

Evaluating Merge/Purge Systems: Part Two

*By Jim Wheaton and Cynthia Baughan Wheaton
Principals, Wheaton Group*

Original version of an article that appeared in the August 1987 issue of "Direct Magazine"

[Note: Despite dramatic increases in raw computing power and a proliferation of end-user software tools since the publication of this series of six articles, virtually all of the content remains highly relevant. The occasional obsolete point is highlighted.]

Statement of Purpose

In a series of six articles, we explain a number of the key concepts that mailers should understand about merge/purge, as well as reviewing (in the first article) a methodology that could be helpful in evaluating the effectiveness of either present or prospective merge/purge systems. While our comments are primarily addressed to mailers, merge/purge vendors can benefit by measuring themselves against the criteria that we have identified as important.

Our objective is to describe new and specific tools that can be used to evaluate and improve the performance of the merge/purge process. Through commentary and examples, we will attempt to translate into layman's terms the technical jargon that baffles many mailers. In the process, practical applications should become apparent.

This month's article, Part Two, focuses on match codes, list selection, system flexibility, and the importance of fine-tuning merge/purge parameters.

Match Codes

We found, surprisingly, that simple match codes are still used by some in the industry. [Subsequent note: This is no longer true.] The problem with match codes is that they analyze only a portion of the name and address information available in a record. A match code frequently is comprised of a fixed number of consonants from the first and last name of the individual, a fixed number of digits from the street numeric, and the entire ZIP Code.

The following are three records that demonstrate the problems inherent in simple match code logic. They are based on the logic currently being used by one of the systems evaluated in Phase 1 of our study:

Record	Match Code
1) <u>Bruce</u> <u>Rose</u> <u>min</u> <u>1143</u> <u>Riverside</u> Avenue Fall River, MA <u>02726</u>	BRSMN143RVR02726
2) <u>Bruce</u> <u>Roz</u> <u>emin</u> <u>1143</u> <u>Riberside</u> Avenue Fall River, MA <u>02726</u>	BRZMN143RBR02726
3) <u>Bruce</u> <u>Rise</u> <u>semanski</u> <u>2143</u> <u>Riverside</u> Avenue Fall River, MA <u>02726</u>	BRSMN143RVR02726

Note that the sections analyzed by the match code have been underlined. In this case, let us assume that two records will be considered duplicates when 15 of the 16 match code characters are identical. We will track the process in which the software will determine whether record #2 and/or record #3 are duplicates of record #1:

- In record #1, the corresponding match code consists of the 16 characters that have been underlined.
- In record #2, we have a slightly different last name and a slightly different street address. These small differences, however, are probably due to input errors. Notice that the match code is different by two characters from that of record number 1.
- In record #3, we have totally different first and last names. In fact, they are not even close. We also have a different house number.

Intuitively, records number 1 and 2 are obvious matches. Number 3, on the other hand, is apparently a second individual. But, let's track how the match code would analyze them:

- Number 2's match code differs from number 1 by two characters. It is therefore not considered to be the same individual as number 1. Both records would be mailed.
- Number 3's match code, however, is identical to number 1. It is considered a match and would not be mailed.

More advanced merge/purge software uses every element of a record to determine whether a match exists. Although vendors guard closely the details of their systems, it is clear that a number have become quite sophisticated, some even comparing phonetics rather than literal spellings. Complex decision trees and mathematical scoring formulas have been developed by the best in the industry, in the never-ending attempt to simulate human judgment.

List Selection

List selection plays a major role in determining the overall difficulty of duplicate problems. When poorly maintained lists result in complex duplicate problems, state-of-the-art software can pay large dividends.

- Subscription lists, including magazines and periodicals, tend to be more accurate and up-to-date than catalog lists. This is because subscribers, who have paid up front for ongoing service, have a vested interest in reporting input errors and address changes.
- Compiled lists, on the other hand, can range from the best to the worst maintained, depending on the compiler's source of names and update schedule.
- Consumer lists can vary in accuracy depending on the tightness or looseness of matching rules applied to them over time.
- Business lists tend to be less accurate than consumer lists, due to frequent employee turnover and job title changes, as well as unique complications that we will discuss in a later article in this series.
- In addition, phonetic problems have been on the increase as telemarketing has assumed major importance for many direct response marketers.

The Importance of Fine-Tuning Merge/Purge Parameters

Parameters are the “rules” within a merge/purge system that define exactly the way in which certain decisions will be made during processing. They can be turned “on” or “off,” as well as adjusted, without programming, depending on the needs of the mailer.

Some vendors require that input forms be filled out prior to a merge/purge, which forces the mailer to think through the decisions that must be made during processing. This helps both parties to focus on the key issues, and provides documentation for future reference.

Unfortunately, not everyone thinks through the options clearly. Often, decisions are made “on the run,” without understanding the implications. For example, the following parameter was recently used by experienced management at an established catalog firm: if the first initial, the last name, and the ZIP Code match, a duplicate situation is assumed to exist as follows:

Record #1	Record #2
Mr. <u>Elizabeth Cooper</u> Apt. 527 114 N. Mt. Carmel Way Wichita, KS <u>67203</u>	<u>E.G. Cooper</u> 1649 N. Charles Wichita, KS <u>67203</u>

- It is extremely unlikely that Elizabeth Cooper is in any way related to E.G. Cooper.
- Under this parameter, however, Elizabeth Cooper’s record was identified as a two-time multi-buyer, and E.G. Cooper’s record was eliminated as a duplicate.
- In fact, if a third record, with Edward Cooper living anywhere in ZIP Code 67203, had been included in the gross input, only Elizabeth Cooper’s record would have been retained. Besides losing two perfectly good, unique prospect names, the mailer would have paid for what was thought to be a valuable three-time multi-buyer.

This one parameter resulted in a large number of similar incorrect matches, with several unfortunate ramifications:

- Multi-buyer counts, which are very important to the mailer, were highly inflated. Significant numbers of three and four-time buyers were seen for the first time.
- Therefore, the quality of the rented names was thought to be unusually high, and the client seized this “opportunity” by adding an extra multi-buyer mailing to the promotional schedule.
- A large number of unique names were improperly eliminated as duplicates, resulting in an unexpectedly low single-buyer count. As a result, many marginal lists were included at the last minute to meet planned mail quantities.

The results were disastrous:

- The response rate for names identified as multi-buyers was very disappointing.
- The incremental mailing to multitis was equally poor, and compounded the problem.
- The marginal single-buyer names that were added performed very poorly, and should not have been mailed.
- Overall, outside list response was depressed by as much as 10 percent, and the cost per new customer increased accordingly.

The important question is, “Could this happen to you?” It should not if good communication has been established with a vendor who assists in carefully examining any parameter changes.

Flexibility

Flexibility can be another important measure of a vendor’s ability to handle a client’s business. Stated another way, the ability of the vendor to meet the client’s needs is directly proportional to the number of parameters that can be adjusted or changed.

- Marketing enhancements can result from an understanding of the options offered by a flexible system. Assume, for example, that an offer has been going to one individual within each household. The product is widely used and in the name selection process one gender has never been specified over another. Through testing, it is discovered that women respond better than men, all other factors being equal. Parameter flexibility would allow the selection of females over both males and “unknowns” in duplicate record situations.
- Parameters can be “tinkered with” over time as the mailing program matures.
- Some parameter options may be worth testing in the mail. Others may warrant adjustment based upon careful quantitative analysis rather than in-the-mail testing.

One example of a situation requiring a parameter decision is when two records with identical names and ZIPs differ because one has a street address and the other has a post office box.

Consider whether the following records are duplicates:

Record #1	Record #2
John Samuelson 122 Clover Street Gadsden, AL 35901	John Samuelson P.O. Box 4337 Gadsden, AL 35901

Gadsden, Alabama is a small town. Therefore, there is probably one John Samuelson.

But, what if these two John Samuelson records were located in New York City? Would they still be likely duplicates? Probably not.

A flexible vendor could set the parameter such that it would call these records duplicates if they exist in rural ZIPs, but not if they exist in urban or suburban ZIPs.

There really is no right or wrong answer. But this type of issue may make a difference – given the offer, its positioning, and the lists that have worked best. It depends on specific needs and individual priorities.

Jim Wheaton and Cynthia Baughan Wheaton are Principals at Wheaton Group, and can be reached at 919-969-8859 or jim.wheaton@wheatongroup.com. The firm specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data warehouses and marts. Jim is also a Co-Founder of Data University www.datauniversity.org.