

Data Cleansing: Bureau versus Web Portal

A guide for New Zealand Businesses

1. Introduction

Provided is a guideline for businesses to assist in evaluating a data cleansing solution, having made a decision to outsource their data cleaning requirements.

It will be particularly helpful for those who are conducting a data cleaning exercise for the first time and are wanting to understand the differences between the two main types of out-sourced data cleaning services available.

The paper helps explain the following:

- A definition of the two services
- Data formatting requirements
- The advantages and disadvantages of each service
- Indicative pricing models
- A suggested approach for on-going cleaning requirements

2. Consideration

For many businesses, this may be the first time they have given any consideration to the level of their data quality. Running a Statement of Accuracy (SOA) is a great starting point, as this will provide an accurate percentage match of data quality as measured against the NZ Post Postal Address File (PAF).

Some SendRight Certified Partners provide SOA's or SOA software for free, which enables businesses to measure their level of data quality at any time. To do this, data would typically be required to be formatted in a particular way.

For some, this may be the determining point as to the approach taken with their data cleaning. If exporting or manipulating a data file into a required format presents a challenge then this may direct you towards one service more than the other.

While this may seem daunting at first, with a little knowledge and understanding, the process of exporting data to send out for cleansing can be very straightforward.

3. Service Definitions

Bureau Cleansing

This type of service involves sending your data to a service provider on disk or attaching it via e-mail, if it isn't a large file, and having it cleaned to NZ Post's SendRight certification standards before being returned with a Statement of Accuracy.

This type of generic cleaning is suitable for most marketing types of mailings and is a simple and powerful solution for one off, or regular cleaning requirements.

A bureau clean provides a greater degree of flexibility around file formats, file output requirements, exception management and also how non-address elements are handled.

Web Portal/Automated Service

An automated service, commonly referred to as a Web Portal, is a standardised self-service data cleaning application which also allows you to easily convert your data and get it cleaned to NZ Post SendRight™ standards.

This would normally involve registering with a service provider and forwarding or uploading your data file over the Internet, downloading it once it has been cleaned. Again this type of generic cleaning is suitable for most marketing types of mailings and is a simple solution for one off, or regular cleaning requirements.

An automated cleansing service provides less, or sometimes no flexibility around file formats, file output requirements and exception management. However what it lacks for in flexibility, it makes up for in its lower cost structure for those businesses which are comfortable and confident in formatting their data to meet the requirements of an automated service.

4. Data Format Requirements

There are many different types of file formats and database systems which data can be exported from.

For obtaining a Statement of Accuracy or sending data for cleaning, the most commonly accepted file formats are comma-separated values (or CSV; also known as a comma-separated list or comma-separated variables) or a delimited text (or txt.) file format, both of which are common on most computer platforms.

CSV file format is very simple and supported by almost all spreadsheets and database management systems. Delimited text files, with a .txt extension, are similar to CSVs (except that they separate fields with something other than a comma “,”; usually a tab or the pipe “|” character), and they are also supported by most spreadsheets and almost all database managements systems.

Files with a txt. extension are text files. In some applications you may need a text editor to be able to create such file. Most Windows platforms come with a simple yet effective program called Notepad which allows you to save files with .txt extensions.

File Structure

While some providers will accept a wide range of file formats and file structures, there is less flexibility when using an automated service and these may require specific file structures or may reject a file if not presented in the correct format.

Typically a file should be formatted with the following fields and may begin with a single header line listing the field names in this order:

Customer Number	Add 1	Add 2	Add 3	Add 4	Add 5	Add 6

Each record of a file must have at least 3 fields (Customer Number, Address Line 1 and Address Line 2).

Each record does not need to have the same number of fields; some programmes may even have a restriction on the number of fields. This would be shown in the instructions or Q&A section of each providers web site.

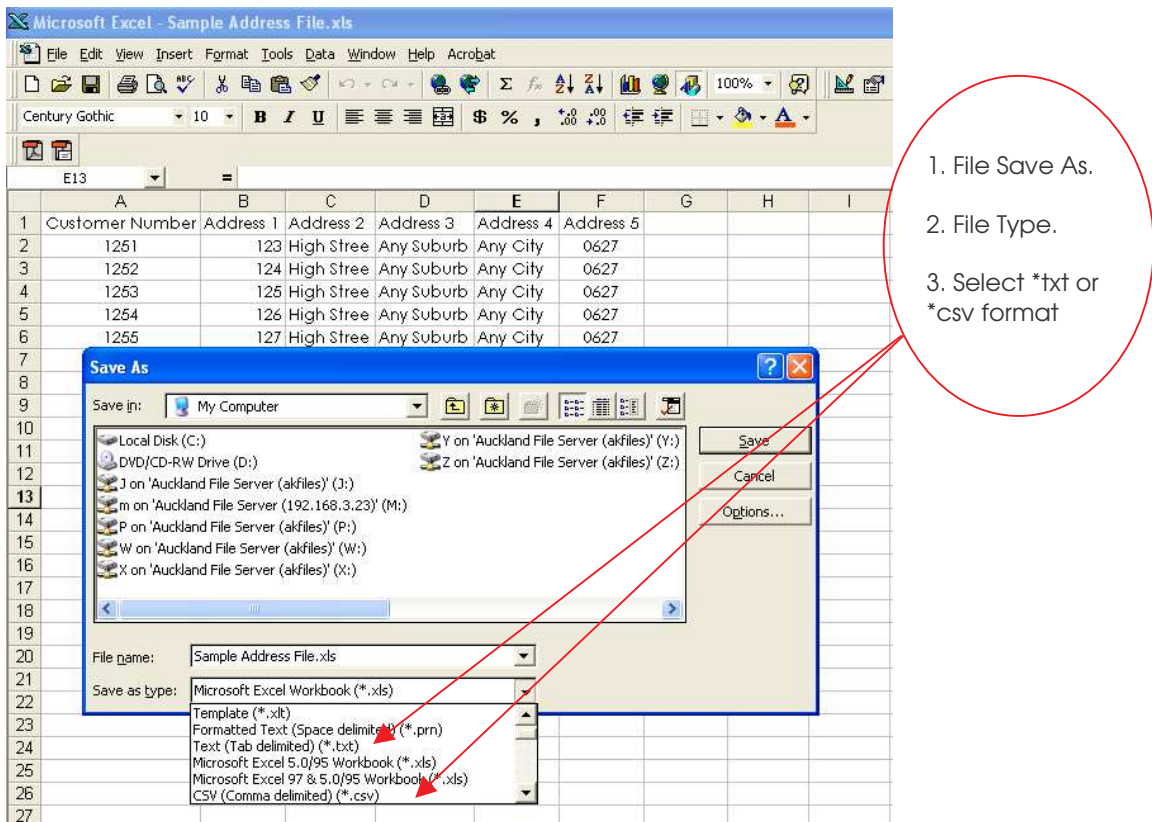
You do not need to have or assign customer numbers in your file, but you must have a field for them. An input file could look like this:

Customer Number	Add1	Add 2	Add 3	Add 4
2177	PO Box 299	Kaiapoi Central	Kaiapoi 7644	

All addresses after July 1st will need to have current postcodes and these need to be positioned immediately after the city. If postcodes are in a separate field, and it's inconvenient to attach these to the end of the city field, these can sometimes be specified as being returned in either format.

Converting to txt or CSV format

While there a multitude of operating systems and CRM applications in use, many with automated exportation and importation functions already incorporated. Converting data from simple windows-based spreadsheet application (Excel/Access Databases) is a relatively straightforward process as shown in the illustration below:



After saving the file, you can try opening the file in a simple text editor such as the Windows Notepad, and if you can see its contents correctly then it has been saved right.

5. Advantages and Disadvantages

Bureau Cleansing

A bureau service enables a greater level of human involvement and the opportunity to engage with a provider for discussing particular or unusual requirements.

Some of these requirements will allow formatting options such as:

- Appending postcodes as separate or combined fields
- Blending of multiple data sources into one completed file
- Removal of recipient data from address fields
- Appending erroneous data eg C/o's into separate columns
- Receiving data in one file format and returning it as another

A bureau clean will also enable some variability to the confidence level applied to matching rates. Some businesses will be very conservative and absolute around the changes made to its data and will require the highest level of quality, ie businesses sending out legal documents or personal financial information. Others may be more relaxed on this need or have a requirement to achieve a higher match rate.



Bureau cleaning can also provide additional data enhancements, matching and de-duping services as well tailored reporting processes which would add value to a business's marketing activities and assist in creating a single customer view.

Some of the data enhancement or matching services available are:

- National Change of Address (New Movers) Matching
- Rural Delivery Database Matching
- Appending Geo coding (lat/long co-ordinates)
- Meshblock information
- Categorisation of common issues or errors to show what exactly needed to be done on non matched data,
- Identifying which addresses are a priority for manual correction

Web Portal /Automated

An automated cleansing service provides less, or sometimes no flexibility around file formats, file output requirements and exception management.

However what it lacks in flexibility, it makes up for in its lower cost structure for those businesses which are comfortable and confident in formatting their data to meet the requirements of an automated service.

Each provider should give detailed information on what is required to upload a file and indicate a time frame for cleaning a file.

As an automated service doesn't require human involvement, it can be run at any time. This provides the benefit of being able to process files at times which cause the least or minimal interruption to normal business activities, ie during evenings or weekends.

The cost structure of an automated service is also typically lower than other types of cleaning services, particularly for lower numbers of records. This brings advantages for businesses that have had its core database cleaned and may have additional or on-going cleaning requirements for additional data. Instances of this could include businesses that:

- Have a very transient customer base
- Regularly acquire a significant number of new customers or,
- Intermittently use external sub sets of data (ie direct mailing lists)

6. Indicative Pricing

Different service providers offer different pricing models and it is important that the differences are made clear up front, prior to making any commitment to having data cleaned.

Costs are generally broken down into the three components:

- Set Up cost
- Per record/change processing
- Statement of Accuracy (SOA)

It is important to understand how processing charges are applied. Costs can be charged either on a per record basis, or on the number of addresses checked and the number of corrections made.

Some providers will charge a set up cost plus a per record charge, others will offer a fixed or minimum cost for data cleaning, particularly through a bureau service.

An example of this would be a poor quality database, requiring multiple changes (ie correction of street name spelling, adding city/suburb relationship and appending new post cost). This sort of job could have different costs applied:

No of Records	4500	Per record charge	Per record &per change
	Set Up Cost	\$350	\$350
	Processing cost	\$40/1000	\$40/1000
	Ave No of changes	-	3
	Total Cost	\$530	\$890

Associations like Chambers of Commerce, Industry bodies and advocate groups will often have formed partnerships or agreements with service providers to offer special member discounts, so its important to inquire with these types of organisations and look out for relevant articles in publications or magazines.

A Statement of Accuracy will typically be charged for and while the cost of this varies, it can be up to a couple of hundred dollars each. From time to time some providers will promote special offers for a period of time, waiving the cost of an SOA.

There are other providers who permanently offer SOA's as a free service through down loading a simple piece of software, so it pays to shop around when you are comparing costs of data cleansing.

Bureau Cleansing

For the reasons already mentioned, a bureau service, which offers a different type of service, would in most cases be more expensive than a web portal service within the approximate region of 20-45%.

Bureau cleaning typically is a fixed cost service with volume price breaks. Pricing will also vary if single or multiple files are being cleaned at the same time.

Web Portal/Automated

A web portal is designed to be self-service. If data can be standardised and provided in the correct format, then cost advantages can be gained.

Again prices will vary between providers, however typically for the provision of address correction, postcoding plus Statement of Accuracy, prices would start from \$30 per 1000 records, or fixed at \$ 295 - \$1000 depending on the provider selected.

Cost Estimates

Some provider's web sites will have cost estimate calculators, allowing businesses to receive an estimate for cleaning services, whether it be through a bureau or automated portal type of service.

7. Suggested Approach

For businesses sending data out for the first time, it can be a learning experience. However, having gone through the process and understood the various requirements, doing this again will be a far simpler process.

Implementing an effective data entry validation system is one way of keeping addresses clean. In the absence of this, the next best option would be to conduct regular cleansing of data.

Downloading Statement of Accuracy software is another simple and effective way for businesses to continue to monitor data quality levels. A Statement of Accuracy is valid for a period of 12 months, unless there are significant changes to the level of data quality within a database during this time. Regularly running a test SOA can signal when changes in data quality levels are occurring.

It is often possible to date stamp when new records were added or records were changed within a database. With this knowledge implementing a regular cleansing

programme becomes more manageable. Understanding how the data cleansing process works, the file format requirements and how cleansed data is returned does make it easier to move to a regular cleaning programme and an automated cleaning service may be the best option for this requirement.

Contributor Acknowledgment

NZ Post would like to acknowledge AddressWorks for their contribution to this article. More information can be found by visiting www.addressworks.co.nz