

## **The Meaning of Merge/Purge**

*By Jim Wheaton  
Principal, Wheaton Group*

*Original version of an article that appeared in the January 1990 issue of "Direct Marketing"*

The power and flexibility of modern Merge/Purge software offers many capabilities that were unavailable with yesterday's primitive software. The goal of this article is to help mailers take full advantage of these capabilities.

A popular misconception is that a Merge/Purge is nothing more than name and address unduplication. In reality, unduplication is just one of five equally important steps that comprise a Merge/Purge:

- 1) Conversion places every name and address from every incoming list into a standard, "internal" format. This is critical for the Unduplication step to work properly. At the same time, the names and addresses are enhanced (e.g., ZIP Code correction, National Change of Address, etc). Optionally, geographic-level demographic data can be applied.
- 2) Unduplication identifies duplicates among names and addresses. Optionally, individual-level demographic data can be applied (i.e., exact age, length of residence, etc.).
- 3) Split and Key divides the "cleaned output" from Unduplication into appropriate panels (i.e., "Label Sets" or "Strings"). In addition, every name and address is assigned an appropriate Key Code (i.e., Source Code or Mail Key).
- 4) Presort organizes each panel out of Split and Key into Carrier-Route qualified, 5-Digit qualified and/or Unqualified ("Residual") groups, to take advantage of postal discounts.
- 5) Print/Reformat translates the service bureau internal name and address format into a format that can be used by the lettershop.

### **Definition of A Duplicate – Key To A Successful Merge/Purge**

The first thing to do when setting up a Merge/Purge is to arrive at an appropriate definition of a duplicate. This is more difficult than it sounds, because there is no such thing as a definite duplicate. One of the biggest fallacies in the Merge/Purge business is that names and addresses can be divided into "definite duplicates" and "definite non-duplicates." The following two names and addresses, for example, are not duplicates:

<b>Name and Address #1</b>	<b>Name and Address #2</b>
James Wheaton	James Wheaton
151 Thurton Drive	151 Thurton Drive
New Canaan, CT 06840	New Canaan, CT 06840

These two names and addresses actually represent two different people: my father and myself, at the address of the house I grew up in. Even though I was named after my father, mail frequently arrived at our home without a suffix; that is, with "Junior" and "the Third" deleted. This often caused confusion.

The point is, short of telephoning every individual who is to be mailed, there is no way to be 100 percent correct when it comes to defining duplicates. Modern Merge/Purge software, therefore, works off percentages. Instead of "definite duplicates" and "definite non-duplicates," direct response advertisers are presented with a spectrum of statistical probability:

- Each circumstance – every different business, every different mailing – requires its own definition of a duplicate. The direct response advertiser, with the assistance of his/her service bureau, must carefully choose from a range of duplicate rules.
- Each rule is a filter that catches names and addresses with a specific statistical probability of being duplicates. At one extreme, there will be a rule that catches only exact name and address matches. This increases the likelihood that duplicates will be mailed. At the other extreme, there will be rules that allow significant discrepancies in both the name and the address, resulting in a smaller number of names to mail.

In other words, modern Merge/Purge technology allows you to dial into the spectrum of statistical probability that is appropriate for your specific circumstances. Mailers must know their needs, and must work carefully with their service bureau's technical people to understand the implications of certain decision rules for that particular mailing.

### **Levels Of Unduplication**

Having arrived at an appropriate definition of a duplicate, the next key issue is the "level" of unduplication for the mailing (i.e., Company, Location, Family or Individual Name). The level of unduplication, in turn, will depend on whether the mailing is to be directed to consumer names, business names, or to both.

Consumer names are found at household locations and business names at business locations (as identified by a company name). Generally, each type requires a different level of unduplication.

For consumer names, modern Merge/Purge software allows three different levels of unduplication: Location, Family and Individual Name. Consider, for example, the following three names and addresses:

Name and Address #1	Name and Address #2	Name and Address #3
James Schneider	Debra Schneider	Oscar Jamieson
23 Adams Lane	23 Adams Lane	23 Adams Lane
Avon, CT 06001	Avon, CT 06001	Avon, CT 06001

- With unduplication at the Location Level, all three records are duplicates, because all three live at the same address. Only one will be mailed.
- At the Family Level, only James Schneider and Debra Schneider are duplicates, because Oscar Jamieson may be nothing more than a friend or a tenant. Therefore, only one Schneider record will be selected, along with Oscar Jamieson.
- And, at the Individual Name Level, none of these people are duplicates; obviously, because each is a different individual. All three will be selected.

For business names, modern software also allows three different levels of unduplication: Company, Location and Individual Name. This is apparent in the following example:

Name and Address #1	Name and Address #2	Name and Address #3
James Schneider	John Holton	Len Dillinger
ABC Marketing	ABC Marketing	ABC Marketing
47 Roberts Road	47 Roberts Road	26 Henry Road
Avon, CT 06001	Avon, CT 06001	Avon, CT 06001

- With unduplication at the Company Level, the three records are duplicates because all three work at ABC Marketing. Only one will be selected; the other two suppressed.
- At the Location Level, only James Schneider and John Holton are duplicates, because they are the only two people who work together at an identical address. Len Dillinger will get a separate mail piece, as will one of the two persons at 47 Roberts Road.
- And, at the Individual Name Level, none of these people are duplicates. Therefore, each will be mailed.

Now that you understand the basics, here are four of the many advanced options available for your consideration:

## **Option A – Efficiently Executing a Mixed Mailing**

Frequently, a mailing is targeted to both consumer and to business names. This is known as a mixed mailing. Under these circumstances, the mailer is faced with a quandary. Often, he or she will want to unduplicate the consumer names at the Family level, and the business names at the Individual Name Level. This is because many believe that prospects at household addresses tend to respond as family units, and prospects at business addresses respond as individuals.

The mixed mailing – with multiple levels of unduplication – is a way of life for most mass market business publications. It is also common among consumer direct response advertisers whose products are targeted to business people. Often, these consumers have personal merchandise delivered to their offices because no one is at home during the day to receive it.

For mixed mailings with multiple levels of unduplication, modern Merge/Purge software allows you to handle this situation very efficiently:

- During the Conversion step, gather the consumer names into a separate "group," so that unduplication can be done at the Family Level.
- At the same time, gather the business names into a second, separate "group," to unduplicate at the Individual Name Level.

This strategy allows two "mini-merge/ purges" to be run simultaneously, thereby avoiding the costly process of passing the mail file multiple times to fill the requirements of the mixed mailing.

## **Option B – Using Duplicates Rates To Predict Test List Performance**

Generally for test lists, the higher the duplicates rate with the house file, the higher the subsequent response rate for that test. This makes intuitive sense. A high duplicates rate with the house file suggests that the individuals on the test list share interests similar to house buyers.

This information can be used to fine-tune the mailing. Test lists with a very low duplicates rate with the house file can be eliminated during the Split and Key step of the Merge/Purge. This avoids incurring in-the-mail costs for test lists that have a low probability of success. This technique is especially useful whenever the net name quantity out of the Unduplication step is larger than the number of promotional pieces that have been printed.

## **Option C – Sophisticated Re-Mailing Of Multi-Buyers**

*[Subsequent, July 22, 2003, comment: The result of the strategy discussed in this section is automatically generated by most if not all of today's Merge/Purge systems.]*

As we have discussed, quality Merge/Purge software must work off percentages, because there is no such thing as a definite duplicate. The corollary is that Overkill – the mistaken identification of unique individuals as duplicates – is unavoidable. As we saw earlier, Overkill can occur even with seemingly identical names and addresses.

Overkill has a cost: a decline in the pool of potential respondents. Response rates are likely to decline as marginal names and addresses are substituted in an attempt to maintain targeted mail quantities. Also, Overkill may affect the overall list cost, depending on the arrangements with list owners for that mailing.

If the direct response advertiser conducts multi-buyer re-mailings, the cost of Overkill can be minimized. This is accomplished by re-mailing the name and address of the duplicate record that was eliminated during the Unduplication step. Here is how:

- For all duplicates pairs identified during the Unduplication step, one record will be eliminated. Before eliminating this record, however, attach its name and address to the retained record.
- During the multi-buyer re-mailing, "write out" this eliminated name and address to the mail-out tape rather than the name originally selected.

In this way, virtually all of the names will be mailed, even those that have been mistakenly identified as duplicates (i.e., "overkilled").

Plan this carefully in coordination with your service bureau because there are a number of details that must be worked out. For example, contacting the "eliminated" name and address should be avoided when doing "sorry we did not hear from you the first time" mailings.

## **Option D – Testing A Location-Level Consumer Merge/Purge**

As we have discussed, most consumer names and addresses are unduplicated at the Family Level. Family Level unduplication, however, will not identify several potential duplicates situations. Only with Location Level unduplication can these situations be identified:

- Female – old surname/new surname (due to differences in change of address timing, one list has been updated to reflect a marriage or divorce, and a second list has not). The same individual – obviously – makes one set of purchase decisions. Therefore a duplicates situation must exist.

- Husband/wife with different last names. Because spouses generally make purchase decisions as a family unit, a duplicates situation is likely to exist.
- Persons with different last names who are living together (either as a family unit or as roommates). Circumstances will determine whether items are bought jointly. Therefore, a duplicates situation will sometimes exist.

The problem with Location Level unduplication is that it is difficult to determine what is a duplicate and what is not. First, there is no way of knowing to which of these three groups a pair of potential duplicates belongs. And, even if the group could be determined, only in the instance #1 could we say with confidence that a duplicates situation exists. (Consider, for example, "Joan Conner" and "J. Jackson" at the identical address, which could be an example of group #1, #2 or #3.)

Therefore, the only way to determine whether Location Level unduplication is appropriate for your business is to do a test. The performance of Location Level "matches" – one sample of which has been unduplicated and one of which has not – can be coded, tracked and compared. The results of this test will indicate whether the majority of these matches are ordering as family units (should be considered duplicates) or as separate individuals (should not be duplicates).

This test can be done at minimal cost without the creation of additional panels with their associated penalties in Presort savings. Because the exact methodology is beyond the scope of this article, contact a technical expert at your service bureau for details.

And finally, be sure that your service bureau has taken adequate steps to ensure that its Location Level unduplication logic is not resulting in excessive Overkill. With some software, Multiple Family Dwelling Units (MFDU's) cause Overkill whenever names and addresses do not contain apartment numbers (i.e., only one record is selected for each address, thereby wiping out entire high rises).

The MFDU problem can be prevented by using information that the U.S. Postal Service includes in its ZIP+4 tables. Associated with each ZIP+4 is one of seven possible "household indicator codes." One of these codes indicates the presence of an MFDU.

This information can be included in the Location Level unduplication logic. Its presence is critical whenever two records contain an identical address, but the apartment number is missing from one or both. Under this circumstance, the records will be considered duplicates only if the address is not an MFDU.

## Summary

Modern Merge/Purge software is both powerful and flexible. The concepts discussed in this article will help the direct response advertiser derive the maximum advantage from future Merge/Purges.

*Jim Wheaton is a Principal at Wheaton Group, which specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data*

*warehouses and marts. He is also co-founder of Data University. Jim can be reached at 919-969-8859, or [jim.wheaton@wheatongroup.com](mailto:jim.wheaton@wheatongroup.com).*