



# DUPED AGAIN

Are duplicate records a blow to your bottom line?

By Ken Brashears

Is the problem of poor data quality an ongoing challenge for your organization? A recent TDWI study reported that organizations lose more than \$611 billion each year due to bad data. Even more shocking, 42% of organizations have made no effort to monitor the quality of their data, according to a study by The Information Difference.

One major root cause of poor data quality is duplicate records. Duplicate records — at any level or amount — can adversely affect the performance of your database, as they can tarnish vital customer contact information. Conflicting data prevents your organization from gaining a single, accurate view of your customer, creating the risk of making poor business decisions. Duplicate records also lead to increased manual labor, lackluster customer response-driven initiatives and wasted postage and production costs.

But there is a way you can turn chaotic data into actionable information: by identifying and eliminating duplicate records.

### The Problem of Finding Non-exact Matches

Deduplication of data through the use of merge/purge solutions is one of the critical components in the data cleansing process. The biggest roadblock to identifying duplicates lies in detecting non-exact, matching duplicate records, data that appears to be multiple sets of unique information but are actually duplicate records. Most data contains these non-exact, matching duplicate records, which are very difficult and tricky to identify.

For example, a “Beth Smith” at “United Data Machines” can be recorded in the same or different database as “Smith, Elizabeth” at “UDM.” In reality, Beth Smith and Elizabeth Smith are the same person, but your organization might identify the data as two different contacts.

To identify these hard-to-spot, non-exact, matching duplicates, the most advanced merge/purge applications utilize “fuzzy matching” algorithms. Fuzzy matching is a mathematical process that determines the similarities between data sets, information and facts where the outcome is neither true nor false, nor 100% certain — hence the word “fuzzy.” The process compares any data type of any length and from any place in a field to find non-exact matches.

### Types of Fuzzy Matching Algorithms

There are several different algorithms that can be implemented as part of the deduplication process:

- **Phonetic matching** — Utilizes the phonetic algorithm to detect “alike-sounding” relationships between words. Phonetic matching allows you to perform approximate searches, instead of just exact matches, thus enabling your organization to find variations of a name.
- **N-gram- or q-gram-based** — The linear n-gram- or q-gram-based algorithm models are primarily used in statistical, natural language processing. An n-gram is a subsequence of “n” items from a given sequence, which can be phonemes, syllables, letters, words or base pairs, as defined by Wikipedia.
- **Jaro-Winkler** — The Jaro-Winkler distance is a measure of similarity between two strings. It is mainly used in the area of record linkage for duplicate detection.

### Finding a Solution that Pushes the Envelope

Most mail management software packages integrate a deduplication function as part of the address hygiene process to not only eliminate bad addresses from the mailing but find and eliminate duplicate records. The combined process ultimately saves the mailer maximum dollars on both printing and postage.

For mailers needing a more potent merge/purge deduping process, they can look for a solution that incorporates the ability to identify the most difficult-to-detect duplicate records via fuzzy matching algorithms while also performing USPS CASS-certified address verification routines.

### Three Steps to Customizing Your Deduplication Efforts

Once you’ve identified a software vendor to implement deduplication technology into your operations, you’re ready for the next step: customizing your entire data quality strategy.

Here is a three-point checklist to follow when customizing your deduplication process:

1. **Standardize your data.** Accurate, consistent, high-quality data is the foundation of a database, so the need to implement sound data standardization processes to ensure the integrity and validity of your data is critical. Your organization cannot build a data warehouse, integrate or migrate data or get a complete view of your customers without first standardizing data. There are data quality solutions in the market that can standardize, verify and validate the contact information in your database, some even in real-time at point-of-entry or in batch mode.
2. **Determine your business needs.** Based on your specific business needs, you can compare your records at once, which is ideal for batch merge/purge suppressing existing data; or you can compare each record as they come in and against a database of already processed records, which is ideal for real-time data entry. Another method called “hybrid deduping” allows your organization to customize how records are processed and stored.
3. **Monitor your data.** Databases that contain contact data are never static because information is constantly changing as customers move or change companies. And new or possibly inconsistent data is coming in all the time from call centers, web forms and data entry by various departments. That is why it is important to add a monitoring package into your operations; this will provide real-time data monitoring in an automated process to immediately recognize and correct issues before the quality of data declines. This approach also helps organizations enforce data governance and compliance measures.

The true value of your database is determined by one fundamental component: the quality of your data. Without data that is reliable, accurate and updated, your organization’s direct marketing and customer communication efforts can never be fully optimized. Detecting the most difficult-to-detect duplicate records and merging/purging them to achieve a unified view of your customer’s buying habits and brand interactions will help you reduce costs and labor and better target and increase relevance for all of your marketing campaigns. ■

Ken Brashears is a product manager at Melissa Data, a provider of data quality and address management solutions. To learn more about how to implement deduplication technology into your operations, go to [www.melissadata.com](http://www.melissadata.com) to download the white paper, “Are Duplicate Records Eroding Your Bottom Line?” or call 800-635-4772 to get a free trial of MatchUp, Melissa Data’s deduplication software.