



Australian Government

Attorney-General's Department



## IMPROVING THE INTEGRITY OF IDENTITY DATA

DATA MATCHING

Better Practice Guidelines

2009



IMPROVING THE INTEGRITY OF IDENTITY DATA

DATA MATCHING

Better Practice Guidelines

2009

## **Acknowledgement**

The assistance of members of the Commonwealth Data Matching Working Group is acknowledged in the preparation of this document and the data matching principles.

Whilst all reasonable care has been taken in the preparation of this publication, no liability is assumed by the Commonwealth of Australia for any error or omission.

## Contents

|  |           |
|--|-----------|
| <b>1. OVERVIEW</b>   | <b>4</b>  |
| <b>IMPROVING THE INTEGRITY OF IDENTITY DATA</b>  | <b>4</b>  |
| <b>2. PRINCIPLES</b>   | <b>5</b>  |
| <b>DATA MATCHING IMPROVEMENT CYCLE</b>   | <b>6</b>  |
| <b>3. GUIDANCE</b>   | <b>7</b>  |
| <b>CATEGORY 1 - DATA TRANSFER</b>  | <b>7</b>  |
| Principle 1 - Improve the process and form of data transfer  | 7         |
| <b>CATEGORY 2 - DATA PRE-PROCESSING</b>  | <b>8</b>  |
| Principle 2 - Retain originally supplied name values but consider the use of standardising approaches to overcome name inconsistencies | 8         |
| Principle 3 - Include a control group  | 9         |
| Principle 4 - Make greater use of deceased status indicators   | 10        |
| <b>CATEGORY 3 - SOLUTION DESIGN (INCLUDING ALGORITHM)</b>  | <b>11</b> |
| Principle 5 - Use name, date of birth, address in the algorithm design   | 11        |
| Principle 6 - Include historical name and address details for each record  | 12        |
| Principle 7 - Ensure the use of a flexible matching algorithm  | 13        |
| Principle 8 - Increase confidence in identity information by confirming data with a number of sources                                  | 17        |
| <b>CATEGORY 4 - MATCH RESULTS AND INTERPRETATION</b>   | <b>17</b> |
| Principle 9 - Use profile groups to stratify data matching results   | 17        |
| Principle 10 - Combine human involvement in the analysis of data matching results when flexible matching has been employed             | 19        |
| <b>CATEGORY 5 – IMPROVED DATA QUALITY</b>  | <b>20</b> |
| Principle 11 - Identify and quantify data integrity issues which affect the ability to match records or data                           | 20        |
| Principle 12 - Employ address validation techniques  | 21        |
| <b>CATEGORY 6 - KNOWLEDGE SHARING AND LONGER TERM DATA MATCHING DEVELOPMENT ISSUES</b>   | <b>22</b> |
| Principle 13 - Share expertise, particularly in specific data matching subject matter areas  | 22        |
| Principle 14 - Identify and evaluate commercial products   | 22        |
| Principle 15 - Apply the lessons learned from data matching to enrolment processes   | 23        |
| Principle 16 - Undertake research and development involving data matching  | 24        |
| Principle 17 - Consider the use of new types of data for data matching   | 24        |
| <b>4. GLOSSARY</b>   | <b>25</b> |

## Attachments

- A. Case Study - Flexible Matching and Standardising
- B. Data Definitions - AS4590 - Standard for Name and Address Data Transfer
- C. Data Matching Profile Template

# 1. Overview

Improving the integrity of identity data for individuals is a key element of the National Identity Security Strategy, as agreed by the Council of Australian Governments in 2007 in response to a revised assessment of Australia's identity management system. Improving the integrity of identity data of individuals also complements efforts to improve the security of identity information, enrolment and systems for verifying the integrity of key identity documents.<sup>1</sup>

Data matching allows information from a variety of sources to be brought together and applied to a range of uses. It is a key tool in ensuring the quality and accuracy of information (i.e. integrity of identity data), which helps to ensure the security of Australia's identity management system, and can be used to facilitate more streamlined and efficient enrolment processes. Variations, however, in the way information is collected, captured, stored and maintained, can impact on the quality of identity data, and can impact on dealings between individuals and government. They can also increase the potential for error, misuse or fraud.

This document incorporates a number of principles to consider in the design, build and analysis of data matching surrounding identity information. The principles range from direction on particular techniques and applications through to encouraging, where possible, the use of standards and to sharing relevant knowledge and expertise. The principles are supported with guidance, which can be drawn on in applying data matching techniques and applications, and to improve the integrity of identity data.

The information contained in this document has been developed with the assistance of the Commonwealth Data Matching Working Group (DMWG), and is largely based on methods applied by experienced data matching practitioners within government. It is intended as a guide only and users should seek professional advice as to their specific risks and needs.

Agencies should note that this document does not address the various legal and policy requirements that will govern data matching activities. In planning and conducting data matching, agencies should refer specifically to the:

- *Privacy Act 1988*
- *Data-matching Program (Assistance and Tax) Act 1990*
- Office of the Privacy Commissioner Guidelines on the Information Privacy Principles
- Office of the Privacy Commissioner Guidelines for the use of Data matching in Commonwealth Administration.

## **IMPROVING THE INTEGRITY OF IDENTITY DATA**

Accurate, up-to-date and complete identity information is essential for several reasons, not least so that government can provide quality services to the public. Some databases or registers can contain inaccurate, out-of-date or false identity information. Records containing names, addresses or dates of birth may be wrong or inconsistent for several reasons. Some of the reasons for this include:

- changes to individuals' personal details over time
- errors in recording and transcribing personal information during enrolment or subsequent updates of identity information
- variances in naming conventions, whether in a social or technical context.

Poor data quality has a number of undesirable consequences. It can lead to situations where records that should match do not (false negatives) or records match that should not (false

---

<sup>1</sup> NISS Report to COAG (April 2007)

positives). It can also impact on the level of confidence that others can have in systems designed to verify the integrity of key identity documents, or to be able to confirm with any certainty that, electronically, a person is who they claim to be. Some of the consequences of this include:

- people not receiving something to which they are entitled
- fraudulent and other criminal activity not being detected
- people being registered with agencies multiple times
- inefficient delivery of services, including delay and unnecessary imposition
- redundant data and inefficient work practices.

## 2. Principles

The data matching principles and guidance developed by the DMWG recognise that no single agency is the authoritative source on data matching, and that agency-centric approaches in the way that data are defined and stored can be problematic. Isolation and the lack of opportunity for any ‘cross-pollination’ of ideas can also result in a constant need to ‘reinvent the wheel’. Even those agencies well practiced, ‘tooled up’ and capable of developing and applying extremely complex and intricate data matching applications, can learn from the experience of others or share their own knowledge or expertise.

There are 17 principles. All are aimed at assisting government agencies to adopt better practice in data matching, with a view to improving the integrity of identity data.

| <b>Data Matching Principles</b> |   |
|---------------------------------|---|
| <b>Principle 1</b>              | <b>Improve the process and form of data transfer</b>  |
| <b>Principle 2</b>              | <b>Retain originally supplied name values but consider the use of standardising approaches to overcome name inconsistencies</b> |
| <b>Principle 3</b>              | <b>Include a control group</b>  |
| <b>Principle 4</b>              | <b>Make greater use of deceased status indicators</b>   |
| <b>Principle 5</b>              | <b>Use name, date of birth, address in the algorithm design</b>   |
| <b>Principle 6</b>              | <b>Include historical name and address details for each record</b>  |
| <b>Principle 7</b>              | <b>Ensure the use of a flexible matching algorithm</b>  |
| <b>Principle 8</b>              | <b>Increase confidence in identity information by confirming data with a number of sources</b>                                  |
| <b>Principle 9</b>              | <b>Use profile groups to stratify data matching results</b>   |
| <b>Principle 10</b>             | <b>Combine human involvement in the analysis of data matching results when flexible matching has been employed</b>              |
| <b>Principle 11</b>             | <b>Identify and quantify data integrity issues which affect the ability to match records or data</b>                            |
| <b>Principle 12</b>             | <b>Employ address validation techniques</b>   |
| <b>Principle 13</b>             | <b>Share expertise, particularly in specific data matching subject matter areas</b>   |
| <b>Principle 14</b>             | <b>Identify and evaluate commercial products</b>  |
| <b>Principle 15</b>             | <b>Apply the lessons learned from data matching to enrolment processes</b>  |
| <b>Principle 16</b>             | <b>Undertake research and development involving data matching</b>   |
| <b>Principle 17</b>             | <b>Consider new or future sources of identity data</b>  |

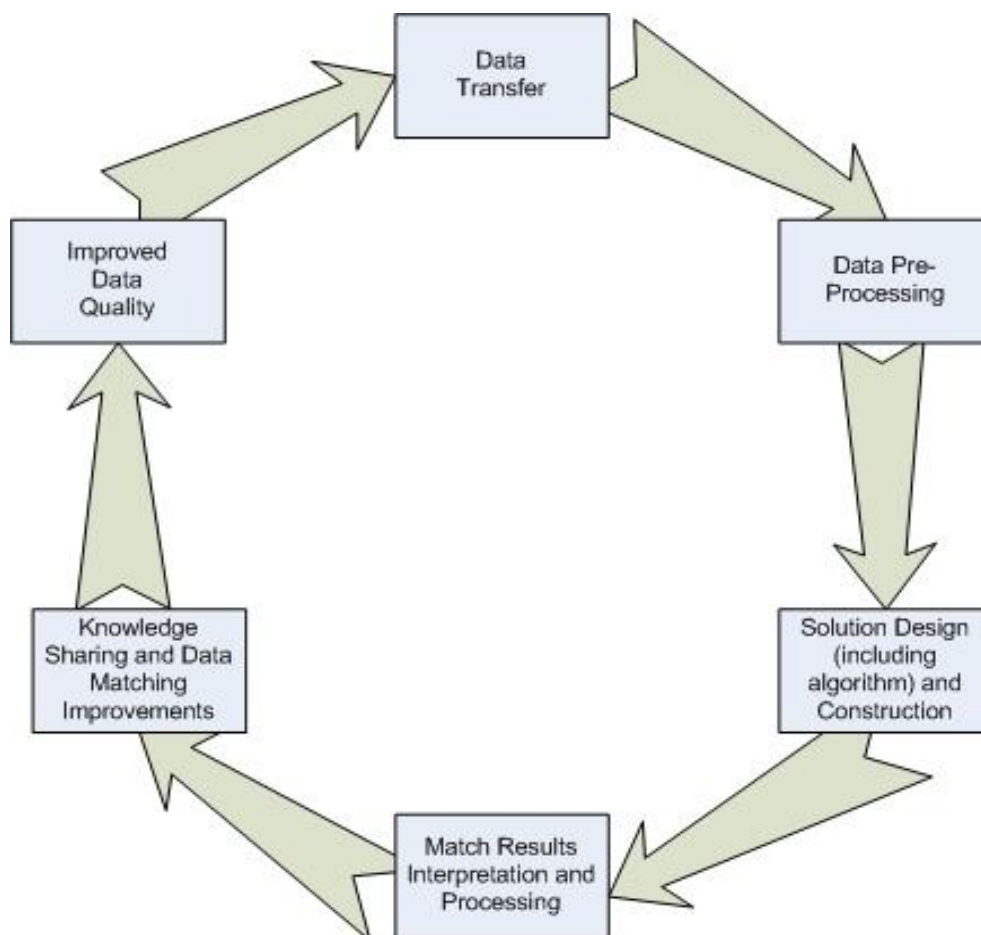
## ***DATA MATCHING IMPROVEMENT CYCLE***

The data matching principles can be applied in the various stages of a data matching cycle. These stages include:

- Data transfer
- Data pre-processing
- Solution design (including algorithm) and construction
- Match results (interpretation and processing)
- Knowledge sharing and data matching improvements
- Improved Data quality

With the completion of each cycle, the standard of data matching should improve as should the quality of identity data. This recursive cycle of improvement is illustrated below in the following diagram.

**Data Matching Improvement Cycle**



### 3. Guidance

In depth commentary and analysis on each of the data matching principles, as grouped within each stage of the data matching improvement cycle follows .

#### ***CATEGORY 1 - DATA TRANSFER***

##### **Principle 1 - Improve the process and form of data transfer**

###### **Data Transfer**

Overseas examples of physical data files being lost, misplaced or stolen emphasise a current exposure to risk involving the physical transfer of data from one agency to another.

- In the UK in 2007 CDs containing the personal details of 7.25 million families (some 25 million people) went missing while being sent by courier. The CDs contained details such as dates of birth, addresses, bank accounts and national insurance numbers for individuals claiming child benefit. Although password protected, the data on the CDs were not encrypted. The loss did not become apparent for three weeks.
- In the US, also in 2007, the Georgia Department of Community Health reported that a computer disk containing names, birth dates and Social Security numbers of 2.9 million Medicaid and children's health care recipients went missing during shipment by an IT services company.

Given the inherent risks associated with the physical transfer of information, agencies not presently doing so, but have the capacity, should consider using government developed site-to-site data transfer solutions such as ICON and FedLink.

ICON (Intra Government Communications Network) is a communications system that provides dedicated point-to-point links for Australian Government agencies based in Canberra.

FedLink is an encryption tool that allows data transmissions between members by encrypting data at the sender's gateway and de-crypting the data at the receiver's gateway. This prevents data being deciphered if it is intercepted while traveling over the public Internet. In the event of a failure, it does not travel at all.

In the absence of such approaches, and with due consideration of the security classification of the data involved, agencies should ensure that care is taken to ensure the security of data physically moved between themselves and other agencies. For example, data should be encrypted to appropriate standards, password protection should be considered and data should be passed by safe hand to any receiving agency using approved couriers meeting appropriate security standards.

###### **Data Format**

Data standards provide agreed definitions and formats of data common to different data users. They enable data to be more 'shareable' by increasing data compatibility, improving the consistency and efficiency of data collection, and reducing redundancy.

Where possible agencies should use recognised data format standards. Standards Australia established 'AS4590 Interchange of Client Information' which, amongst other things, specifies particular attributes to use when dealing with names and addresses. It was developed so that Australian industry, community and governments can avoid the need to develop multiple

variations of name and address fields in electronic data and transaction systems. Working within a standard allows for the avoidance of duplication, which can lead to error and the need to reformat data.

As an example, the AS4590 standard specifically details how to identify client information for the purpose of data interchange between organisations, and allows for more uniformity in the identification of clients by person and organisation including specifications around the details of person, organisation, telephone and address. AS4590 specifically includes requirements for the following data items, which allow for interchange of client information:

- (a) identification number - Australian Company Number (ACN), Australian Registered Body Number (ARBN), Customer Reference Number (CRN) and the like
- (b) name of client -
  - (i) person - title/suffix/family/given including alias, preferred, maiden, legal, professional/stage; and
  - (ii) organisation - company, partnership, trading
- (c) date of birth
- (d) sex
- (e) marital status (partner relationship)
- (f) occupation
- (g) country of birth
- (h) country of citizenship
- (i) industry
- (j) organisation type - business, non-profit bodies
- (k) telephone number details, facsimile, electronic addresses
- (l) address details - physical and postal

More detail in relation to Person, Person Name and Address attributes is contained in **Attachment B - Data Definitions - AS4590 - Standard for Name and Address Data Transfer**.

Some additional questions to consider in relation to working with AS4590, or any other standard, include:

- are the data being stored already standardised?
- are the data being stored already AS4590 compliant?
- if not, does the owner agency have the expertise to convert them?
- will adoption of standards result in the loss of detail in the search for commonality?

## ***CATEGORY 2 - DATA PRE-PROCESSING***

### **Principle 2 - Retain originally supplied name values but consider the use of standardising approaches to overcome name inconsistencies**

Name inconsistencies for the same individual across different databases can occur for numerous reasons. Some of the reasons include:

- the enrolment process may have been flawed and name details transcribed incorrectly, e.g. mis-spelling, “name swapping” (first name recorded as middle name and vice versa)
- many agencies have implemented data file designs based on the traditional Anglo-Saxon naming convention, where an individual has a first name, a middle name (or names) and a family name. This construct does not always lend itself well to names of different

ethnicities. The result is that the same name can be treated in different ways by different agencies and produce inconsistent final forms across agencies

- individuals choose to be known by different names with different agencies, resulting in non-consistent data values for the same individual. For example, John may have a passport in the name of 'John' but his driver's licence may state 'Johnny'.

In formulating approaches to these aspects of data matching, it is important to ensure that the original and complete name values (as they were initially recorded) are retained. The 'unprocessed' format or 'raw' nature of the name may or may not be the best format for the available data matching techniques and technologies. Some commercial data matching software solutions are designed to use the raw name values.

Agencies using commercial matching solutions alone will likely not need to consider the use of standardised values. For others, both standardised and raw values may be used. Standardising does not replace raw name data. Standardising produces new field values which can be used in the matching process to help avoid subsequent non-matching resulting from the existence of inconsistent data.

Standardising may involve the removal of non-alphabetic characters like hyphens, spaces and apostrophes to produce a 'standard format'. As an example, in instances where "OConnor" would normally not match with "O'Connor", standardising would result in a record in each file with the value "OConnor" which would then produce matches.

Standardising might employ applications which enable the successful matching of given name values that although not character-consistent, relate to the same name. In many cases, this can be attributed to the "anglicising" of non-Anglo-Saxon names such as Jo, Joseph, Joe, Giuseppe. The applications can be built 'in-house' (Centrelink and the Australian Electoral Commission have developed such applications) or be part of a commercial product .

It should be remembered that it is the role of an agency's business experts to decide which names are to be accepted as equivalent in any standardising solution. Equally, business experts should provide advice to technical staff in the use of commercial matching products in determining what constitutes matching values.

Agencies considering standardising name values may also wish to consider how to best interpret the types of results obtained with this approach. For instance, a match obtained using raw data may be considered of higher quality than one obtained with standardised data due to the inherent data consistency. Consequently, agencies may wish to use 'blended' approaches, where both raw and standardised values can be matched and the results ranked or rated (i.e. a 'cascading' interpretation of results).

Further information on the use of standardised data values is contained in **Attachment A - Case Study - Flexible Matching and Standardisation**.

### **Principle 3 - Include a control group**

The use of a control group<sup>2</sup> of records can assist in the development of data matching applications and in interpreting the results of data matching activity. By including a control group with known characteristics in the data passing through the data matching application and observing the results, the effectiveness of the application can be reviewed and refined. This concept is also useful when paired with **Principle 9 - Use profile groups to stratify data matching results**.

---

<sup>2</sup> Refer Glossary

## **Control Group Example**

A recent multi-agency cross data matching exercise demonstrated the potential usefulness of control groups in data matching. The aim was to help identify ways that data matching could be used to improve the integrity of identity data stored by government agencies; this includes the identification of fraudulent identities.

The inclusion of records of 65 known identity fraud cases in effect created a control group within the greater sample file of data matched against agency records. Based upon the degree of correlation between matching records (or lack of correlation, if no matching record was found), records in the sample, including the control group, were allocated to particular profile groups.

The combination of using a control group in the data and allocating the records to profile groups based on the match results was most informative. Approximately 86 per cent of the known fraudulent identities included in the sample group were allocated to two particular profile groups, though the great majority were allocated to just one of them. This result provided indications regarding which profile groups represented the highest risk of identity fraud. Future data matching efforts intended to identify potentially fraudulent identities can be better designed by including the characteristics which resulted in the allocation of the control group records to the apparently higher risk profile group.

Results such as this allow designers to confirm the validity of the design and also help direct subsequent work efforts with greater assurance and efficiency. In a compliance context, greater efforts and resources can be directed towards higher risk cases than cases exhibiting lower risk characteristics.

## **Principle 4 - Make greater use of deceased status indicators**

Including deceased status indicators in source data used for matching can help agencies highlight unknown deceased individuals with records in the matching agency's database. A common task for many government agencies is to identify records of deceased individuals within their holdings. When agencies are not notified of death through normal channels, data matching is a way to supplement existing procedures.

Many agencies update their records by matching their own data against fact of death information. Each agency uses the techniques and technologies known and available to them. However, despite a number of approaches available, efforts are inconsistent. Consequently, cases detected by one agency may be missed by the next, though it is in the interests of all agencies to identify such cases.

This raises the risk of records being targeted for 'hijacking' and potentially used for fraudulent purposes such as to continue government payments posthumously. As another example, identities can be 'resurrected' some time after death has occurred, as in cases of 'tombstone' identity frauds. Hence, the inclusion of a deceased status indicator is a useful 'backstop' in ensuring deceased individuals are identified by agencies.

A deceased indicator field may take no more than a single byte of space, meaning that the cost of inclusion in a data matching application is likely to be minimal. Normal processing can take place but match results will now have the potential of highlighting cases of unknown deceased status. If the matching agency is able to identify such examples, the agency can apply lessons from this to improving their fact of death matching efforts.

## **CATEGORY 3 - SOLUTION DESIGN (INCLUDING ALGORITHM)**

### **Principle 5 - Use name, date of birth, address in the algorithm design**

In designing identity data matching algorithms and applications, designers should consider the use of name(s), date of birth and address, as using multiple aspects of record detail in compared data enables greater flexibility in determining what constitutes a match. Consideration may also need to be given to the use of the sex field, although many agencies consider the susceptibility to mis-coding of this value may negate its overall usefulness.

As an example, records showing consistent date of birth and address detail, but some slight name variation, could be considered non-matches due to that name inconsistency. However, the combination of an 'acceptable' degree of name inconsistency with the other consistent field values may indicate that the records do in fact pertain to the same individual and that the match should be retained as such. Such cases may be the result of, for example, poor enrolment practices, where transcription errors and the like affect the eventual data quality.

Another important aspect to consider at the solution design stage is whether a name value should be used for matching in its complete or 'raw' form, or whether the name should be broken into components or parts. A single name may take many different forms due to the inconsistent nature of collection and distribution into smaller individual name fields or tokens.

Generally, names are recorded in agency databases in an Anglo-Saxon format. This format typically uses a 3 part naming convention: first name, middle name and family name. Not all ethnic naming conventions follow or suit this format, and attempts to provide for all names to fit such constructs can often lead to data error. In some ways, this can be seen as an attempt to squeeze a square peg into a round hole. The result will be an unpredictable interpretation of the supplied name and consequently an unpredictable final form in which the details of a name may exist in a broken format.

Further information in relation to ethnic naming conventions can be found in "A Guide to Ethnic Naming Practices", which is a document produced by Centrelink. See also **Principle 6 - Include historical name and address details for each record** and **Principle 15 - Apply the lessons of data matching to enrolment processes**.

As a general principle, the most prudent advice that can be offered here is, include all name data available at the time so that there is the best possible chance of matching a record. Of course, it is recognised that, this may not be practical in all cases.

Data matching is best served with the use of the full name, and this is the preferred way to make use of customer data when using commercial matching products. Some agencies, however, are better placed to make use of 'unbroken' or 'raw' name than others (i.e. some will have access to techniques and technologies designed for matching data in an 'open form').

For agencies not working with commercial data matching products, the complete or raw name may not be a viable option. In these situations, some current concepts proving useful in the 'in-house' construction of data matching solutions include:

- the first n characters of a name, where n ranges from 5 to 15, are matched rather than the whole name
- if address elements are to be used in the matching process, some choice regarding the specific elements may be required (e.g. use of postcode or both street number and street name). This choice should be made following analysis of source databases to determine which option is preferable.

For example, common street names (such as Main and Park) can occur in some agency databases hundreds of times within the same postcode. This must be balanced with the

preponderance of common family names (such as Smith, Jones or Nguyen) within a particular postcode. Preparatory analysis of database contents will assist in this choice.

- middle name or initial may be used to filter potential false positive matches - for example, if both exist but are unequal - delete match
- select age range tolerances that minimise the chances of parent / child match pairs being selected (e.g. no more than 15 years difference to be accepted, even if all other match key fields match)
- first and middle names can sometimes be incorrectly (or at least inconsistently) transposed. Therefore matching of first against middle and middle against first can sometimes produce a match where other match options have failed
- name differences may be the result of inconsistent use of accents, in which case the use of a standard such as the Unicode Standard may assist
- data matchers should be aware that in relation to date of birth, and due to reasons such as ethnic conventions, some values are simply not known and standard 'estimates' recorded in their place (e.g. by far the most frequently recorded date of birth in many databases is 01 January)
- combinations of current and historical names (e.g. current legal, previous legal, maiden, known alias etc.) and addresses should be included as a separate record within each of the tables or files being matched to maximise chances of finding a valid matching record. See also **Principle 6 - Include historical name and address details for each record.**
- if disk space is a consideration when using this approach, then multiple records need not exist permanently but could be created prior to any data matching efforts. The effect of using multiple records for each individual will be simpler processing and design of the data matching application, although at least for the duration of the matching process more disk space will be required to contain the additional records
- match results may be classified and ranked according to the type of match obtained (i.e. scored). As a general guide, the score or inherent value allocated to matches obtained using historical or 'derived' data should be lower than those obtained using clean, complete and current information.

### **Principle 6 - Include historical name and address details for each record**

The use of historical name and address data assists to find matching records where there is no alignment with an individual's current information. Name and address detail is the basis of an agency's identification of, and ongoing contact with, the individuals dealing with that agency. However, identification and contact details can 'decay' over time as individuals change their name and address. Some agencies are informed before others; some are never informed. This has obvious effects on data consistency across agencies and impacts on data matching performance.

#### **Example - Use of current and historical detail to derive records for matching**

In the example below, an agency has recorded in its database an individual whose current details are as follows:

- Name: Mary Jones
- Date of Birth: 01/01/1960
- Postcode: 2168

Mary's previous surname was Davis and her previous postcode was 2906. The agency's Unique Reference Number (URN) for Mary is 123. To maximise the chances of matching a record for Mary in other databases, the file to be created for the matching process could contain a record for each unique combination of name and postcode. The records would all share Mary's URN but with the inclusion of two additional fields, Historical Name Value and Historical Address Value, to ensure that each record can be uniquely identified. The most

recent values for name and postcode can be allocated a value of '1', while the next oldest can be allocated '2' and so on. With this approach, Mary's records would look something like the following:

| URN | Surname | Forename | DOB        | Postcode | Historical Name Value | Historical Address Value |
|-----|---------|----------|------------|----------|-----------------------|--------------------------|
| 123 | Jones   | Mary     | 01/01/1960 | 2168     | 1                     | 1                        |
| 123 | Jones   | Mary     | 01/01/1960 | 2906     | 1                     | 2                        |
| 123 | Davis   | Mary     | 01/01/1960 | 2168     | 2                     | 1                        |
| 123 | Davis   | Mary     | 01/01/1960 | 2906     | 2                     | 2                        |

The retention and use of previous names and addresses (i.e. historical information) can be a crucial element in data matching. As discussed in **Principle 5 - Use name, date of birth, address in the algorithm design**, combinations of current and historical names and addresses should be included as a separate record within the tables or files being matched. This may take some preparation but will increase the chances of finding valid matching records, should they exist.

It should also be remembered that names can have the added dimension of having a name type or category such as legal, maiden, alias, preferred, tribal names. This can impact on design choices or the interpretation of results.

**Principle 7 - Ensure the use of a flexible matching algorithm**

Name matching should optimally employ orthographic<sup>3</sup>, linguistic or phonetic (or any combination thereof) fuzzy logic pattern matching. Flexibility in the values of all other matching elements such as date of birth and address should also be considered.

Whether a matching solution has been developed in-house or is a commercial product, developers will need to determine what constitutes a match. Developers will have decided upon those fields they wish compared and the form in which they will be compared (e.g. standardised, raw, complete string, broken into component values). Some of these decisions will be influenced by data format and the type of infrastructure that is available to an organisation.

Agencies will also need to decide on the degree of field value correlation they are willing to accept in the matching process as constituting a match. If two records have largely consistent, but not exact, field values in those areas being compared (e.g. name, date of birth, address), the developer, in conjunction with business analysts, will have to establish the boundary between acceptable difference and unacceptable difference, a decision that will also need to take into account the risks posed by the various options.

**Example - Identity Detail Variation and Choice of Algorithm**

An agency may have a record with the following details:

---

<sup>3</sup> Refer Glossary

- Name: Cheryl Donna Blewitt
- Previous Name: Cheryl Donna Andolini
- Date of Birth: 12 / 11/ 1978
- Current Address: 4 / 20 Benjamin Street, Barton, ACT, 2600
- Previous Address: 165 Lucas Road, Liverpool, NSW, 2170

A matching agency may choose to accept any of the following records as a match against this record:

#### Case 1

- Name: Cheryl Donna Blewitt
- Date of Birth: **11 / 12/ 1978** (*day / month transposition*)
- Address: 4 / 20 Benjamin Street, Barton, ACT, 2600

#### Case 2

- Name: **Sheryl** Donna Blewitt (*first name variation*)
- Date of Birth: 12 / 11/ 1978
- Address: 4 / 20 Benjamin Street, Barton, ACT, 2600

#### Case 3

- Name: Cheryl Donna Blewitt
- Date of Birth: 12 / 11/ 1978
- Address: **20 Benjamin Street, Barton, ACT, 2601** (*missing unit number, postcode variation*)

#### Case 4

- Name: **Cheryl Blewitt** (*no middle name*)
- Date of Birth: 12 / 11/ **1979** (*year of birth difference of 1 year*)
- Address: 4 / 20 Benjamin Street, Barton, ACT, 2600

#### Case 5

- Name: **Donna Cheryl** Blewitt (*name transposition*)
- Date of Birth: 12 / 11/ 1978
- Address: 4 / 20 Benjamin Street, Barton, ACT, 2600

#### Case 6

- Name: Cheryl **Andolini** (*no middle name, previous family name*)
- Date of Birth: 12 / 11/ 1978
- Address: **165 Lucas Road, Liverpool, NSW, 2170** (*previous address*)

The matching solution must then cater for each of these possibilities, and any other that fits within the accepted 'tolerance' determined by the matching agency.

However, they may choose not to accept the following as a match:

#### Case 7

- Name: **Donna** Blewitt (*middle name as first name, no middle name*)
- Date of Birth: 12 / **04/ 1984** (*different month and year of date of birth*)
- Current Address: **120 Jensen Street, Parkes**, ACT, 2600 (*same postcode but different street address and suburb*)

In the instance of Case 7 above, an agency may decide that the degree of flexibility required to make the match is too great and there is an unacceptable risk of obtaining a false positive result. This highlights the need for the matching agency, through investigation and research, to determine the acceptable levels of tolerance and the means of achieving them in their data matching solution.

Given the complexities of naming conventions, the most accurate identity data matching solutions are probably commercial software products which have been specially designed for this purpose. Such solutions employ orthographic, linguistic or phonetic principles to verify the ethnic origin, as well as the variations and meanings of names. Many also are able to detect nicknames and titles within the name values recorded. More information about this is contained in **Principle 14 - Identify and evaluate commercial products**.

For agencies not in a position to employ such solutions, the in-house solution should cover multiple match key combinations, from exact or relatively 'tight' at one extreme, to less exact or 'looser' combinations at the other. The result of the match may be the allocation of some kind of ranking or score, denoting the nature of the match obtained. Generally, if a record matches with more than one match key combination, that match key considered the 'highest', or closest to exact matching should be the result allocated to the matched record.

An example of the use of a multiple match key algorithm which produces ranked or scored results appear in the following example:

### Example - Multiple Match Key Algorithm

|                          |          |                                   |           |             | Current Client Data  | Historical Client Data  |
|--------------------------|----------|-----------------------------------|-----------|-------------|--|---|
| Components of Match Keys |          |                                   |           |             | Score if matched with <i>current</i> matching agency name data | Score if matched with <i>historical</i> matching agency name data |
| surname                  | forename | ddmmyyyy                          | Street No | Street name | A  | B   |
| surname                  | forename | ((dd or mm) & yyyy)               | Street No | Street name | C  | D   |
| surname                  | forename | ddmm & (yyyy +/-1 to yyyy +/-5)   | Street No | Street name | E  | F   |
| surname                  | forename | ddmm & (yyyy +/-6 to yyyy +/-10)  | Street No | Street name | G  | H   |
| surname                  | forename | ddmm & (yyyy +/-11 to yyyy +/-15) | Street No | Street name | I  | J   |
| surname                  | -        | ddmmyyyy                          | Street No | Street name | K  | L   |
| -                        | forename | ddmmyyyy                          | Street No | Street name | M  | N   |
| surname                  | -        | ddmm & (yyyy +/-1 to yyyy +/-10)  | Street No | Street name | P  | Q   |
| surname                  | forename | ddmmyyyy                          | -         | -           | R  | S   |
| surname                  | forename | ddmm & (yyyy +/-1 to yyyy +/-5)   | -         | -           | T  | U   |
| -                        | -        | -                                 | -         | -           | 0  |   |

The above algorithm uses surname, forename, date of birth, street number and street name as the match key components. It differentiates between matches achieved with current data and those achieved with historical data, allocating higher 'scores' to the former. Also, those records exhibiting the closest degree of correlation in data field values receive a higher 'score' than those with more variation. Those records failing to meet any of the 20 potential match key combinations (i.e. failed to match) are allocated a score of zero.

Another design feature worthy of consideration is the use of middle name to first name comparisons. Experienced data matching agencies report that this occurs with some regularity. Also, middle name or initial can be useful in discounting potential matches. For example, an agency may decide against classing Cheryl Donna Blewitt and Sheryl A. Blewitt as a match because of the inconsistency of the middle name / initial.

The agency may, however, class Cheryl Donna Blewitt and Sheryl Blewitt as a (lower quality) match due to the lack of contradictory evidence in the middle name/initial. They may just as likely discount it though for the lack of confirming evidence. In the end, it is a business decision. The decision on exactly where to draw the line on matching tolerance would be related to the degree of risk posed by either accepting or rejecting matches of that kind.

### **Principle 8 - Increase confidence in identity information by confirming data with a number of sources**

Australia's identity management system does not rely on a single identifier. A variety of sources of identity data, whether internally or externally sourced, is often required to confirm or verify key identity details. In an identity context, a greater number of data sources will provide better and more meaningful results.

As an example, where agencies are attempting to identify fictitious identities within their databases, 'non-matching' of identities across many databases may be indicative of greater 'identity risk' than 'non-matching' against a single database. Data matching applications employing this principle are used in the detection of identity fraud.

Considering the use of a variety of sources of data would also mean the involvement of a greater variation in the number and types of approaches to data matching, meaning exposure to a greater number of techniques and technologies.

See also **Principle 9 - Use profile groups to stratify data matching results** and

**Principle 17 - Consider new or future sources of identity data.**

## ***CATEGORY 4 - MATCH RESULTS AND INTERPRETATION***

### **Principle 9 - Use profile groups to stratify data matching results**

Often, the reason for undertaking data matching is to be able to make certain business decisions based on the outcomes of the data matching process, where different results may require different types of action. Allocating records to a range of profile groups based on data matching results allows records to be categorised according to specific business rules or needs. It also allows for the subsequent processing of matching results to be tailored in order to maximise efficiencies.

As examples:

- in a compliance context the allocation of records of the greatest potential risk to a particular group can assist agencies to better target resources in terms of investigative and corrective procedures. More effort can be directed towards those examples which meet the highest potential risk and less towards those exhibiting low or no risk

- in an enrolment or ‘facilitation’ context, agencies can use profile groups to streamline ‘customer’ processing, again with more appropriate use of agency resources. Data matching can determine additional information regarding the customer, on their behalf, and use that information to direct a request for service to the most appropriate area. Allocation to a particular group, based on the set of match results, brings consistent treatment and processing options for individuals when dealing with agencies and leads to improved levels of service
- in a data cleansing context, data matching enables the frequency of different categories of data anomalies (e.g. missing values, invalid formats, such as zero-filled date fields etc.) to be determined by allocation to set risk groups. The first step to cleansing and improvement is identification and data matching can assist in this process.

Most commercial data matching solutions can be used to direct records to specific profile groups if they are correctly tuned. Manual in-house solutions require the application developer to implement business rules to interpret the results and allocate each record to the most appropriate group.

### **Example - Potential Profile Groups**

Agencies deciding to use profile groups in the interpretation of data matching results will need to design the groups in a way that maximises benefit. The design will be related to the aims of the data matching exercise and subsequently required business activities. The following profile group examples have been designed to suit a data matching application aimed at assessing the potential ‘identity risk’ posed by identities in a particular data base and at the same time, to help identify data integrity issues that may affect data matching results. The groupings in this example lend themselves best to comparison of the source data with data from multiple other agencies.

#### Profile Group 1 - Identity is Highly Confirmed

Records in the source file which found matching records in a wide range of other databases would be allocated to this profile group.

#### Profile Group 2 - Identity is Confirmed to a Substantial Level

Records in the source file which found matching records in a substantial range of other databases (but less wide-ranging than those in Profile Group 1) would be allocated to this profile group.

#### Profile Group 3 - Identity is Partially Substantiated

Records in the source file which found matching records in a small number of other databases would be allocated to this profile group.

#### Profile Group 4 - Identity is not Substantiated

Records in the source file which failed to find matching records in any other databases would be allocated to this profile group.

#### Profile Group 5 - Identity is substantiated but also uses another name

Records with a matching record in another agency database containing a new or updated name would be allocated to this group

#### Profile Group 6 - Identity recorded as deceased with at least one agency

Records with a matching record in one or more other databases indicating that the individual is deceased would be allocated to this group.

Each group is informative in its own way. As examples:

- in an enrolment or facilitation context, identities allocated to Profile Group 1 may be directed to more streamlined processes, as much information about them is recorded online
- in a compliance context, identities allocated to Profile Group 4 may be directed to the initial steps in a process of confirmation of the supplied identity details. If no subsequent identity confirmation is possible this group may represent a higher risk of identity fraud
- records allocated to Profile Group 6 could represent a possible data integrity issue. Where practical, deceased individuals should be identified and records updated accordingly.

In each case, subsequent business activity is better targeted and more efficiently delivered.

### **Principle 10 - Combine human involvement in the analysis of data matching results when flexible matching has been employed**

One of the efficiencies deliverable with the use of data matching is the ability to automate particular actions or activities depending on the results obtained. Such automated ‘cause and effect’, or ‘lights-out’, systems are based on the perceived accuracy (or believability) of the results obtained and the low risk involved in automating subsequent business activity. As examples, data matching can be used to facilitate online enrolment or used to pre-populate forms.

However, at least initially, data matching based on flexible or fuzzy techniques should not initiate automatic actions due to the greater chance of false positive or false negative results. Human evaluation of results not only confirms the validity of any matching that has taken place but the analysis and evaluation involved provides recursive advice for improved data matching. This approach is applied in the design of tools aimed at this type of matching.

The amount of human involvement required in the analytical process may depend on a number of factors, including the importance of the business aims of the application, the number of analysts available and the effort involved in result interpretation.

Once completed, the initial ‘product’ of the data matching process using flexible or fuzzy logic can be a candidate list of matched records which is directed to a human analyst. Each record pairing in the list will have a probability relating to the likelihood of the two matched records relating to the same individual.

This probability may be in the form of a score or a ranking, and may have been determined by the correlation of the values in each record of the pair using, for example, direct character comparisons or perhaps rules of an orthographic, linguistic or phonetic nature. The score may also allow allocation of the results to one of high, medium or low threshold risk groups. This process should be conducted within the right conceptual and perceptual parameters, and appropriate risk management approach.

Determining the correct conceptual and perceptual parameters is a role best performed by specialist analysts who have a good understanding of both the data and the business underpinning the application. Given the intelligence invested in modern matching toolsets, it would be unwise to have the matching engines return results to an untrained user to assess in isolation. In any case, results are best analysed in terms of functionality and risk rather than cost. Agencies requiring post-match results analysis should also consider developing and implementing formal procedures for this process.

As an example, agencies requiring strong negative search strategies (like those applied to watch lists) may need to introduce quite intensive post-matching processes including manual analysis of results, as the case of not detecting a match may have severe consequences. As another example, de-duping (i.e. remove duplicate records) mailing lists would not require the same efforts, less the results would be completely outweighed by the cost of the data matching activity.

## **CATEGORY 5 – IMPROVED DATA QUALITY**

### **Principle 11 - Identify and quantify data integrity issues which affect the ability to match records or data**

Data integrity and data content can impact on data matching results. If the results of data matching are to be fully understood and appreciated, it would benefit those performing the exercise to better ‘understand’ the data with which they are working. Analysis of data before matching begins, or in light of the results obtained during any testing period, can be informative and recursively aid refinement of data matching applications and the subsequent interpretation of results.

The context of observed characteristics is also important. In relation to identity data, analysis would benefit from an understanding that some conventions surrounding name and dates of birth, for example, can result in certain data characteristics. This understanding helps prevent the labeling of certain data topology<sup>4</sup> trends as data integrity issues. Date of birth, for instance, is often a problematic data field. In some cultures for instance, there are only two dates of birth, 1 January and 1 July, depending on the half of the year in which an individual is born. Similarly, in some cultures it is possible for an individual to be known as a single name. The effect of recording this non-Anglo-Saxon name in a database conforming to Anglo-Saxon constructs (first, middle and family name) is that two fields will be left blank and will appear to be a data integrity issue.

See also **Principle 6 - Include historical name and address details for each record** and **Principle 15 - Apply the lessons learned from data matching to enrolment processes.**

Fields may also contain invalid or non-sensical values. For example, dates of birth may contain zero-filled values, which can have a direct affect on the ratio of non-matches obtained. Efforts should be made to identify and quantify the prevalence of such characteristics.

Knowing the preponderance of various data anomalies and characteristics would assist in better understanding the data matching results obtained and more correctly interpreting their significance. This is illustrated in the following two scenarios:

- failure to match is due to the fact that there exists no record for that identity in the other databases
- a record exists for the same identity in the other databases but there is a failure to match because the date of birth for one record is zero-filled.

If, for example, an aim of a data matching exercise was to determine which identities in a particular database exhibit higher ‘identity risk’ by not appearing in other databases, the inclusion of records from both of the above scenarios in the same category of output skews any real understanding of the problem.

A preliminary analysis of data quality can help place subsequent results into context. Invalid, missing, duplicate or otherwise ‘incorrect’ values can be identified prior to matching. It can also be the first step to data cleansing and corrective activity; identification of missing or incorrectly filled fields can be productive. Match results can be better interpreted with such knowledge or alternative actions implemented in light of it. (The adoption of data standards<sup>5</sup> may also assist in the improvement of data quality prior to its use in data matching.)

On a related issue, agencies transferring data from legacy systems to new ones should also obtain a better understanding of the quality of the data they are about to move. Failure to do so

---

<sup>4</sup> Refer Glossary

<sup>5</sup> Refer Principle 1 – Improve the process and form of data transfer

may have unforeseen consequences for downstream data matching activities, as well as other applications once the data ‘lands’ in its new database.

## **Principle 12 - Employ address validation techniques**

As with naming, address details can be misinterpreted, corrupted, incorrectly transcribed or intentionally misrepresented. Add to this the changes that occur from time to time to postcodes or suburb boundaries, and the likelihood of anomalous address data is very real. The effects can be just as problematic for data matching as poor quality name detail.

Address validation will improve data quality, consistency and accuracy and therefore the quality of data matching results. Address-cleansing or validation software solutions can help to ensure that:

- incorrect postcodes are amended
- delivery point identifiers are added
- missing address elements are added
- spelling errors are corrected
- verified addresses are consistently formatted and standardised

Validations, for example, can allow for comparisons with a known quality reference set such as Australia Post’s Postal Address File (PAF) or Australia Post’s Delivery Point Identifier (DPID). Address data can also be validated and made more useful with the use of geocodes. Geocodes have potential applications in general business, management information and compliance, including identity fraud detection.

The use of address validation measures at the data collection point (i.e. during enrolment) can help to ensure that intra-agency searching, matching and general application of address detail is more reliable and predictable. See also **Principle 15 - Apply the lessons learned from data matching to enrolment processes.**

When an agency validates and standardises all recorded address detail, data becomes consistent, duplicate records can be better detected and corrected, and databases can become more accurate and reliable. With cross agency adoption of address validation, address detail becomes consistent across agencies and results in better data matching outcomes for whole-of-government. A further advantage is that validation allows for enrolment or registration to be simplified by reducing the number of key strokes required, which also reduces the potential for error.

## **CATEGORY 6 - KNOWLEDGE SHARING AND LONGER TERM DATA MATCHING DEVELOPMENT ISSUES**

### **Principle 13 - Share expertise, particularly in specific data matching subject matter areas**

It makes sense, particularly in a whole-of-government context, that agencies share relevant data matching expertise. The development of data matching applications and techniques is a niche area, with increases in demand for skills in this area. There are several ways that this could be resolved, including to establish communities of interest such as with the DMWG.

It should also be acknowledged that some agencies are in a better position to purchase quality data matching and data integrity-oriented 'tool kits'. However, there may be short term improvements possible with the sharing of quality 'in-house'-developed applications. Examples of applications of proven quality include those produced by Centrelink for name and address standardising and the AEC's phonetic name comparison routines. Further information in relation to this using the AEC as a model is contained in **Attachment A - Case Study - Flexible Matching and Standardising**.

Better practice can also be shared through the use of specific tools such as the template at **Attachment C**, which was designed for use by the DMWG to gain a broad understanding of various approaches to data matching. Readers may wish to use the template to assess their own data matching needs.

Data matching is quite specialised. Consideration should be given to recruitment, training and the retention of staff with the appropriate skills and experience. Consideration should also be given to developing skills and succession planning.

### **Principle 14 - Identify and evaluate commercial products**

Commercially produced software products are available which can:

- assist or improve data matching efforts by helping to identify and correct data quality or data integrity issues, or
- perform data matching

Agencies may choose to invest in either or both of these types of complementary products.

#### **Data Integrity Products**

Commercial products designed to assist in the improvement of data quality are able to quantify the frequency of certain data characteristics such as:

- null values
- zero counts / percentages
- low / high / average / median values and counts
- minimum / maximum / average / median string length
- blank count / percent

Commercial products can also allow profiling of data through pattern analysis to identify previously unknown data quality issues, which can inform application development by allowing corrective actions or changes in design. Further information about this is contained in **Principle 11 - Identify and quantify data integrity issues which affect the ability to match records or data**.

## Data Matching Products

Commercial data matching software is often the outcome of extensive research and development. Complex mathematical modelling and research into cultural and linguistic traits has resulted in the development of applications that are able to mimic informed human perception. Even minor inconsistencies, such as a missing character or value which human observers can often discern, may not always be highlighted with in-house applications.

Cultural and linguistic differences can also make name matching extremely challenging for in-house developers. Some commercial products contain references to literally millions of name variations covering all major ethnic groups and geographic regions. Names from some non-Anglo-Saxon cultures contain a variety of complicated elements not part of ‘traditional’ Western names including references to gender, profession, religion, military status, culture, age and country.

## Existing In-House Applications v. Commercial Products

Data matching efforts involving the use of applications developed in-house are currently producing savings for Government valued at many millions of dollars annually. These applications also help to ensure correctness of government records. They are fast and efficient.

However, although speed and efficiency are important considerations for any data matching application, so is the quality of results produced. In deciding whether or not to invest in commercial data integrity or matching products agencies should ask themselves: “*What are we not getting with our current data matching applications?*”

Some agencies using commercial products have been able to realise noticeable improvements in data matching performance compared to previously existing applications. Agencies that are able to use some of the best commercial products have even been able to observe significant capability improvement even when upgrading between software versions.

Although not all agencies are in a position to identify and test commercial software, agencies should consider ways that commercial products can be evaluated and the results of these evaluations made available to whole-of-government. Promotion of agency findings surrounding particular products may enable other agencies to identify potential solutions which may meet their own needs, or to further investigate the possibilities of testing software they may not have otherwise considered using.

There may also be scope for agencies to see in use, or perhaps even to use for testing purposes, the commercial products in use with other agencies. This may allow some informed comparison with their own existing data matching processes.

## Principle 15 - Apply the lessons learned from data matching to enrolment processes

The quality of identity data can be adversely affected by events occurring during enrolment processes or during subsequent updating; transcription errors are examples. Many of the principles outlined in this document can be used to improve the quality of identity data after this has occurred. They can, however, also be applied during the enrolment or updating processes, thereby negating the need for subsequent correction. Early application of these principles *prevents* the occurrence of problems, which is preferable to the need to correct problems which have already occurred. Preliminary application of these principles also ensures improved subsequent data matching performance by improving the quality of data being matched.

One of the proven ways of raising matching performance is in the area of data validation. A good example is address validation, as detailed in **Principle 12 - Employ address validation techniques**. Address validation can help to ensure that the quality and accuracy of personal

data is improved at point of capture, increasing efficiency, particularly in matters involving data transfers or cross-agency data matching.

Another benefit of applying data matching principles during enrolment is that they can be used to facilitate better, more streamlined service for those being enrolled. For example, a minimal number of supplied personal details can be used to initiate data matching of available data sources in order to allow, pre-population of forms. This minimises the imposition upon the individual being enrolled and ensures greater consistency of data by using other existing data sources. It reduces the work for front line staff and makes enrolment a faster and more accurate process. These types of systems could also be used to improve business by automatically directing information to particular areas most suited to an individual's requests.

## **Principle 16 - Undertake research and development involving data matching**

Data matching can highlight cases of a known type or which meet specific and well-defined criteria when the source data is split over two or more sources. It can also identify and quantify previously unknown areas of interest or concern.

Data matching applications can be used to produce particular outputs such as:

- determine if an individual is entitled to a particular service or benefit
- detect fraud
- identify individuals who appear on particular lists such as watch lists

Data matching applications with much broader aims can also be designed. However, in the case of data matching designed to help improve identity data quality, applications would need to be wide-ranging in design. The design would need to test a number of aspects, including matching techniques and approaches to data handling, and consideration would need to be given to the analysis of the data matching results and the matching process. The design would also need to allow for anomalous results to be identified and quantified to determine their significance and possible methods of redress.

Other areas of research and development to consider include:

- geo-spatial analysis and profiling tools
- improved or automated enrolment procedures using data matching facilitation
- data validation ratings
- comparisons of data definitions to enable improved cross-agency data matching.

## **Principle 17 - Consider the use of new types of data for data matching**

In terms of future data matching developments, agencies should be open to the identification, adoption and use of new and emerging data sources in a data matching context. They should also be attuned to possible advancements in data matching, both 'here and now' and 'over the horizon'. Data matching results may be improved with the use of new data sources and types that meet appropriate standards. Awareness of, and adaptation to, new opportunities will help 'future proof' agency data matching efforts and keep them relevant and effective.

As an example, some agencies have already implemented techniques and applications in relation to the use of biometrics. As with **Principle 15 - Apply the lessons of data matching to enrolment processes** and issues in relation to name, consideration should be given to the quality of biometric data. Consideration should also be given to the use of interoperable technologies early on as part of any implementation phase, to avoid unnecessary limitations in the use of biometric data.

## 4. Glossary

|                                    |   |
|------------------------------------|---|
| <b>Address elements</b>            | The individual component elements / fields of an address string e.g. street number, street name, street type, town / suburb   |
| <b>Algorithm</b>                   | A set of logic rules determined during the design phase of a data matching application. The ‘blueprint’ used to turn logic rules into computer instructions that detail what steps to perform in what order   |
| <b>Application</b>                 | The final combination of software and hardware which performs the data matching   |
| <b>Business rules</b>              | Rules of a business nature to help guide the design of a data matching application  |
| <b>Control group</b>               | In a data matching context , a set of records of a known type (e.g. previously identified fraudulent identities, deceased individuals) which are used to better interpret data matching results               |
| <b>Cross agency data matching</b>  | The matching of data from one agency with those of one or more other agencies   |
| <b>Database</b>                    | A structured collection of records or data that is stored in a computer system  |
| <b>Data cleansing</b>              | The proactive identification and correction of data quality issues which affect an agency’s ability to effectively use its data   |
| <b>Data consistency</b>            | In a data matching context, the degree of correlation in design and content of different data stores  |
| <b>Data error</b>                  | The allocation of incorrect values to specific database fields e.g. a family name stored in the ‘Given Name’ field  |
| <b>Data field</b>                  | A physical unit of data such as name, date of birth or address  |
| <b>Data file design</b>            | The criteria used by agencies to describe the ways they will store individual data fields within their databases  |
| <b>Data integrity</b>              | The quality of correctness, completeness and compliance with the intention of the creators of the data i.e. ‘fit for purpose’   |
| <b>Data matching</b>               | The bringing together data from different sources and comparing it  |
| <b>Data topology</b>               | The order or relationship of specific items of data to other items of data  |
| <b>Deceased status indicator</b>   | A field value to indicate whether the agency has been notified of the death of the individual to whom the record has been allocated   |
| <b>Enrolment</b>                   | The process of an individual to enrolling with an agency. Involves the initial collection of identity details   |
| <b>False negative</b>              | In a data matching context, a result indicating that no matching record exists for a particular individual’s record, when in fact a matching record does exist but the matching application failed to find it |
| <b>False positive</b>              | In a data matching context, when the record for one individual has been incorrectly matched with that of another  |
| <b>Field values</b>                | The information stored in data fields   |
| <b>Flexible matching</b>           | Matching which allows some tolerable degree of difference in the values being compared and still constitute a match   |
| <b>Full name</b>                   | The complete set of name detail of an individual  |
| <b>Historical (name / address)</b> | Detail which an individuals do not currently use to declare themselves to an agency but have done so in the past (e.g. previous addresses and names)  |
| <b>In-house</b>                    | With reference to data matching / data quality software, a product developed internally by an agency to meet its business needs   |
| <b>Linguistic</b>                  | In a data matching context, rules relating to language  |

|                             |   |
|-----------------------------|---|
| <b>Match key</b>            | The combination of data fields which are the basis of comparison in a data matching application   |
| <b>Match results</b>        | The set of matched records produced by a data matching application  |
| <b>Matched records</b>      | Two or more records brought together as a match   |
| <b>Name inconsistencies</b> | When the same individual is recorded with varied identity detail by different agencies  |
| <b>Name token</b>           | A component of the full or raw name such as family name, first given name or title  |
| <b>Name type</b>            | Describes the nature of a name used currently or previously by an individual such as legal, maiden name or an alias   |
| <b>Non-matched records</b>  | Records for which a data matching application failed to find a matching record in one or more other data files. NB. This is not to say that a record for the individual does not exist elsewhere, only that the application failed to find one  |
| <b>Orthographic</b>         | A principle used in data matching where correct or accepted spelling and characters are used to determine the results   |
| <b>Profile groups</b>       | In the interpretation of identity data matching results, the allocation of matched records to particular groups depending on the ways in which matching was obtained. Used to better allocate resources to subsequent processing of results   |
| <b>Phonetic</b>             | Agreeing with or corresponding to pronunciation   |
| <b>Raw name</b>             | The name in its full form (i.e. not broken into tokens or elements)   |
| <b>Standardise</b>          | A process performed prior to data matching intended to overcome some data inconsistencies which may adversely affect data matching results such as removal of hyphens, apostrophes  |
| <b>Score</b>                | A value (perhaps numeric or alphanumeric) allocated to matched records to describe the degree of value correlation that exists between those records e.g. a score of 100 may indicate that all elements of the match key were exactly the same, while a score of 90 may indicate some slight variance |
| <b>Stratify</b>             | The splitting of the set of match results into profile groups   |
| <b>Unicode Standard</b>     | A character code of 1-4 bytes that defines every character in most of the speaking languages in the world   |

## Case Study – Flexible Matching and Standardising

The following approach, adopted by the Australian Electoral Commission (AEC), is a useful and informative case study:

Part A - Name, Date of Birth and Address Handling

Part B - Name Key Generation Rules

Part C - Phonetic Name Matching Scoring

This approach has proven to be successful and agencies may consider adopting similar solutions.

### Part A - Name, Date of Birth and Address Handling

#### Name Matching

The following approaches to non-exact matching are applied to name fields:

##### *In-house phonetic analysis algorithms*

The basic process is as follows;

- i. All source and target 'records' are allocated a 10 character NAME-KEY based on the persons surname, 1<sup>st</sup> given name and date of birth (sic).  
Eg. Michael Nielsen 1957 becomes NLSN MSL57  
The name key generation rules are set out in Part B.
- ii. When matching a name, all data records with the same NAME-KEY are retrieved and 'scored' out of a maximum 99 according to the rules set out in Part C. These rules also may vary slightly depending on the data source.
- iii. Once this score has been calculated then any candidate records who score under the threshold (adjustable by the user) will be eliminated from further processing.
- iv. All candidate records remaining may then be ;
  - a. in an online scenario, presented to the user in online screens , sorted by score (highest first) and by name within score OR
  - b. in a batch scenario, assessed further depending on the batch runs requirements (may discard all, print all or other).

## ***Name Interchange***

One of the major problems encountered in a name search is sequence errors, which could be caused by any of the following factors:

- Cultural difference in the presentation of name. Most Asian names are written with surname first and followed by given names;
- Use of second and third given names; and
- Transcription errors.

The software can swap any part of a name to locate all possible matches on the database as follows:

- Interchange first two given names,
- Interchange surname with first given name, and
- Interchange surname with second given name.

## ***Name Substitution***

A name substitution facility is required to overcome the problem of related but not 'sounds-like' names; for instance, **Bill** is a nickname for **William** and they do not sound like each other. This coupling process is extremely useful to retrieve all target electors with related names. At present, the AEC has an on-line function to maintain a list of name substitutes, which has over 1500 substitutes for surnames and given names. This list can be added or deleted at any time without impacting the name key structure of the database.

This facility has the ability to substitute surnames or given names and further enhances the accuracy rate of the name matching software, and can be turned on or off by the user. First given name substitution is used as a default option.

## **Date of Birth Matching**

### **a. Fuzzy Date of Birth comparison**

An additional facility within the in-house name matching software is 'fuzzy' matching on the whole date of birth.

If the source date of birth is entered as a part of the search criteria and the 'Fuzzy Date of Birth' option is selected, the software performs the following analysis of the Date of Birth and 'scores' as shown. This score is added to the score totalled after the regular Name matching.

|   |         |
|---|---------|
| DoB Exact match   | score 9 |
| Day and Month transposed<br>i.e. 04/05/57 and 05/04/57                    | score 8 |
| Day exact match and month digits transposed<br>i.e. 04/01/57 and 04/10/57 | score 7 |
| Month exact match and day digits transposed<br>i.e. 12/05/57 and 21/05/57 | score 7 |

|  |         |
|--|---------|
| Day exact match and month single digit match<br>i.e. 04/05/57 and 04/09/57                 | score 6 |
| Month exact match and day single digit match<br>i.e. 23/05/57 and 27/05/57                 | score 6 |
| Century and year match   | score 5 |
| Century, month and day match but year single digit match<br>i.e. 04/05/1957 and 04/05/1967 | score 5 |
| Century, month and day match but year digits transposed<br>i.e. 04/05/1957 and 04/05/1975  | score 4 |

### Address matching

The AEC maintains an Australian 'Address Register'. No enrolment may be processed unless the address to which the elector is assigned is known to the Address Register. An address may not be added to the Register unless its existence and land-use type has been proven to the satisfaction of electoral officers.

Addresses may be used to help identify electors when there are very similar looking or sounding person's names.

AEC only uses direct text comparison for address matching processing. No phonetic or 'fuzzy' name matching is used.

- Generally matching will occur on *locality, street name, street number and flat/unit number/habitation* name if they exist.
- *Street type is only used if there is more than one street with the same name (say Street and Place) in the locality.*

## Part B - Name Key Generation Rules

Listed below are the steps taken to turn a Surname and 1<sup>st</sup> Given name and Date of Birth into a NAME-KEY of 10 characters.

1. Remove all non alphabetic characters from the name.
2. Replace all 'EE' with 'Y'.
3. Replace consecutive repeated characters with one character.
4. Check the first 3 characters and make the following substitutions  
'CHL','CHM','MAC','SCH' is replaced by  
'CLL','CMM','MCC','SSS' respectively
5. If no 3 character match is found then check the first 2 characters and make the following substitutions  
'CH','CZ','DJ','KN','PH','TS','TC','TZ','WR','YV' is replaced by  
'SS','SS','GG','NN','FF','SS','SS','SS','RR','AV' respectively.
6. If no 2 character match is found then check the first 1 character and make the following substitutions  
'E','I','J','K','O','Q','U','Z' is replaced by  
'A','A','G','C','A','C','A','S' respectively.
7. Check the last 2 characters and make the following substitutions

'CE','DT','EE','EH','ET','IE','ND','NT','RD','RT','SE','SS','YE','YI','ZE','ZZ' is replaced by  
 ' ','DD','YY','YY','DD','YY','DD','DD','DD',' ',' ','YY','YY',' ',' ' respectively

8. If no 2 character match is found then check the first 1 character and make the following substitutions  
 'I','S','Z' is replaced by  
 'Y',' ',' ' respectively.
9. Examine the 3 middle characters of the name starting from position 2 and replace them if they match one of the specified substitutions;  
 'SCH','VSK' is replaced by 'SSS','SSC' respectively
10. If no 3 character match is found, examine the 2 middle characters of the name starting from position 2 and replace them if they match one of the specified substitutions;  
 'CH','DG','EV','KN','PH','PS','SH' is replaced by  
 'SS','GG','AF','NN','FF','SS','SS' respectively
11. If no 2 character match is found, examine the middle character1 of the name starting from position 2 and replace it if it matches one of the specified substitutions;  
 'J','K','M','Q','Z' is replaced by  
 'G','C','N','C','S' respectively.
12. Replace consecutive repeated characters with one character.
13. Drop all vowels.

The above 13 steps are performed against the Surname and 1<sup>st</sup> Given name independently.

The Surname produces a 5 character key component.

The first Given Name produces a 3 character key component, truncated if necessary.

14. Set the last 2 characters of the NAME-KEY to the year of birth. If unknown set to blank.

## Part C - Phonetic Name Matching Scoring

When scoring on **Surname** look for an **exact match**,

if not found then look for a match on the **first 3 characters**

if not found then look for a match **on first vowel and first 3 consonants (excluding 'y')**

|                                  |          |
|----------------------------------|----------|
| Surname matches                  | score 50 |
| Surname first 3 characters match | score 35 |
| First vowel matches              | score 3  |
| First consonant matches          | score 10 |
| Second consonant matches         | score 5  |
| Third consonant matches          | score 4  |

When scoring on **First Given Name** looks for an **exact match**,

if not found then look for a match on the **first 3 characters**

if not found then look for a match **on first vowel and first 3 consonants (excluding 'y')**

|   |          |
|---|----------|
| First given name matches (or blank)       | score 35 |
| First given name first 3 characters match | score 30 |

|                                  |          |
|----------------------------------|----------|
| First given name initial matches | score 25 |
| First vowel matches              | score 3  |
| First consonant matches          | score 10 |
| Second consonant matches         | score 5  |
| Third consonant matches          | score 4  |

When scoring on **second given name** look for an **Exact Match**  
if not found then look for a match on the **first 3 characters**

|  |         |
|--|---------|
| Second given name matches                  | score 5 |
| Second given name is spaces                | score 3 |
| Second given name first 3 characters match | score 3 |

When scoring on Date of Birth

|                        |         |
|------------------------|---------|
| Year of birth matches  | score 9 |
| Year of birth is blank | score 9 |

Other year of birth score is calculated as below:

$$((100 - \text{year-diff}) * \text{year-diff} * 0.5) * 0.09$$

# Data Definitions - AS4590 - Standard for Name and Address Data Transfer

Drawn from the 'AS4590 - Interchange of Client Information', the information below is in three parts as follows:

AS4590 – Person Attributes

AS4590 – Person Name Attributes

AS4590 – Address Attributes

This information is provided to assist, but is only a guide. It is meant to be provided as an example of a standard that could be broadly adopted to help ensure common format of identity data transfer between agencies.

## AS4590 – PERSON ATTRIBUTES

### Person Attributes

| Attribute   | Details   |
|---|---|
| <p><i>public</i> <i>Date</i><br/><b>Birth</b></p>               | <p>&lt;&lt;BCC&gt;&gt;<br/> <i>Range:0 to 1</i>businessTerm=Date of Birth, DOB<br/>           Definition=The date when a person was born as recorded on the birth certificate or other official documents.<br/>           RepresentationLayout=YYYYMMDD<br/> <i>Notes: Verification Rules and Examples</i><br/>           20060217, where<br/>           20 = Two digits Century value (YY)<br/>           06 = Two digits Year value (YY)<br/>           02 = Two digits Month value (MM)<br/>           17 = Two digits Day value (DD)<br/> <i>Constraints:</i><br/>           OCL self.Content.MaximumLength(8)<br/>              </p>   |
| <p><i>public</i> <i>Code</i><br/><b>Birth Country</b></p>       | <p>&lt;&lt;BCC&gt;&gt;<br/> <i>Range:0 to 1</i>businessTerm=Country of Birth, Birthplace<br/>           Definition=A code representing the person's country where a person was born.<br/>           RepresentationLayout=AN(4)<br/> <i>Notes: Verification Rules and Examples</i><br/>           NOTE: If this field is blank the address is by default an Australian address. The field size accommodates different length codes used by different standards. ISO 3166 allows for both 2 or 3 character codes while some others are 4 characters.<br/><br/>           The recommended Domain Values is the list of Country Name Codes in ISO 3166. Other Domain Values may be used though should be accompanied by metadata to perform the conversion from the abbreviation to the full description as agreed by the involved parties.<br/><br/>           Examples:<br/>           AUSTRALIA AUS<br/>           AUSTRIA AT<br/>           NEW ZEALAND NZ<br/><br/>           Discussion AUSTRALIA should not be printed on domestic mail.<br/><br/>           Mail for Australia Island Territories (e.g. Christmas Island, Norfolk Island) is treated as Australian domestic mail with the name of the island included as the Locality information.<br/> <i>Constraints:</i><br/>           OCL self.Content.MaximumLength(4)<br/>              </p> |
| <p><i>public</i> <i>Code</i><br/><b>Citizenship Country</b></p> | <p>&lt;&lt;BCC&gt;&gt;<br/> <i>Range:0 to 1</i>businessTerm=-<br/>           Definition=A code name representing a country that has conferred a person's citizenship.<br/>           RepresentationLayout=AN(4)<br/> <i>Notes: Verification Rules and Examples</i><br/>           NOTE: If this field is blank the address is by default an Australian address. The field size accommodates different length codes used by different standards. ISO 3166 allows for both 2 or 3 character codes while some others are 4 characters.<br/><br/>           The recommended Domain Values is the list of Country Name Codes in ISO 3166. Other Domain Values may be used though should be accompanied by metadata to perform the conversion from the abbreviation to the full description as agreed by the involved parties.<br/><br/>           Examples:<br/>           AUSTRALIA AUS<br/>           AUSTRIA AT<br/>           NEW ZEALAND NZ</p>   |

|   |  |
|---|--|
|   | <p>Discussion AUSTRALIA should not be printed on domestic mail.</p> <p>Mail for Australia Island Territories (e.g. Christmas Island, Norfolk Island) is treated as Australian domestic mail with the name of the island included as the Locality information.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(4)</p>   |
| <i>public</i> <b>Date</b><br><b>Death</b>               | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:</i>0 to 1<i>businessTerm</i>=Date of Death<br/><i>Definition</i>=The date when a person died as recorded on the death certificate or other official documents.<br/><i>RepresentationLayout</i>=YYYYMMDD</p> <p><i>Notes:</i> Verification Rules and Examples<br/>20060217, where<br/>20 = Two digits Century value (YY)<br/>06 = Two digits Year value (YY)<br/>02 = Two digits Month value (MM)<br/>17 = Two digits Day value (DD)<br/>NOTE: In cases where all or some of the Death Date components are not known or where the Death Date represents an estimated value derived from for eg. Forensic evidence, a valid calendar date must be recorded together with and agency / industry specific date accuracy indicator.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(8)</p>   |
| <i>public</i> <b>Code</b><br><b>Gender</b>              | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:</i>0 to 1<i>businessTerm</i>=None<br/><i>Definition</i>=A code indicating the distinction between male and female as declared by a person who, for whatever reason prefers to state their gender as opposed to their sex.<br/><i>RepresentationLayout</i>=A(1)</p> <p><i>Notes:</i> Verification Rules and Examples<br/>NOTE: Code 3 - Intersex or Indeterminate (ABS 1286.0) should only be used if the person or respondent volunteers that the person is intersex or where it otherwise becomes clear during the collection process that the individual is neither male nor female. Indeterminate is normally used for babies for whom sex has not been determined for whatever reason.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(1)</p>   |
| <i>public</i> <b>Code</b><br><b>Occupation</b>          | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:</i>0 to 1<i>businessTerm</i>=Profession, Trade<br/><i>Definition</i>=A code to identify a person's occupation.<br/><i>RepresentationLayout</i>=N(6)</p> <p><i>Notes:</i> Verification Rules and Examples<br/>The coding structure has five level hierarchal levels:<br/>Level 1 – Major (1st Digit)<br/>Level 2 – Sub-Major (2nd Digit)<br/>Level 3 – Minor (3rd Digit)<br/>Level 4 – Unit Group (4th Digit)<br/>Level 5 – Occupation (5th &amp; 6th Digits)<br/>A lower numbered level is required for the next level to exist. A Level 5 occupation cannot exist without a Level 4 Unit Group; a Level 4 unit group cannot exist without a Level 3 Minor.<br/>Examples:<br/>230300 – Specialist Medical Practitioner<br/>230311 – Anaesthetist<br/>230325 – Paediatrician.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(6)</p> |
| <i>public</i> <b>Code</b><br><b>Relationship Status</b> | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:</i>0 to 1<i>businessTerm</i>=Marital Status</p>   |

|  |  |
|--|--|
|  | <p>RepresentationLayout=A(1)</p> <p><i>Notes:</i> D = Divorced<br/> E = Engaged<br/> F = De Facto<br/> M = Married<br/> N = Single (excludes D, S and W)<br/> S = Separated<br/> U = Civil Union (the same sex couple)<br/> W = Widowed</p> <p><i>Constraints:</i><br/> OCL self.Content.MaximumLength(1)</p>  |
| <p><i>public</i> <u>Code</u><br/> <b>Sex</b></p> | <p>&lt;&lt;BCC&gt;&gt;</p> <p><i>Range:</i> 0 to 1businessTerm=None</p> <p><i>Definition:</i> A code indicating the biological distinction between male and female as reported by a person or as determined by an interviewer. A person's sex may change during their lifetime</p> <p>RepresentationLayout=A(1)</p> <p><i>Notes:</i> Verification Rules and Examples<br/> NOTE: Code 3 - Intersex or Indeterminate (ABS 1286.0) should only be used if the person or respondent volunteers that the person is intersex or where it otherwise becomes clear during the collection process that the individual is neither male nor female. Indeterminate is normally used for babies for whom sex has not been determined for whatever reason.</p> <p><i>Constraints:</i><br/> OCL self.Content.MaximumLength(1)</p> |

## AS4590 – PERSON NAME ATTRIBUTES

### Person Name Attributes

| Attribute  | Details  |
|--|--|
| <p><i>public Text</i><br/><b>Family Name</b></p> | <p>&lt;&lt;BCC&gt;&gt;<br/>businessTerm=Surname, Last Name<br/>Definition=A person's name that is either: The hereditary or tribal surname of a person's family, Acquired by a person in accordance with a due process defined in a State or Territory Act. Any other name distinguished from a person Given Name<br/>RepresentationLayout=(40)</p> <p><i>Notes: Verification Rules and Examples</i><br/>There are no universal verification rules for a person Family Name.<br/>NOTE: This data item may be repeated if a person offers more than one Family Name.<br/>A useful resource when capturing ethnic names is the annexed Ethnic Names Condensed Guide for recording ethnic names produced by Centrelink, Canberra, AGPS</p>  |
| <p><i>public Text</i><br/><b>Full Name</b></p>   | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to 1</i>businessTerm=None<br/>Definition=The full name of a person.<br/>RepresentationLayout=(500)</p> <p><i>Notes: Verification Rules and Examples</i><br/>This field allows the full name of a client to be interchanged as a string of text, including, but not limited to, family name and given names. A great deal of information about a person is able to be extracted from a full name, information which would otherwise be lost when a Western style format is imposed. This is particularly important with such names as Arabic and Chinese. Additionally, by using Unicode, names can readily be interchanged as represented in the native language.<br/>This field should only be used when agreement has been reached, between the organisations interchanging information, that use of the field is appropriate and needed for their purpose.<br/>There are no universal verification rules for a Full Name.<br/>A useful resource when capturing ethnic names is the annexed Ethnic Names Condensed Guide for recording ethnic names produced by Centrelink</p> |
| <p><i>public Text</i><br/><b>Given Name</b></p>  | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to *</i>businessTerm=First Name, Forename, Christian Name, Middle Name, Second Name, Other Given Name<br/>Definition=A person's name that is either: Assigned by a person parents shortly after birth or adoption. Acquired by a person in accordance with a due process defined in a State or Territory Act. Attained by a person within the family group.<br/>RepresentationLayout=(40)</p> <p><i>Notes: Verification Rules and Examples</i><br/>There are no universal verification rules for a person Given Name.<br/>NOTE: This data type may be repeated if a person offers more than one Given Name.<br/>If a person has only one name it should be recorded as the Family Name not the Given Name.<br/>A useful resource when capturing ethnic names is the annexed Naming Systems of Ethnic Groups produced by Centrelink, Canberra, AGPS.</p>  |
| <p><i>public Text</i><br/><b>Name Suffix</b></p> | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to *</i>businessTerm=Post nominal<br/>Definition=An affix which follows the element to which it is added. Honours, awards and other denominations that follow a person name, usually an acronym or abbreviation<br/>RepresentationLayout=(12)</p> <p><i>Notes: Verification Rules and Examples</i><br/>NOTE: This data type may be repeated where more than one Name Suffix is associated with a person, eg. Queens Counsel and Justice of the Peace (QC, JP).</p>   |
| <p><i>public Code</i><br/><b>Title</b></p>       | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to *</i>businessTerm=Salutation<br/>Definition=A prefix to a person name. An honorific form of address commencing a name, used when addressing a person by Family Name, whether by mail, phone or in</p>   |

|  |  |
|--|--|
|  | <p>RepresentationLayout=(12)</p> <p><i>Notes:</i> Verification Rules and Examples<br/> The Name Title should not be confused with a person Job Title.<br/> NOTE: This data type may be repeated where more than one Name Title is associated with a person, eg. Honourable Doctor (Hon Dr),</p>  |
| <p><i>public Code</i><br/> <b>Usage Type</b></p> | <p>&lt;&lt;BCC&gt;&gt;</p> <p><i>Range:</i> 0 to 1businessTerm=None</p> <p>Definition=The code for the usage type of a person Family Name and Given Name that enables to differentiate between the roles of each recorded or interchanged person's name.</p> <p>RepresentationLayout=A(3)</p> <p><i>Notes:</i> Verification Rules and Examples</p> <p>Also Known As (or Aliases).<br/> This type denotes any other name that a person is also known by, or has been known by in the past. This includes misspelt names or name variations that are to be retained as they have been used to identify this person. More than one alias name may be recorded for a person.</p> <p>Maiden Name.<br/> To be used for females only</p> <p>New Born<br/> This type is reserved for the identification of unnamed newborn babies.</p> <p>Preferred Name<br/> This type is to be associated the name by which the person chooses to be identified.</p> |

## AS4590 – ADDRESS ATTRIBUTES

### Address Attributes

| Attribute   | Details   |
|---|---|
| <p><i>public Code</i><br/><b>Status</b></p>               | <p>&lt;&lt;BCC&gt;&gt;<br/> <i>Range:0 to 1</i>businessTerm=Address Official Standing Code<br/>           Definition=To identify the category of the address, as officially assigned by an addressing authority authorised by the Jurisdictional State or Territory, or Australia Post for postal type addresses.<br/>           RepresentationLayout=A(3)<br/> <i>Constraints:</i><br/>           OCL self.Content.MaximumLength(3)</p>  |
| <p><i>public Code</i><br/><b>Client Currency</b></p>      | <p>&lt;&lt;BCC&gt;&gt;<br/> <i>Range:0 to 1</i>businessTerm=None<br/>           Definition=Details the relationship between the client and the associated address at the point in time of interchange<br/>           RepresentationLayout=A(1)<br/> <i>Notes: Verification Rules and Examples</i><br/>           The examples provided list the recommended values. Use of additional or alternative domain values should be accompanied by metadata as agreed by the involved parties.<br/>           Examples<br/>           Prior: P<br/>           Current: C<br/>           Future: F<br/>           Temporary: T<br/> <i>Constraints:</i><br/>           OCL self.Content.MaximumLength(1)</p>  |
| <p><i>public Code</i><br/><b>Country Name</b></p>         | <p>&lt;&lt;BCC&gt;&gt;<br/> <i>Range:0 to 1</i>businessTerm=None<br/>           Definition=A code indicating the country, territory, colony or dependency for an address.<br/>           RepresentationLayout=AN(4)<br/> <i>Notes: Verification Rules and Examples</i><br/>           NOTE: If this field is blank the address is by default an Australian address. The field size accommodates different length codes used by different standards. ISO 3166 allows for both 2 or 3 character codes while some others are 4 characters.<br/>           The recommended Domain Values is the list of Country Name Codes in ISO 3166. Other Domain Values may be used though should be accompanied by metadata to perform the conversion from the abbreviation to the full description as agreed by the involved parties.<br/>           Examples:<br/>           AUSTRALIA AUS<br/>           AUSTRIA AU<br/>           NEW ZEALAND NZ<br/>           Discussion AUSTRALIA should not be printed on domestic mail.<br/>           Mail for Australia Island Territories (e.g. Christmas Island, Norfolk Island) is treated as Australian domestic mail with the name of the island included as the Locality information.<br/> <i>Constraints:</i><br/>           OCL self.Content.MaximumLength(4)</p> |
| <p><i>public Identifier</i><br/><b>Delivery Point</b></p> | <p>&lt;&lt;BCC&gt;&gt;<br/> <i>Range:0 to 1</i>businessTerm=None<br/>           Definition=A number created by Australia Post for an address.<br/>           RepresentationLayout=N(8)<br/> <i>Notes: Verification Rules and Examples</i></p>   |

|  |   |
|--|---|
|  | <p>Example:</p> <p>77220761</p> <p>Delivery Point Identifier for 321 Exhibition St, MELBOURNE VIC 3000</p> <p>The DPID is the intellectual property of Australia Post and may only be assigned to an address using a current AMAS approved product. The DPID is used in the process of bar coding mail. For postal purposes, the DPID should be re-validated every 3 months.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(8)</p>   |
| <p><i>public</i> <u>Date</u><br/><b>Start</b></p>  | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:</i>0 to 1businessTerm=None<br/>Definition=The time the address began to be used.<br/>RepresentationLayout=HHMMSS</p> <p><i>Notes:</i> Verification Rules and Examples<br/>The numeric value nine (9) may be used where all or part of the time is unknown.<br/>The time is to represent the hours, minutes and seconds past midnight. The hour is to be recorded using 24 hour notation.<br/>Validation rules should be applied to ensure that the date is valid, in terms of format and value correctness.<br/>Examples are as follows:<br/>163752: This indicates that the start time for the address record is the 16th hour, 37th minute and 52nd second.<br/>129999: This indicates that the start time for the address record is the 12th hour with unknown minutes and seconds.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(6)</p>        |
| <p><i>public</i> <u>Date</u><br/><b>End</b></p>    | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:</i>0 to 1businessTerm=None<br/>Definition=The time the address stopped being used.<br/>RepresentationLayout=HHMMSS</p> <p><i>Notes:</i> Verification Rules and Examples<br/>The numeric value nine (9) may be used where all or part of the time is unknown.</p> <p>NOTE: The time is to represent the hours, minutes and seconds past midnight. The hour is to be recorded using 24 hour notation.<br/>Validation rules should be applied to ensure that the date is valid, in terms of format and value correctness.<br/>Examples are as follows:<br/>163752: This indicates that the end time for the address record is the 16th hour, 37th minute and 52nd second.<br/>129999: This indicates that the end time for the address record is the 12th hour with unknown minutes and seconds.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(6)</p> |
| <p><i>public</i> <u>Text</u><br/><b>Line 1</b></p> | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:</i>0 to 1businessTerm=Exceptional address line 1.<br/>Definition=The first line of an unstructured Australian address.<br/>RepresentationLayout=X(50)</p> <p><i>Notes:</i> Verification Rules and Examples<br/>NOTE: Where there is a need to transfer an Australian address that does not conform with the commonly used address format, the Unstructured Address Lines may be used. These fields should not be used except when it is impossible to use of the other more structured address fields.<br/>While 4 lines have been provided for unstructured Australian address details, not all lines need to be used. Unstructured Address Lines, when used, should contain the entire address.<br/>Examples:<br/>Cabin 44 Block 7 (Unstructured Address Line 1)<br/>UMAS Watson (Unstructured Address Line 2)</p>   |

|  |  |
|--|--|
|  | <p>Watson Bay Wharf (Unstructured Address Line 3)<br/>WATSONS BAY NSW 2030 (Unstructured Address Line 4)</p> <p>TSS 5 AVN REGT (Unstructured Address Line 1)<br/>RAAF BASE TOWNSVILLE (Unstructured Address Line 2)<br/>TOWNSVILLE QLD 4810 (Unstructured Address Line 3)</p> <p>Joes Fruit Juice Shop (Unstructured Address Line 1)<br/>Food Court (Unstructured Address Line 2)<br/>Chadstone Shopping Centre (Unstructured Address Line 3)<br/>CHADSTONE VIC 3148 (Unstructured Address Line 4)<br/>The examples above demonstrate poor addressing and should be avoided.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(50)</p> |
| <p><i>public Text</i><br/><b>Line 2</b></p>              | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to 1</i>businessTerm=Exceptional address line 2.<br/>Definition=The second line of an unstructured Australian address<br/>RepresentationLayout=X(50)<br/><i>Notes:</i> Verification Rules and Examples<br/>Refer to Line 1.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(50)</p>  |
| <p><i>public Text</i><br/><b>Line 3</b></p>              | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to 1</i>businessTerm=Exceptional Address Line 3<br/>Definition=The third line of an unstructured Australian address<br/>RepresentationLayout=X(50)<br/><i>Notes:</i> Verification Rules and Examples<br/>Refer to Line 1.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(50)</p>  |
| <p><i>public Text</i><br/><b>Line 4</b></p>              | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to 1</i>businessTerm=Exceptional Address Line 4<br/>Definition=The fourth line of an unstructured Australian address<br/>RepresentationLayout=X(50)<br/><i>Notes:</i> Verification Rules and Examples<br/>Refer to Line 1.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(50)</p>   |
| <p><i>public Text</i><br/><b>Locality Name</b></p>       | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to 1</i>businessTerm=Suburb Name<br/>Definition=The name of the locality/suburb of the address.<br/>RepresentationLayout=X(46)<br/><i>Notes:</i> Verification Rules and Examples<br/>Examples:<br/>RICHMOND<br/>KIPPA-RING</p> <p>Usage: For mailing purposes the format of this field should be upper case. Refer to Australia Post Presentation Standard. Any forced abbreviations shall be done by truncation from the right.<br/>Discussion: Official locality names are assigned by relevant state naming committees/protocols.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(46)</p>   |
| <p><i>public Text</i><br/><b>Location Descriptor</b></p> | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to 1</i>businessTerm=None<br/>Definition=A free text field to describe the position of the address relative to</p>   |

|  |   |
|--|---|
|  | <p>RepresentationLayout=X(30)</p> <p><i>Notes:</i> Verification Rules and Examples<br/> Examples:<br/> NEAR THE NORTHBRIDGE OVERPASS<br/> Via Blackmans Rd<br/> OFF PRINCESS ST<br/> Rear 150 Smith St<br/> OVER SWANPORT BRIDGE<br/> 3km PAST THE BLACK STUMP SIGN<br/> DIAGONALLY OPPOSITE TOWN HALL<br/> CORNER SMITH STREET</p> <p><i>Constraints:</i><br/> OCL self.Content.MaximumLength(30)</p>  |
| <p><i>public Text</i><br/> <b>Postal Delivery Number</b></p> | <p>&lt;&lt;BCC&gt;&gt;<br/> businessTerm=None<br/> Definition=Identification number for the channel of postal delivery<br/> RepresentationLayout=AN(11)</p> <p><i>Notes:</i> Verification Rules and Examples<br/> NOTE: Is used in conjunction with a Postal Delivery Type Code. For display purposes, in the format, Postal Delivery Type Code &lt;space&gt; Postal Delivery Number.<br/> Examples::<br/> PO BOX C96<br/> RMB 123<br/> (Postal Delivery Number is C96)<br/> (Postal Delivery Number is 123)<br/> NOTE: Not all postal delivery types have a postal delivery number. A postal delivery number is mandatory for all postal delivery types other than:<br/> CARE PO<br/> CMA<br/> CMB<br/> CPA<br/> No associated postal delivery number<br/> No associated postal delivery number<br/> Optional<br/> No associated postal delivery number<br/> Verification Rules and Examples<br/> Discussion: This is used where mail is being sent to an area where normal mail delivery is unavailable, or not preferred. Additionally it may be used in some rural areas where no other formal addressing structure exists to identify delivery addresses.</p> <p><i>Constraints:</i><br/> OCL self.Content.MaximumLength(11)</p> |
| <p><i>public Code</i><br/> <b>Postal Delivery Type</b></p>   | <p>&lt;&lt;BCC&gt;&gt;<br/> Range:0 to 1businessTerm=None<br/> Definition=Identification for the channel of postal delivery<br/> RepresentationLayout=A(11)</p> <p><i>Notes:</i> Verification Rules and Examples<br/> The recommended Domain values is the list of Postal Delivery Type Codes in the Australia Post Address Presentation Standards.<br/> Other Domain values may be used though should be accompanied by metadata to perform the conversion from the abbreviation to the full description as agreed by the involved parties. This may be required where non Australia Post deliveries are performed. Eg. Private Box at a Service Station.<br/> Usage: Used where mail is to be delivered to a box, bag or agent for pick-up by the intended recipient or to the rural mail box number where no other address exists.<br/> Discussion: This is used where mail is being sent to an area where normal mail delivery is unavailable, or not preferred. Additionally it may be used</p>  |

|   |  |
|---|--|
|   | <p>in some rural areas where no other formal addressing structure exists to identify delivery addresses.</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(11)</p>   |
| <p><i>public Identifier</i><br/><b>Postcode</b></p> | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to 1</i>businessTerm=None<br/>Definition=The four digit numeric identifier used for Postal purposes.<br/>RepresentationLayout=N(4)</p> <p><i>Notes:</i> Verification Rules and Examples<br/>NOTE: Printed Postcodes should display leading zeroes.<br/>Examples:<br/>(postcode for BRUNSWICK, VIC)<br/>Verification Rules and Examples<br/>0800 (postcode for DARWIN, NT)<br/>Australian postal addresses should include a valid Postcode.<br/>Refer to the Australia Post Address Presentation Standard for rules on presentation and positioning of postcodes on mail.<br/>For a full list of Australian Postcodes visit the Australia Post website:<br/><a href="http://www.auspost.com.au">www.auspost.com.au</a></p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(4)</p>   |
| <p><i>public Code</i><br/><b>Purpose</b></p>        | <p>&lt;&lt;BCC&gt;&gt;<br/><i>Range:0 to 1</i>businessTerm=Address Use Code<br/>Definition=The role or use of the address in relation to the client.<br/>RepresentationLayout=A(3)</p> <p><i>Notes:</i> Verification Rules and Examples<br/>The Address Purpose Code can only exist if an associated address has been entered.<br/>The examples provided list the recommended values. Use of additional or alternative domain values may be used and should be accompanied by metadata as agreed by the involved parties.<br/>Examples include:<br/>Primary property address: The property address normally used by the client.<br/>NOTE: where this is the principal place of residence of the client the Residential code should be used.<br/>Secondary property address: The address of an additional property attached to the client.<br/>Residential: The address of the principal place of residence for the client.<br/>Temporary Accommodation: The address where the client is resident for a temporary period.<br/>Business: The address of the principal place of business for the client.<br/>Overseas address: The address used by the client when overseas.<br/>Delivery address: The address used for goods delivery purposes.<br/>Postal/Correspondence: The address used by the client for receipt of correspondence.<br/>Where the Address Purpose is not stated or unknown a NULL entry will be recorded. This indicates that the address can be expected to be used as a residential and a correspondence address by default.<br/>An address may have more than one purpose.<br/>Example codes:<br/>Primary property address PR<br/>Secondary property address SEC<br/>Residential RES<br/>Temporary Accommodation TEM<br/>Business BUS<br/>Overseas address OVS<br/>Delivery address DEL<br/>Postal / Correspondence address POS<br/>Not stated / unknown (NULL)</p> <p><i>Constraints:</i><br/>OCL self.Content.MaximumLength(3)</p> |

|  |   |
|--|---|
| <p><b>State Territory</b></p>              | <p><i>Range:0 to 1businessTerm=None</i><br/> Definition=The State or Territory code of the address.<br/> RepresentationLayout=A(3)</p> <p><i>Notes:</i> Verification Rules and Examples:<br/> Examples Australian Capital Territory<br/> Tasmania<br/> ACT<br/> TAS</p> <p><i>Constraints:</i><br/> OCL self.Content.MaximumLength(3)</p>   |
| <p><i>public Code</i><br/> <b>Type</b></p> | <p>&lt;&lt;BCC&gt;&gt;</p> <p><i>Range:0 to 1businessTerm=None</i><br/> Definition=This element is to distinguish between a physical address or other.<br/> RepresentationLayout=A(3)</p> <p><i>Notes:</i> Verification Rules and Examples<br/> The Address Type Code can only exist if an associated address has been entered.<br/> The examples provided list the recommended values. Use of additional or alternative domain values may be used and should be accompanied by metadata as agreed by the involved parties.<br/> Examples include:<br/> Physical: A property address (e.g. not a Post Office Box)<br/> Other: Not a physical address (e.g. PO Box)<br/> Where the Address Type is not stated or unknown a NULL entry will be recorded.<br/> Example codes:<br/> Other<br/> Physical address<br/> Not stated / unknown<br/> Examples:<br/> 12 Smith St<br/> PO Box<br/> OTH<br/> PHY<br/> (NULL)<br/> PHY<br/> OTH</p> <p><i>Constraints:</i><br/> OCL self.Content.MaximumLength(3)</p> |

**Data Matching Working Group**  
**Agency Data Matching Activity Profile**  
for  
Sample Agency

**Table of Contents**

1 ..... NAME OF AGENCY

2 ..... AGENCY MISSION STATEMENT (FUNCTION OF AGENCY)

3 ..... PURPOSE OF DATA MATCHING

4 ..... CONSEQUENCES OF MATCHING ERROR

5 ..... SOURCE OF DATA

6 ..... TYPE OF MATCHING TECHNIQUES USED

7 ..... DATA MODELS

8 ..... ASSESSMENT OF TECHNIQUES USED

9 ..... FUTURE PATH OF AGENCY

## **Name of Agency**

*The name of the Agency*

## **Agency Mission Statement (Function of Agency)**

*The purpose of the Agency*

## **Purpose of Data Matching**

*Why the agency performs Data Matching and how it relates to the Agencies function.*

## **Consequences of Matching Error**

*What happens if data matching errors occur.*

*For example:*

### ***False Negative Matches***

- *Implications include...*

### ***False Positive Matches:***

- *Implications include...*

## **Source of Data**

*Listing of the sources of data used in matching;*

*Also give an indication as the 'quality' of data received and any comments about this data which may be useful to other agencies.*

## **Type of Matching Techniques Used**

*Detail the various data matching techniques used.*

*Example*

*Name Matching*

*Address matching*

## **Data Models**

Data models/descriptions of the data involved.

Examples

***In-House Data***

| <b><i>Data 'Field'</i></b>        | <b><i>Maximum Size</i></b> |
|-----------------------------------|----------------------------|
| <b><i>Name</i></b>                |                            |
| <i>Surname</i>                    | 25                         |
| <i>Given Names</i>                | 25                         |
| <i>DoB</i>                        | 8                          |
| <i>Sex</i>                        | 1                          |
| <b><i>Residential Address</i></b> |                            |
| <i>Habitation Name</i>            | 25                         |
| <i>Flat Number</i>                | 6                          |
| <i>Street Number</i>              | 6                          |
| <i>Street Name</i>                | 25                         |
| <i>Street Type</i>                | 7                          |
| <i>Locality</i>                   | 25                         |
| <i>Postcode</i>                   | 4                          |
| <i>State</i>                      | 3                          |
| <b><i>Postal Address</i></b>      |                            |
| <i>Line 1</i>                     | 36                         |
| <i>Line 2</i>                     | 25                         |
| <i>Line 3</i>                     | 25                         |

***External data***

| <b><i>Data 'Field'</i></b>        | <b><i>Maximum Size</i></b> |
|-----------------------------------|----------------------------|
| <b><i>Name</i></b>                |                            |
| <i>Surname</i>                    | 40                         |
| <i>Given Name 1</i>               | 40                         |
| <i>Given Names other</i>          | 0 – 40                     |
| <i>DoB</i>                        | 8                          |
| <i>Sex</i>                        | 1                          |
| <b><i>Residential Address</i></b> |                            |
| <i>Habitation Name</i>            | 40                         |
| <i>Flat Number</i>                | 4-10                       |
| <i>Street Number</i>              | 4-10                       |
| <i>Street Name</i>                | 40                         |
| <i>Street Type</i>                | 5-10                       |
| <i>Locality</i>                   | 40                         |
| <i>State</i>                      | 3                          |

## **Assessment of Techniques Used**

*An assessment of the different matching techniques used including the good and bad points of each techniques and how they could be improved.*

## **Future Path of Agency**

*Where the agency would like to go in the future with data matching. Include in this anything already in the planning stage or just plain aspirations.*