Data Cleansing and Maximizer

If no one person is responsible for the maintenance of a database, data cleansing is occasionally required. Undertaken with a suitable degree of forethought the results can be high and can help you comply with legislation such as the Data Protection Act for example.

The biggest pitfall in this process, is listening to the data cleansing vendors promises of what can be done with your data without giving sufficient thought to how the results of their work will be re-integrated into your database. Don't forget that Maximizer keeps lots of lovely history linked to names and addresses and in most cases you'll want this to stay linked to the appropriate cleaned data. On the rare occasion when you are prepared to start from scratch, accepting the results of data cleansing is relatively easy.

This working paper is intended to provide you with a good basis for considering how to approach the task, how to communicate your requirements to your cleansing vendor and then how to integrate the changes they supply.

1. Types of Cleansing

What exactly does data cleansing mean?

This is the first thing you must clearly identify. You need to clarify exactly what you want to have achieved by the end of the process. Often in database cleansing the administrator may just want to correct address details. However, data cleansing may also include the addition of new information and even new contacts.

There are a number of levels of cleansing which you may wish to undertake. Each of these different types requires a different level of effort by the cleansing vendor. It is also worth considering whether this is a one off exercise or the start of an on-going process.

In the follow sections we consider the different types of cleansing you may be considering.

1.1 Data Enhancement

With Data Enhancement you will be asking your cleansing vendor to add additional information to your database based on your existing data. This type of cleansing is often simply based on adding new information to your existing companies and contacts. In some cases your cleansing supplier will also offer to use a profiling system to select many new prospects for you based on a similar marketing profile in their research database using an ACORN or similar coding system.

This type of checking of names and addresses may be undertaken by a telemarketing team and could be a one off or on-going process.

1.1.1 One-Off

As a one off process you will need to take particular care to consider all the different items of information that may be changed as a result of the cleansing work. Your cleansing/telemarketing supplier will be doing a lot of work for you on each address and careful consideration is required for how every item changed will get back into Maximizer.

One off processes are usually done as a batch job and the cleansing vendor will probably require to load your key Maximizer data into their own systems to undertake the cleaning task.

1.1.2 On-going cleansing

If you are considering an on going cleansing or telemarketing process then Maximizer offers a unique possibility..... Put the cleansing/telemarketing company on Max Exchange and suggest that their staff work directly on your Maximizer data.

With the cleansing/telemarketing company working on a copy of your database, any changes can be synchronised to your office as frequently as required.

If using Max Exchange is not a possibility then you'll need to design a 'batch' process to allow you to send parts of the database and know subsequently which records in Maximizer have been updated and which have not.

1.2 Data Checking

With Data Checking the main aim is to validate and correct the data that is already in your database. This activity will often be undertaken by the cleansing vendor using software to automatically update your data. Examples of this type of cleansing include checking and adding postcodes.

1.2.1 Cleansing in situ

Some data checking programs may be able to make use of the Maximizer ODBC driver to work directly on address data in your database. Check carefully that products used for this kind of update are able to match the output field formats and lengths required by Maximizer. You may want the ask the cleansing vendor if the product provides any kind of trace or log of the changes made in case you later want to see what was changed.

1.2.2 Cleansing Off-line

Most cleansing vendors will want to extract the address information from your database for cleansing. It will generally be very easy to extract the data for them. Getting the changes back in to Maximizer, however, can be quite another matter without the right tools and a little planning. This aspect of the cleansing project is often one that is given insufficient consideration by cleansing vendors and database owners at the start of the project.

1.3 Data Tidy-up

Tidying up data is usually only aimed at improving the format of the data. For example, you may want to re-format address data that has been entered or imported all in capital letters. This type of data improvement can also involve moving data (such as town names) to the correct database fields in order to improve searching.

2. Output

When your cleansing vendor has worked on your data you will have two main options as to what you'll do with the data when returned to you.

2.1 New Database

You may be prepared to scrap your old Maximizer database and simply start again. In some cases this may be a perfectly good approach, which is quite simple to undertake. More than likely, this process will only require the standard features of Maximizer to complete. If you adopt this approach you will make your task much easier by providing the cleansing vendor with the Maximizer field list and Maximum field widths as a part of the output specification.

2.2 Updates to the original database

In most cases you will have lots of valuable history attached to the addresses and contacts in your Maximizer database. In order to integrate the results of the cleansers work on your data back into your original, we need to provide a mechanism for getting the changes back into Maximizer and match them up with the right history data. Before the data is sent to the

cleanser vendor it will need to be extracted with the identification reference numbers from Maximizer. The identification numbers are the unique reference numbers Maximizer gives to each company in your database. If retained by the cleanser these identification numbers will provide a link that will make the updating of the original records possible.

2.2.1 Method

There are several possible methods for getting your updates back into Maximizer. One of the easiest is to use CABC's Max Feed product. If you have a Maximizer integrators toolkit it may also possible to use Maximizer's proprietary MTI file format.

2.3 Format

When handling your exported data be especially careful if you open the files in Microsoft Excel. If you do this to check and review the files, be especially careful not to save the file at the end of your review. In general avoid using Excel if possible and in preference use Microsoft Access to open the data as it is very easy to inadvertently change a phone number for example from 01635 570970 to 1.6E+09 !!!!

3. Planning Issues

In this section we consider some of the issues that need addressing in order to successfully complete a cleansing project. Many of these items will be will be very dependant on how you use your database and you will need to consider carefully exactly what you expect your cleansing vendor to do with your data. Most importantly these items are a checklist to help insure you've thought about how the results of the cleansing work will be re-incorporated.

3.1 records

The first thing is to be sure about which records from your database are to be processed. Consider whether this exercise will be just be for address updates or will it include company and contact names?

3.1.1 Individuals and/or Contacts

Records that record the details of people in Maximizer are made up of two distinct type of records, Contacts and Individuals. Will you send both of these? They have different fields and may require different processing.

If you are cleansing contacts, then your existing database will often start with multiple contacts for each company. Will you send all these to the vendor? or just one or two? If you don't want to send them all how will you select the ones that are to be processed?

3.1.2 Addresses only

If you are planning to clean address information then you may need to give consideration to your current use of Alternate/Mailing addresses in Maximizer (these are the ones you can add on the Mailing Address tab of a Maximizer Address book record). If you make considerable use of these addresses then you will need to be especially careful when considering how to incorporate changes back into your database.

3.2 New info

Will the cleansing Vendor be adding new information to your data?

3.2.1 New Contacts

If you are intending to generate new contacts for existing companies in your database, these new contacts will need to refer to the correct Company Client ID in order for you to be able to import them.

3.2.2 Changed contacts

If your cleansing process will be revising the details of contacts, the Company Identification may not be sufficient to enable you to update the changes. Internally, Maximizer gives contacts a sub-reference called the contact number. This is a unique reference to a contact at a company. Knowing this number when the data is returned to you would allow for

example, the name of the Managing Director of a company to be changed. If you don't have a key like this, the original contact name will need to be used for the matching.

3.3 Deletions

Always ask your cleansing vendor to provide the list of records, which they recommend you delete.

Unless you are building a completely new database then you will need a list of companies and contacts that are to be deleted, along with their original IDentification numbers. These will allow you to add a field to these records in Maximizer, marking them for removal. After you've merged back the new deletion information, you can search for these records and then archive/delete them.

3.4 Merges

If your data cleanser will be de-duplicating your data you will need to give consideration to how you will update Maximizer. If you need to merge the records in order to keep notes and documents, you will need to provide details of the old record IDs and/or contact names that should merge into which IDs. If there are many of these it could be a big task for you as they will need to be entered onto the system by hand.

If you are prepared to loose history on the records that are being merged into other records, simply ask the cleanser to present any merges as the resulting new records and deletions.

3.5 Overseas addresses

If your database contains overseas addresses you may need to exclude these from your list for cleansing, or make special provision for checking them.

3.6 Duration of process

If there is a long time between sending data to the cleanser and its return, some data will not necessarily update correctly when returned as you may have altered the original records in the mean time.

4. Guidelines

These documents are intended to help you with instructing a data cleanser.

4.1 Fields

This is the reference list of Maximizer field details which is provided in Annex A. You'll need to give a copy of this to the data cleanser.

4.2 Instructions to the cleanser

This is a data sheet (Annex B) explaining to the cleanser the formatting required and the importance of retaining your key field information.

4.3 Export Definition

This is the start of the definition of your export that you will give the data cleanser. (Annex C)

You may also want to agree in advance the format of the data they will return data in a similar way.

5. Useful Products from CABC

The following CABC products can be useful with data cleansing. Please contact CABC for more details.

5.1 Max Feed

Max Feed will help you update your database when your cleansing company returns the tidied data files.

Max Feed can also help de-duplicate your data in an automated fashion.

5.2 Max Tidy

Max Tidy is only available on a service basis. This product will read through all of your data improving the formatting a (e.g. character case) and positioning of address fields (e.g. moving address components back to the right fields).

5.3 Max Splitter

This product will assist with the majority of conversion of the Individual client records into separate companies and contacts. This is very useful if in the past data was poorly imported.

5.4 Max Contact Tidy

This product can remove duplicates of the same contact name at the same company throughout you database in an automated fashion. This is very useful if in the past data was poorly imported.

5.5 Max Postcode

This product can help you with the correct entry of new addresses in Maximizer using the Royal Mail PAF files.

Prepared By Ian C. Wallace
Principle Consultant
CABC Ltd
6 West Mills Yard
Kennet Road
Newbury
Berks
RG14 5LP

Tel: 01635 570970
Fax: 01635 570971

ANNEX A Data Field Definitions

Contact Details

Field Contents	Max Length	Comment	Field Ref No.
Identification	23	Maximizer Internal Reference	1000
Contact Reference	100	Original Full Name or a Contact Number	1001
Mr/Mrs	39		1
First Name	39		2
Initial	1		3
Last Name	59		4
Position	59		5
Salutation	39	(caution coding scheme!)	6
Email address	118	Personal	7
Web Site	118		8
Phone 1	21	Often Min switch board can be DDI	9
Phone 1 Description	5	Phone 1 description E.g. 'Main'	10
Phone 2	21	Often main/personal Fax no	11
Phone 2 Description	5	Typically 'Fax'	12
Phone 3	21	Often DDI number	13
Phone 3 Description	5		14
Phone 4	21	Often Cell phone number	15
Phone 4 Description	5		16
Other UDFs	Variable	May be, free text, Dates, Numbers or lists	100 to 300

Company Details

Field Contents	Max Length	Comment	Field No
Identification	23	Maximizer Internal Reference	1000
Company Name	59		40
Email Address	118	General	41
Web address	118		42
Phone 1	21	Main Switchboard No	43
Phone 1 Description	5	Phone 1 description E.g. 'Main'	44
Phone 2	21	Often used for Main Fax No	45
Phone 2 Description	5	Typically 'Fax'	46
Phone 3	21		47
Phone 3 Description	5		48
Phone 4	21		49
Phone 4 Description	5		50
Other UDFs	Variable	May be, free text, Dates, Numbers or lists	400 to 500

Address Details

Field Contents	Max Length	Comment	Field No
Department	39	Often empty. Do not use for street/property info	30
Division	39	Often empty. Do not use for street/property info	31
Address Line 1	39	Property & Street	32
Address Line 2	39	Locality	33
Town	39		34
County	39		35
Postcode	19		36
Country	39		37

Annex B - Notes for data cleansers working with Maximizer Data

This document assumes you are working with data extracted from Maximizer for the purpose of cleansing. The exact nature of the cleansing will be agreed with you separately. This document is concerned with defining which data is being provided and ensuring that you are able to return it to us in a format that can be used to update our database.

You will be provided with an extract of data which will include the fields listed on the Extract Sheet attached. Please read this in conjunction with the field list.

When processing our data it is vital to us that you maintain the Identification field information with every address record. If you will be changing contact names then please maintain the original contact name and/or the contact ID that has been provided in our extract.

Format of returned data

We would prefer data to be returned in a tab delimited file format without string delimiters.

The updated data should be returned using the field format provided (i.e. Company, Department, Division, 2 address lines, town, county, postcode). Please do not load the property or street details into the Department or Division fields.

The Maximum field lengths specified must also be observed for each field.

For each data set please return the data in two files:

New and updated records

For updated records the ID and original contact name (Full Name) or contact ID must also be returned with each company/contact

E.g. updated record

Field	Supplied	Returned
ID	011213044071788442303C	011213044071788442303C
Company	CABC	CABC Ltd
Original Full Name	I wallace	I wallace
Contact No (if provided)	2	2
Update First Name		Ian
Updated Last Name		Wallace
Position		Director
Department		
Division		Maximizer CRM Systems
Address Line 1	1 West Mills Yard	6 West Mills Yard
Address Line 2		Kennet Road
Town	Newbury	Newbury
County		Berks
Post Code		RG14 5LP

New records will obviously not have an original contact name or number,,however if you are providing us with new contacts at an existing address please ensure the contacts have the correct company Identification attached.

Deletions

If as a result of the data cleansing you recommend we delete records, these must be listed in a separate file and not simply removed from the returned data set.

E.g. Deletions File contents

Field	
ID	0112130440717884423O3C
Original Full Name	I Wallace
Contact No (if provided)	2

For companies you need only return Identification numbers.

Merged Records

If you are de-duplicating the data we may require you to provide details of suggested merges in a similar format i.e. which IDs should merge with which. We may be happy to accept these simply as the resulting new/deletions and will have advised you on this.

Annex C - Export Data Definition

Our Database extract contains the following fields in a tab delimited format.

Please refer to the standard field list for Maximum lengths and other comments using the field referenced.

Field	Field Ref No	Comment
Identification	1000	
Company Name	40	
…etc.		