

Data Quality: What do you consider a duplicate record?

By Tyler Anderson IT Director, immedia

At first glance snowflakes often seem identical, and yet we know each is unique in structure. The same can be said for how people view duplicates in data files and how they decide these dupes should be compared and removed. When reviewing data files for duplicates what one person sees as similar others may see as totally different. It's all about perspective.



Reviewing files for duplicates and actually purging them can present many hazards. For example, look at the way MS Excel stores records in cells over columns and rows. If a column is improperly indexed or sorted the other columns won't change with respect to that index resulting in a corrupted database. Checking for duplicates in Excel can be time prohibitive and expensive. MS Access has a query system for duplicate detections that can be processed on any number of fields, but the data must be **identical**.

The key is to be able to find data that matches exactly or very closely — at least closely enough for the person reviewing the data to positively say that two records are duplicates.

Let's look at some actual scenarios that have presented themselves to me over the years. Say you have a database for a doctor that contains records for each of his patients. It's safe to say that there will be a record for each family member since most nuclear families go to the same family doctor. We could then conceivably have a record for Jim Smith, Carol Smith, Cary Smith and Clark Smith all at the same address. If we were to do a mailing, how would we want to dupe detect a file of this type? Your first inclination might be to say that they are all duplicates, but others may feel that they are not. It all depends on how you look at the data, what you are mailing to them, and what you hope to achieve. If this doctor is going to mail out new forms for each patient to sign he would definitely want each family member to get one. However, if he wanted to do a mailing letting his patients know his office was going to move, a single letter to the entire family would probably be sufficient.

What if we added someone else to this household with a last name such as Jimmy Cobb? As we all know, family dynamics have changed considerably over the past 20-30 years. Household members, even married couples, don't always have the same last name. These types of family dynamics force us to treat data very carefully and review who the audience is and what they are receiving. Often times we receive requests for dupe detects to be done "one to an address", the most destructive to a file. Later the client finds the results are not



what they expected. For example, the Janitor received an invitation to the Spring Fundraising Gala when it should have gone to the President/CEO.

This leads me to my next point: criteria used within the process. I can't stress the importance of criteria enough. It is an extremely dynamic feature that can be "tweaked" and manipulated to no end. If you only check for duplicates based on ZIP Code and address then you won't have any control over which person in the house or business gets one. One to an address wipes out everything at that address except **ONE** record. By simply adding last name to the mix you essentially add one more criteria and effectively reduce the number of dupes that get expunged from the file. This may or may not be a good thing. If you look at my previous example of the Smiths and Jimmy Cobb, we may have wanted to send him his own letter about the doctor moving his office. Maybe he's a relative, but maybe he's renting the apartment above the garage. Who knows? It may be safer in this instance to mail a letter to the Smiths and one to Jimmy Cobb. There may be other data present in the file that could provide additional clues such as age. Age would tell us if he is old enough to read his own letter or if he should be included with the rest of the clan.

Another factor to consider in this scenario is this: if duping on one to an address and last name, then which of the Smiths will get one? The answer is: the first one in the data file. This may be Cary Smith who is four years old. Maybe it would be better to address the mailing panel to read "Jim, Carol, Clark & Cary Smith" or "The Smiths". This would certainly let the Smiths know that everyone in their nuclear family should review the information. But what do we do with the other three records in the file? They simply get removed from the data file and not mailed to. The software is able to combine the first names through a process called *Group Posting* whereby it recognizes exact same last names and combines the first names into a single record. This is also known as *Householding*.

Criteria can be based on anything from ZIP codes, addresses, titles and names, to account numbers, dates, social security numbers, branch locations, sales amounts or even revenues. It doesn't just have to look at mailing address information for the software to consider it a dupe. Within the specified criteria of fields, you can select a level of intensity to base the checks on. For example you may want an exact match on ZIP

Features of Duplicate Detection Software

- Offers purging of unwanted duplicate records, to eliminate extraneous or out-of-date customer data
- Matches customer data within a single source or across diverse sources, even when wide variations, such as different spellings are involved; consolidates customer data into one complete record, based on your specifications
- Increases processing efficiency and decreases costs with "one-pass" processing and n-per firm capabilities
- Provides further identification and fielding of address, name, and firm name to standardize data internally prior to the matching process
- Provides reports based on individual records, input mailing lists, and output mailing lists
- Uses suppression lists to eliminate identified records based on specific business rules or remove people who have requested to be removed from the mailing file
- Processes up to 2000 different input lists with priority handling for each



and address, but only a tight match on last name. These matching options can even be translated into a percentage. Duplicate detection software simply uses a set of algorithms to make comparisons in a mathematical sense. If it feels that a comparison is close enough mathematically then it considers it a dupe.

Extremely sophisticated software such as our Match/Consolidate program can even check for transposed letters, firm abbreviations such as IBM vs. International Business Machines and name variations such as Jim versus James.

The software also matches initials to the name such as J. F. Smith versus James Francis Smith and matches on hyphenated last names such as Mary Anderson-Olson versus Mary Anderson.

Whatever the outcome is, your final goal should be to gain a cleaner list that meets your needs, gets the information to your clients the most effective and efficient way, and saves postage and printing costs.