate:

7. Recommendations of things to do differently next time
8. Other options that were thought of (or tried) and rejected.

Important information to include in each section

1. Summary: There should be a short summary of the matching methodology at the beginning for the reader who doesn't want to go through everything, and which can be copied for other documents that report on the project. This should be accessible to the interested lay person: keep technical jargon to a minimum and explain any technical words that are needed.

The next two sections should relate to what is relevant for the match, and do not necessarily need to be full detailed descriptions.

2. The reason for the integration: The matching needs to be put in context for a reader who otherwise knows nothing about it. It is important to state what the objectives of the whole project are, otherwise one can't make any judgement about whether the process is adequate, or if the quality of the match is good enough; that is, fit for the intended use.

3. A description of the input datasets: This should include where they come from (such as an internal collection, external agency, or some other source) and what the primary purpose of each input dataset is. The focus is on the variables available for the match. Include any relevant information about how the variables get to the source data file, such as things that might lead to anomalies in the coding, or very high-quality coding. There should be an indication of the editing or standardisation carried out before the variables are used in the matching. If extensive editing has been used it may be summarised here and placed in an appendix, or referenced.

4 and 5. Matching methods and results: These sections should contain enough detail for a statistician to peer review the methodology. Important details include reasons for decisions to proceed in a particular way, blocking and matching variables used in each pass, m and u values used and how they are calculated, cut-off values, and tables of match rates. Some detail can go in an appendix so the body of the paper is readable.

8. Other options: The report should stay reasonably focussed on the final decisions. Discussion of various avenues that were explored before a final decision was made should only be a summary and needs to be concisely written. A reference to more detailed work that has been written up elsewhere may be included.

3. Presentation

Title: Use a meaningful title, such as *Technical description of the record linkage methodology for student loans.*

Authorship: For external audiences, it is best to include a general contact address (for Statistics New Zealand, info@stats.govt.nz) to make sure that the right contact can be found, since staff often leave or change projects.

Referencing: Treat the technical description as a formal documentation with full referencing. If information from other documents is needed then the key points should be summarised in the report, with a named source. References should be at the end of the main paper.

Direct links to internal documents should be avoided because the technical description often needs to be printed, sent to external people, or placed on the web. As noted above, the document should be self-contained.

Computer output: Programs and output are often necessary, but the amount of detail kept in the main text should not make the paper unreadable. Large tables or long computer listings should be in an appendix, not in the middle of the text. The text can highlight main points, or summarise in a small table, and refer to a longer program or output in the appendix.

Structure: The list of contents above is a guide to what should be included, but is not meant as a strict template. However, the report should be have a clear structure, with numbered sections, tables and figures numbered and with headings, and a contents list.

4. Exclusions from the technical description

Detailed investigation into the quality of the resulting linked files should be reported in detail separately. A summary of findings or reference to the quality investigations may be included. Detailed information about the source datasets that is not relevant to the methodology, such as quality information about variables not used for matching, should also be documented clearly and separately.

5. Guidelines for peer reviewers

As explained above, peer review is an important part of the project and its documentation. A reviewer needs to examine the technical aspects of the project and make sure that there are no methodological issues. They should also check the technical description includes all necessary information and can be understood by somebody else in the future.

The reviewer can assess the report on the following criteria:

1. Completeness

Are all aspects covered? See the suggested content in the guidelines above.

2. Presentation

There is no requirement for a particular template to be used, as data integration projects can have quite different issues and challenges. Each writer has their own particular style, but the technical writing should be:

- well-structured
- clearly written
- have technical concepts explained
- include references.

Appendixes should be used as necessary for large tables, programs, and so on to improve the readability of the report.

3. Sound methodology

The method developed for the data matching should be technically sound. The reviewer needs to assess whether:

- the type of match is explained and is appropriate (eg, exact/probablistic; de-duplication, union or intersection of data sources)
- the relative importance and impact of false negative and false positive errors is explained (as far as possible)

- variables from the source data have been edited or derived appropriately
- good use has been made of blocking and matching variables, and of different passes
- any underlying assumptions about the data that influence the matching strategy are stated
- m and u values are appropriate
- cut-off values are appropriate. Cut-offs should reflect the relative importance of false negative and false positive errors; there needs to be ‘enough’ clerical review of possible links around the cut-off values so there is an understanding of what types of records are being linked, and what are not being linked.

4. Main issues highlighted

Key problems are discussed, and either resolved satisfactorily, or signalled as being outstanding.