The intricacies of deduplication

Executive Summary

Since 1996 CCR have been working with clients to provide the most effective de-duplication routines possible. This is an evolving process which varies for each piece of work undertaken. In terms of data quality and data cleaning we all have our own goals and requirements and an integral step to meet these requirements is the accurate deduplication of data.

This document aims to explain the intricacies of deduplication:

- Why de-duplicate data
- De-duplication levels
- Master record selection
- Managing Duplicates

De-duplication - The reasons why

De-duplication and matching should be at the heart of any data oriented system, whether it be a data warehouse used for business intelligence or a CRM system (Customer Relationship Management) such as Sales*force* or Microsoft Dynamics. For example in the world of consumer database marketing it helps identify and group members of the same family which is the first step to data cleaning and creating a view of the behaviour at household, family and individual levels.

The benefits of accurate de-duplication are:

- The ability to correctly identify returning customers
- Identify and group customers that exist in the same dwelling
- Reduce wasted direct mail budget
- Merging of multiple data sources
- Create a single customer view to aid reporting, analytics and business intelligence
- Reduce brand degradation
- Accurately calculate ROI for marketing campaigns
- Accurate data cleaning and data
- Net names agreement

CCR Data Limited, Unit 4, Minton Distribution Park, London Road, Amesbury, Wiltshire, SP4 7RT
☎ 01980 592999 Tel ☎ 01980 592929 Fax ✉ info@ccr.co.uk www.ccr.co.uk

The intricacies of deduplication

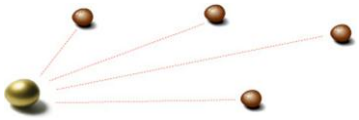
De-duplication levels

There are a number of differing levels which can be applied to ensure accurate de-duplication. The decision regarding which levels to use is dependent on a number of factors such as the types of source data files i.e. B2C or B2B and the campaign type i.e. customer contact or prospect mailing.

Individual – To be identified as a duplicate the surname, forename or initials and address details must be deemed to be the same

Mr Andrew Smith, 1 The Road, London, AB1 2CD

Mr A Smith, 1 The Rd, London



Mr Andrew Smith, 1 The Road, London, AB1 2CD

Mrs Barbara Smith, 1 The Rd, London

Family – To be identified as a duplicate the surname and address details must be deemed to be the same

Household – To be identified as a duplicate the address details must be deemed to be the same

Mr Andrew Smith, 1 The Road, London, AB1 2CD

Miss P Jones, 1 The Rd, London

CCR Data Ltd, Unit 4 Minton Distribution Park, Amesbury, SP4 7RT

CCR, 4 Minton Distribution Park, London Road, SP4

Business – To be identified as a duplicate the company name and address details must be deemed the same

Contact – To be identified as a duplicate the surname, forename or initials, company name and address details must be deemed the same

Mr Andrew Smith, CCR Data Ltd, Unit 4, Minton Distribution Park, SP4 7RT

Mr A Smith CCR, 4 Minton Distribution Park, London Road, SP4

CCR Data Limited, Unit 4, Minton Distribution Park, London Road, Amesbury, Wiltshire, SP4 7RT
☎ 01980 592999 Tel ☎ 01980 592929 Fax ✉ info@ccr.co.uk www.ccr.co.uk

The intricacies of deduplication

Master Record Selection

In order to create a single file and complete the deduplication process a master record must be selected. If you have 3 records each defined as the same, which record to you keep? This process is different for each project and more than one rule can be used, examples rules are shown below.

Random master selection
A record is taken at random from the matching set.

Record age
Master record is selected on the basis of addition date to the database.

Records history
Number/Value of purchases.

Campaign history
A record is selected on the basis of the number of campaign responses.

Field based
The number and depth of field population is used to determine the most complete record which is then defined as the master record.

Field length
Selection of master records is completed using fields such as name where the longer of the two is determined to be the most complete. I.e. E Spicer or Edward Spicer.

Record source
Records are selected in order of priority to be the master record on the basis of which source file they originate from.

Managing Duplicates

Once duplicates have been identified at the required level and master records selected duplicates need to be managed. This can be done in many ways and again will vary dependant on a client's requirements. Do you?

- Flag & exclude duplicates from future campaigns
- Delete all duplicated from the system
- Merge duplicate records

The most effective method is to merge duplicates. This not only reduces wasted direct marketing spend but enables the following:

Accurate, cost effective data cleaning
Create and maintain a single customer view
Retain valuable contact history
Keep field variables i.e. email

Managing duplicates is an on-going process and should be assessed at every juncture to ensure that the knock on effects caused by duplicate records can be limited.

To assess your data and gain an accurate picture of your duplicates contact CCR for a free data audit.

CCR Data Limited, Unit 4, Minton Distribution Park, London Road, Amesbury, Wiltshire, SP4 7RT
☎ 01980 592999 Tel ☎ 01980 592929 Fax ✉ info@ccr.co.uk www.ccr.co.uk