Batch cleansing of postal addresses

A guide for New Zealand businesses

New Zealand Post

1. Introduction

This is a guideline for businesses evaluating options to conduct batch cleansing of postal addresses. It will be particularly useful for those who are conducting batch cleansing for the first time, where the level of changes could be relatively high, compared with subsequent cleansing further down the track.

The paper helps to explain the following:

- How batch address cleansing software works
- Suggested framework for cleansing
- Dealing with low confidence and non matches to reference data
- Options for keeping addresses clean

2. How batch address cleansing software works

The Postal Address File (PAF)[1] is New Zealand Post's definitive listing of postal addresses in New Zealand. This file can be used as a reference data set by address cleansing software to cleanse (i.e., standardise, validate and postcode) mailing addresses.

You can choose to build your own cleansing application. However, the decision to do this should be weighed against the economics of licensing proven products from data service providers.

Many of those providers are able to offer value added services such as house-holding, single customer views, de-duping, validating records against national change of address and deceased databases, and mesh-blocking. These functions[2], along with address cleansing, can be licensed as software, accessed via online 'data portal' services, or run by a 'bureau service'.

The processing of data by some of the more common cleansing software and services typically follows three basic steps:

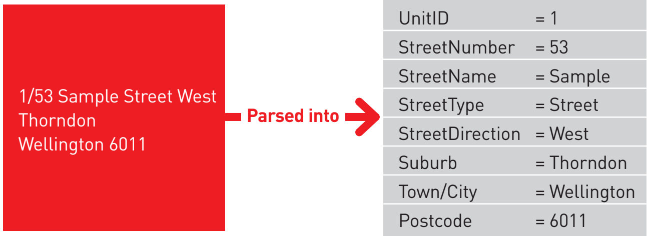
1. Parse an input set of addresses data into their basic elements

2. Standardise, validate and postcode the data against the PAF

3. Output the data, and indicate which records are:
   - Matched with and without changes to the original record
   - Unmatched with and without changes to at least improve the data

Parsing

Parsing an address into its basic elements such as street number, street name, suburb, etc., allows it to be easily matched with the parsed data structure of the PAF. For example:



1/53 Sample Street West
Thorndon
Wellington 6011

Parsed into

UnitID	= 1
StreetNumber	= 53
StreetName	= Sample
StreetType	= Street
StreetDirection	= West
Suburb	= Thorndon
Town/City	= Wellington
Postcode	= 6011

Validating

Logic built into many commercially available cleansing software handles most of the commonly encountered issues encountered in low quality addresses. For example, missing or duplicated address elements, common misspellings or abbreviations, and non-address information included in the address fields.

When a successful match has been made, the PAF Delivery Point Identifier (DPID) can be assigned to the matched record. The DPID is a 7-digit number that uniquely identifies each postal delivery address.

Output

It is common practice for the original data to be returned along with the cleansed output. This allows for the following:

- Visual verification of changes to ensure they are correct and logical
- Comparison of the original input data with the same record in the source system to check if anything has changed (e.g. customer changed residence recently) during the time of the cleansing

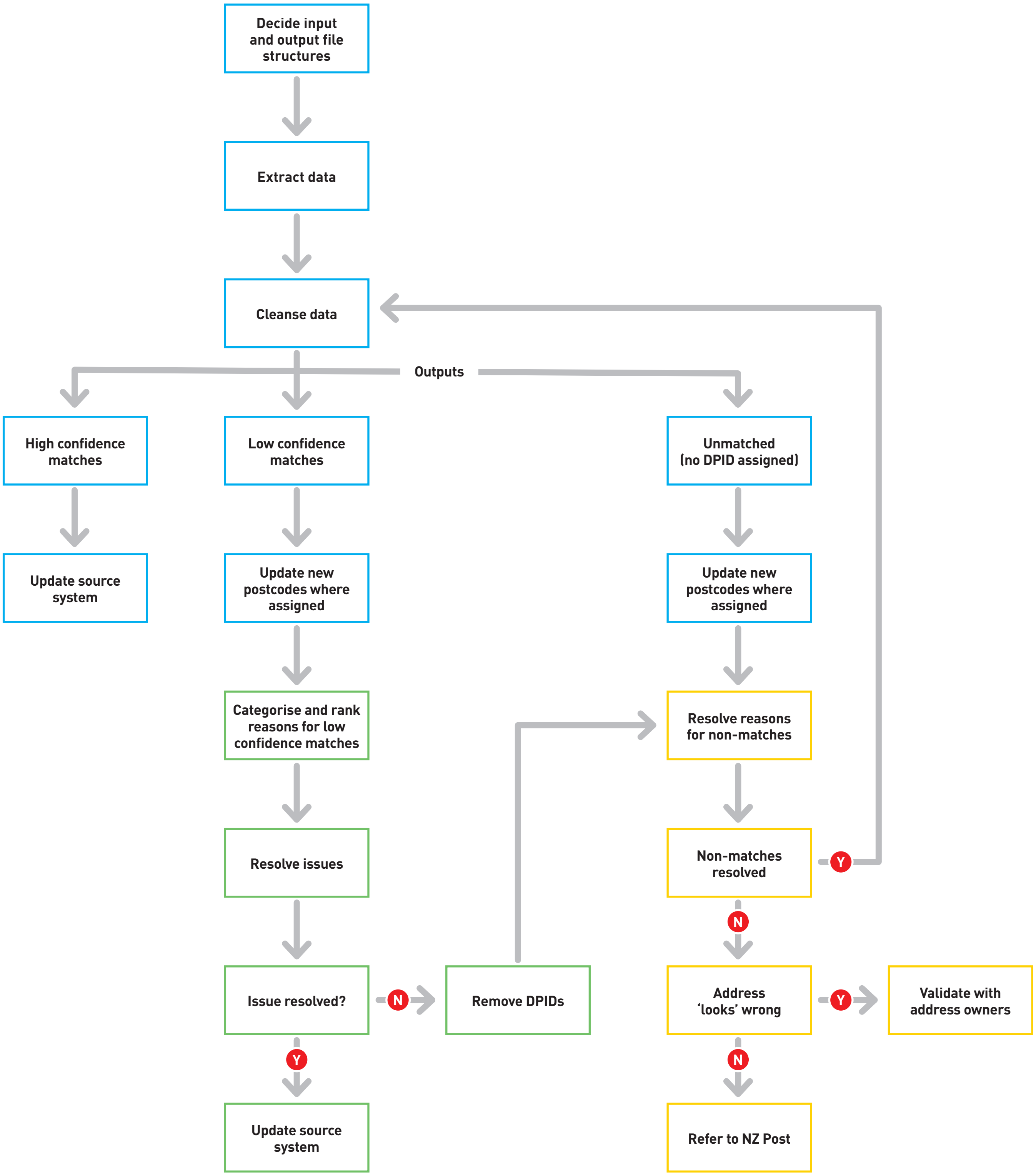
The last point may be particularly relevant for systems where an IT change-freeze is not feasible during the cleansing process.

[1] For more details on the PAF, please refer to www.nzpost.co.nz/sendright.

[2] A listing of Data Industry Support can be found at www.nzpost.co.nz/sendright.

3. Suggested framework for address cleansing

Cleansing a database containing several thousand records, let alone a few hundred thousand records can seem like a daunting task. The following is a suggested framework for tackling this task, based largely on the learnings from within New Zealand Post, and consultation with the data industry.


Decide input and output file structures
Extract data
Cleanse data
Outputs
High confidence matches
Low confidence matches
Unmatched (no DPID assigned)
Update source system
Update new postcodes where assigned
Update new postcodes where assigned
Categorise and rank reasons for low confidence matches
Resolve reasons for non-matches
Resolve issues
Non-matches resolved
Y
N
Issue resolved?
N
Remove DPIDs
Address 'looks' wrong
Y
Validate with address owners
Y
N
Update source system
Refer to NZ Post

This approach can just as easily be applied to a test run, pilot, or the full production cleanse of an address database. It is a phased approach where:

🟦 Phase 1 – Cleansed, high confidence matched addresses, as well as postcodes are updated into the system.
🟩 Phase 2 – Low confidence matches are resolved and updated into source system.
🟨 Phase 3 – Unmatched records are resolved and re-cleansed.

The phased approach enables quick wins to be achieved and where postal discounts and other benefits are concerned, reap business benefits very early on. For a description of business benefits please refer to 'Quality addressing – a guide for New Zealand businesses'.

3.1 Decide input and output file structures

The first step of the process involves deciding the structures of input and output data. Some of the more sophisticated software in the market allows full flexibility – including free format to fully parsed input and output. No two source address systems are usually the same and so the advantage of having such flexibility is obvious.

It is often only after conducting test cleansing that a decision can be made on the best input and output structures to use. Some of the considerations include:

- Is the address data stored in free format fields, fully parsed or something in between (e.g. Address Line 1, Address Line 2, Suburb, City, Postcode)?
- Is there a preference for receiving fully parsed output to provide full flexibility in rebuilding the data back into the source system?
- Should related non-address fields (such as recipient details) be included in the input data, just in case some valid address information may have been accidentally stored in those fields?
- Should the suburb name in your database be included in the output, or will you use the postally preferred suburb from the PAF?

3.2 Extract data

The second step of the process involves taking an extract of the addresses from the source system. Most address cleansing software would be able to deal with common extract file formats such as pipe separated text, comma separated text, or spreadsheet files.

One of the decisions to make at this stage is whether the cleansing is required to be done on all records in the database. For example, to meet our SendRight™ requirements for bulk mail products, only active postal New Zealand addresses are required to be accurate.

Besides the address elements, the primary key or identifier for each address record (e.g. account number) will normally need to be extracted. The purpose of this key is to enable cleansed addresses to be updated to the correct records back in the source system.

A change freeze for the source system may need to be in place until the time cleansed records are updated. Alternatively, a process can be implemented to check if there are differences between the source records in the system with the extracted records just before the updates occur. If there are any, filter those records out for re-extraction and re-cleanse.

3.3 Cleanse data

It is recommended that test cleansings are conducted for:

- Gauging the initial data quality
- Reducing the risk of making incorrect changes to the address database
- Ensuring the process is well understood
- Assisting in scheduling the resources required

By visually checking the test results, it can be determined if the results are expected and output formats are correct. The results may also highlight any portion of the data that can be targeted for 'quick fixes' or need to be tidied up beforehand to maximise the match rates from the cleansing.

Some of the data issues that can be tidied up beforehand include the presence of phone numbers (e.g. search for seven or more number strings), email addresses (e.g. search for '@' character), and incorrect use of abbreviations (e.g. search for capitalised words).

As discussed earlier, the cleansing process typically involves parsing, cleansing and output.

Parsing is the routine of breaking down an address into its basic elements. The New Zealand Post PAF file address elements are:

- Street Number
- Street Alpha
- Unit Type
- Unit Identifier
- Floor Number and Identifier
- Building Name
- Street Name
- Street Type
- Street Direction
- Delivery Service Type
- Box/Bag Number
- Box/Bag Lobby
- Suburb Name
- Rural Delivery Identifier
- Town/City/Mailtown
- Postcode

The more advanced software may be able to further parse non-address elements such as:

- Company Name
- Department
- Recipient Name
- Phone Number

This ability to parse non-address elements will be very useful for systems that are likely to have non-address elements mixed into address fields.

Records that have been found to fully match the ones in PAF can be assigned a DPID.

Note that for the purpose of obtaining a SendRight™ Statement of Accuracy, only base address matching[3] is required. Base address matches are not assigned PAF DPIDs.

As a guideline, high confidence matches are generally those that have been matched with no changes required to the original records or where corrections have been made to obvious mistakes. Some of those corrections include:

- Deleting duplicate information, such as city information entered twice
- Correcting obvious spelling mistakes, e.g. from Fielding to Feilding
- Changing street type, e.g. street to avenue

[3] A base address match excludes Unit, Floor, Building Name, Street Alpha, Street Direction. For example, Flat 1 63A Smith Street West, Thorndon, Wellington could achieve a 'base' match as 63 Smith Street, Wellington.

- Adding extra elements such as a suburb, or postcode to complete an address
- Correcting delivery service types, e.g. from Private Box to PO Box
- Correcting wrong suburb/city relationship, e.g. Takapuna/ Auckland to Takapuna/North Shore City

High confidence results can be immediately updated back into the source system. It may also be useful to save the DPID information for future reference.

Low confidence matches are those where the software could only match addresses to the PAF by significantly changing the original input data. It would thus be prudent to conduct an analysis of these changes before committing them.

Non-matches are those where the address quality is so low that no DPID could be assigned to those addresses.

Tactics to deal with low confidence matches and non-matches are detailed in Sections 4 & 5 of this document.

3.3.1 Some Cleansing Considerations

Alternate suburbs: Although it is preferable that commonly used suburbs are used, it may sometimes be inappropriate to change the customer's preference. In such situations, it is acceptable to allow any suburb name, unless it is an invalid alternative as listed in the PAF.

Use of abbreviations: Acceptable abbreviations for unit type, floor type, commonly used street type and street direction are provided by New Zealand Post[4]. This information will be particularly useful for systems that have limited space in their address line fields.

Building names: Building names are not required for postal addresses if a street number is present. However, more often than not, street numbers for buildings such as shopping malls and schools are usually not provided. For such situations, the building name should be retained to aid mail delivery.

Overseas addresses: If the cleansing software is unable to cleanse overseas addresses, it would be advisable to filter them out before cleansing.

3.4 Updating the source system

Updating the source system would involve matching the primary keys of the cleansed records with the records in the source system.

A key consideration at this stage is the currency of the extracted and cleansed data. For example, if customers have moved recently, they may have requested changes to their addresses while the cleansing is still taking place. To avoid overwriting new data with the 'outdated' cleansed data, any one of the following strategies can be considered:

- Conduct the data extraction and cleansing during non-working hours
- Implement a change freeze to normal data entry processes while cleansing is happening
- Implement a process to check for differences between the source records in the system with the extracted records just before the updates occur and filter out the relevant records for re-extraction and re-cleanse at a later date

The decision will largely be based on balancing between minimising system downtime, costs, and rework.

4. Dealing with low confidence PAF matches

Low confidence matches occur when it is unclear as to whether important information may have been changed or removed to obtain a match to the PAF. Some of the issues and tips to resolve them are listed in the following table.

Issues	Tips
Important non-address elements such as 'care of' information, company or department, etc., may be removed during cleansing from the address fields.	Where the software is capable, configure the output format to include extra fields to capture this information during the parsing. Next, instigate a separate routine to move these to the appropriate fields in the source system.
The inclusion of a physical address with a postal address, e.g., 7 Waterloo Quay Private Bay 39990 Wellington 5045 Some software by default will remove the physical address.	Where the software is capable, configure the output format to include extra fields to capture the physical address information during the parsing. Instigate a separate routine to move these to the location address fields in the source system.
Where numbers in the format 'xx–xx' are encountered, it is ambiguous as to whether they should be treated as street number range, or unit/ street number. Note this would not be an issue if the front-end data capture system restricts the entry of street number ranges and forces users to use the '/' to indicate unit type address.	Occasionally, a unit address can be 'guessed' at by using the following business rules to decide if it is not a street range: • the second number is lower • one number is odd and the other even. Looking up the address in the phone book or business directory may help resolve most issues, or as a last resort, validate the address directly with the owner.

Where large databases are concerned, 'divide and conquer' can be a good approach rather than trying to validate all the above at once:

- Assign flags based on the above issues against all the low confidence matches
- Do a high-level sort and count the records based on the flags (see table below)
- Prioritise actions based on number of records and ease of change
- Develop post-cleansing routines specific for each of the top categories

For example:

Description	Count	%	Post-cleansing routine
Low Confidence Matches			
Matched; Company/Recipient Name Encountered.	210		Move Company/recipient name into Organisation field.
Matched; Ambiguous Street Range or Unit Address.	104		Check with the online postcode finder, and Telecom Whitepages or UBD directory.
Matched; Other Non-address Elements Encountered.	51		Do further analysis.
Matched; Ambiguous Street Range or Unit Address; Company/Recipient Name Encountered.	1		Conduct manual correction.
Subtotal	366	3%	

Note that there are data service providers who provide similar, if not more sophisticated, sets of analysis as an option.

[4] Listed in the 'Quick guide to addressing your letters and parcels' and 'Address and layout guide'.

5. Dealing with non-matches to PAF

Non-matched records, as the name suggests, are those which could not be matched to any records in PAF and therefore not assigned a DPID. In many cases, postcodes could still be assigned to unmatched records from a valid combination of street name, box/bag number, suburb, box lobby or city. It is recommended that where new postcodes are assigned, these are accepted.

Non-matches could be due to the reasons in the following table:

Issues	Tips
If an address is missing a street number, street name or PO Box/Private Bag number, it is almost impossible to obtain a complete match. Similarly, for nationally duplicated street names, missing suburb, town, and/or postcode information would cause similar problems.	In most cases it is necessary to verify the address with the customer or look up the phone book or business directory.
Some rural postal addresses do not have a street number allocated by the Local Authority. These addresses are currently excluded from the PAF due to the Privacy Act, as they require the name of the registered occupant to make them unique.	Over time, this is will be less of an issue for rural street addresses as street or rapid numbers become more common place for rural addresses. In the meantime, if the correct postcode could at least be assigned, it would ensure that the mail gets sorted to the correct rural delivery round.
Counter Delivery addresses are currently excluded from the PAF as they require the name of the registered person to make them unique, but due to the Privacy Act, those names cannot be included in the PAF.	A listing of all the counter delivery locations along with their postcodes can be downloaded from our website.
Community Mailbox (CMB) addresses are currently not contained in the PAF.	A listing of all the community mail box locations along with their postcodes can be downloaded from our website.
New street numbers, streets, PO Box/ Private Bag have not yet been included in the PAF.	Look up the Address and Postcode Finder on the New Zealand Post website. This reflects more current data and is thus more likely to include new addresses. If those addresses still cannot be found, refer them to the New Zealand Post Customer Service Centre.
Invalid delivery address.	If we confirm that an address is not a valid postal address and therefore not included in the PAF, check with the address owner if an alternative postal address is in use. If the customer is unsure of the correct street address they may need to check the correct legal address that the local authority provides to the owner of the property. For the rural round and mail town, or box or bag address, the owner can check the contract with us to get the correct details. It is quite common to find mail for physical locations that have been 'redirected' to go either PO Box, Private Bag, Community Mailbox, or Counter Delivery addresses. Those physical locations would thus not be in the PAF as valid delivery addresses. Sometimes the problem could be due to misspellings, including incorrect insertion or omission of spaces or hyphens in compound names e.g. street or suburb names.

6. Options for keeping address databases clean

Address cleansing is just one part of quality address management. Initially it may seem like a huge task, especially for large databases but once the initial cleanse has been completed, all that needs to be done is to keep incremental changes to address databases clean and to keep up with the updates of postal address changes either via the periodic cleansing or the online address and postcode finder.

Implementing an effective data entry validation system is the ideal way to keep addresses clean. In the absence of this, the next best option would be to conduct frequent cleansing of the database. Upfront training of system users will also help minimise the amount of cleansing work required.

We provide a 'Quick guide to addressing your letters and parcels', available from our website, that can assist in the training. It shows address data-entry rules including the correct presentation layout for addresses.

Another tool available is our online Address and Postcode Finder. This can be used to look up the correct suburb, city and postcode for a postal address. It also provides the correct address presentation layout as well as the option to see the delivery location on an electronic map.

In addition, the Postcode directory available from Postshops, our website and the New Zealand Post Customer Service Centre is a convenient method for manual lookups of postcodes.

Other resources to verify address include business or telephone directories and maps, and property information on Local Authority websites.