

Evaluating Merge/Purge Systems: Part One

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

Original version of an article that appeared in the July 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 will be explaining 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.

What Is a Merge/Purge?

Before discussing the methodology of our study, a basic definition of the merge/purge process is appropriate:

A merge/purge uses software to combine records from any number of sources, each of which usually is composed of a name and address of a company and/or individual. These sources can include:

- Outside rental lists.
- House lists.
- Suppression lists (e.g., nixies, bad debt and, for prospecting efforts, current customers).

The software compares these records with each other in an attempt to identify and eliminate multiple occurrences. The resulting cleaned output of records is frequently used as a master list for direct mail promotions, from which mailing labels or inkjet addressing is generated.

There are four basic steps to a merge/purge:

- Edit converts each record into a standard format. Generally, the following are also performed during the edit step:
 - Deletion of unwanted or invalid records and characters.
 - ZIP correction.
 - Coding of remaining records for information such as sex, job function and consumer versus business address.
- Unduplication subjects the standardized records to a number of logic tests to identify duplicate situations.
- Split and Key divides the cleaned output from unduplication into groups (strings) of records, and applies the appropriate key code to each record. Separate strings are necessary when different promotional packages are used within one mailing.
- Presort organizes records to take maximum advantage of postal discounts.

Many direct marketers are unaware that the edit step is just as crucial to a successful merge/purge as unduplication. A merge/purge frequently consists of many lists, sometimes numbering into the hundreds, each of which can have a unique record format. Identical information is often recorded in very different ways. For example:

List #1	List #2
M (i.e., "Male")	Mr. Robert James Wheaton
Wheaton	Apt. #4-I
Robert	850 North Dewitt Place
James	Chicago
000850	Illinois
N Dewitt Pl	60611
4-I	
Chicago	
IL	
60611	

It is crucial that list formats be standardized, because the best unduplication software is worthless if a faulty edit results in two different address elements being compared with each other.

The edit step, in conjunction with unduplication, is at the heart of all merge/purge software. It will therefore be the focus of our attention in this series.

Let's attempt to answer the obvious questions:

What Is a Duplicate?

This is the first question to be addressed in every merge/purge. The specific definition will vary according to the mailer's needs and the abilities of the software being used.

For these two pairs of names, which contains a duplicate and which represents separate individuals?

Pair #1	Pair #2
James Wheaton 23 Adams Lane New Canaan, CT 06840	Beth Wilson 52 Devils Garden Road Norwalk, CT 06854
James Wheaton 23 Adams Lane New Canaan, CT 06840	Buffy Walters 52 Devils Garden Road Norwalk, CT 06854

It appears obvious that Pair #1 contains the duplicate record, and the Pair #2 represents two different individuals at the same address.

Surprisingly, however, Pair #2 contains the duplicate:

- Pair #1 consists of two different people, a father and his son, with their respective suffixes (Jr. and III) deleted. It was a constant source of confusion for one of the authors while growing up, as is the case with any son who is named after his father.
- In Pair #2, the first record represents a married woman's professional name, comprised of her given name and maiden surname. As with many women, Beth did not change her name professionally when she married. She decided to retain the name that all of her co-workers and associates were familiar with.

The second record, however, is Beth's nickname and married surname. In her personal life, she opted to take her husband's last name. She picked up the nickname as a child. Most of her friends are not even aware that Beth is her legal first name.

Any merge/purge system, however, that properly identified these two pairs would not be rated highly by any direct marketer.

The point is, there is no way to be 100 percent correct when it comes to merge/purge. Short of contacting every individual who is to be mailed – obviously an impractical approach – a direct marketer can never be absolutely certain what is a duplicate and what is not.

- Every system, therefore, must work off percentages.
- These percentages can be altered by manipulating the parameters that control the unduplication software.
- Each circumstance – every different business or mailing – may require the software to work differently.

The merge/purge user best understands the way his or her business works, and must therefore cooperate closely with the vendor to maximize results for his or her specific needs. Far from being an active partner, however, many direct marketers know little about the merge/purge process. They treat list unduplication as a black box and respond to vendor requests for direction with insights such as, “Do what you think is best,” or “Set it up the way most mailers do for our kind of work.”

During the first half of 1986, we were involved in a detailed comparison of merge/purge systems. We learned many of the specifics that can be used by direct marketers to better understand this process.

Subsequently, the impact that merge/purge can have on a mailer has been reinforced through our work with other clients. We have been amazed at the lack of understanding among otherwise savvy direct marketers. We want to share the insights we have gained in order to facilitate a more intelligent choice of vendors and, therefore, a better working relationship with that vendor to produce the best possible merge/purge.

Both of us were direct marketing managers prior to our consulting work. We wish that we knew then what we know now about the merge/purge process.

Introduction to the Study

Our merge/purge evaluation was based on a thorough comparison of five different systems after an initial review of 25 leading suppliers. The objective was to identify the system that would best meet the specific and complex needs of our client, who has asked to remain anonymous.

It is important to note that our study was not designed to identify the best vendor for all mailers. In fact, we would be willing to guess there is no one vendor that is the best in all facets of performing a merge/purge.

In our review of major merge/purge systems, two considerations became important in selecting the vendors for participation:

- Because the client was looking for an in-house installation, vendors who would not sell their software were not invited to participate.
- Solid business-to-business capability was required in addition to consumer matching. This eliminated a number of vendors who specialize in one or the other.

The lessons we learned can be applied to an evaluation of any situation, whether consumer or business, in-house software or outside vendor, or evaluating possible new suppliers.

The methodology of the study, which was divided into two phases, was straightforward:

In Phase 1:

- Twenty-five potential merge/purge candidates were identified and sent identical requests for system information.
- Those vendors who met preliminary qualifications, and who were interested in participating, were extensively interviewed by telephone.
- The system documentation we received was then thoroughly reviewed.
- Finally, vendors were selected for Phase 2 participation.

In Phase 2:

In Phase 2, consumer and business-to-business capabilities were tested separately. Each test included the following steps:

- The unduplication requirements of the client were documented and approved by client management. These included current practices as well as a wish list of future capabilities not available to the client at that time. By addressing both, the client would maximize the usefulness of the study.
- Detailed job instructions were drafted that reflected these requirements.
- Rented lists and house suppression files chosen were representative of a typical prospect mailing for the client.
- Names were pulled from selected ZIPs and SCFs. This made the existence of duplicates among the various lists more likely, given the limited sample size of the test. We ensured that these selected ZIPs were also a representative mix of the urban, suburban and rural ZIPs encountered in a typical prospect mailing.
- Seed names were inserted to test the full range of each system's unduplication capability. These unique names and addresses, containing specific unduplication problems, were created by Kestnbaum & Company [where the authors worked at the time]. (Seed names will be reviewed in more detail later.)
- Each vendor was given a copy of all tapes, along with instructions on how to run the job.

Vendor Evaluation

We used certain key criteria to evaluate vendors upon receiving their test output:

- Ability to follow detailed and complex instructions.
- Reporting ability and flexibility, which can and did vary significantly from vendor to vendor.
- Overall counts from all phases of the merge/purge, such as the number of names removed during the edit process, names matched to house suppression lists, and rented names identified as duplicates. In the consumer test, records identified and eliminated as business names were included. Likewise, the business test identified and eliminated consumer records.
- Net output, or names available to mail, as defined by each vendor.
- Overkill, derived for both suppression names and duplicates within rented lists.
- Underkill, for these same types of records.
- Consistency of results throughout all ZIPs, to ensure that there had been no hand-manipulation of the data.
- Ability to find seeds.

The names on the rented and house suppression lists supplied by the client filled an important role. They were used to determine the ability of each system to handle our client's current business requirements.

- Typical duplication problems were, by definition, included in the test because the rented lists and suppression tapes were those generally used by our client.
- Because the client used a high proportion of clean lists, we were concerned that the system may not be confronted with a sufficiently large number of difficult duplication problems.

Seed Names

By using carefully constructed seed names, we were able to probe the ultimate potential of each system and its ability to perform some of the functions on the client's wish list.

- Within the seed names, there was a uniform distribution of duplicate problems throughout three basic levels of difficulty. This was unlike naturally occurring circumstances, where simple problems predominated.

- By using a relatively high proportion of difficult seeds, performance differences between systems were accentuated. Also, the client could be assured that difficult problems would be handled properly were they to be encountered in the future.

Eighty-one different types of duplicates were tested via seed names in the consumer test and forty-eight different types in the business test. Our seed names:

- Employed actual street names and numerics, so they could not be eliminated as invalid by sophisticated software.
- Were grouped by problem type. Some seeds contained more than one problem.
- Were reformatted to match the different record layouts of the test tapes and inserted into several different rented lists and house files, making them difficult to identify visually if someone attempted to do so.

Seeds were rated according to three levels of difficulty:

- 1) There were obvious duplicates, such as this one with transposed consonants in the last name:

Record #1	Record #2
Leonard Heal <u>t</u> ey 83 Marble Road Ashburnham, MA 01430	Leonard Heat <u>l</u> ey 83 Marble Road Ashburnham, MA 01430

Either spelling of the last name could be correct, but it would be unlikely to have two different individuals with such similar last names and the identical first name residing at the same address.

- 2) Some were difficult to identify, such as this one with a dropped consonant in both the last name and the street name:

Record #1	Record #2
Leo Ham <u>l</u> iton 3 Car <u>t</u> er Street Leominster, MA 01453	Leo Hailton 3 Cater Street Leominster, MA 01453

In this case, these could be two different individuals with similar last names on similar streets. However, the additional circumstance of identical first names, house numbers and cities make this unlikely.

- 3) Finally, there were those seed name records that would be considered overkill if the system identified them as duplicates. The following example has a number of problems with the last name, street name and street numeric:

Record #1	Record #2
Mickey Kau <u>l</u> life	Kouff <u>l</u> ife M.S.
20 Oliv <u>i</u> er St.	20 <u>0</u> Oll <u>i</u> ver
Boston, MA 02148	Boston, MA 02148

These overkill seeds were constructed such that they might be picked up by a manual inspection, but not by a mathematical or match code system comparison. Thus, they also served as a control device against hand manipulation by the participants.

After all seed names were located in vendor output, the systems were rated on the consistency of performance for each type of duplicate problem. In our scoring system vendors received:

- No credit if they never handled a problem correctly.
- A score of one if they partially or sometimes handled the problem correctly.
- A score of two if they handled the problem correctly in every example.

Scores for difficulty and performance were combined, and a penalty was imposed for overkill; that is, the score was subtracted rather than added to the overall total. This gave us an important basis for comparison of performance.

We believe that when performance on actual lists is comparable between vendors, and the costs of two systems are similar, the one with better seed performance should be chosen. This ensures that software provides flexibility to meet future requirements and is more likely to catch the difficult duplication matches.

A careful evaluation for our client required weeks of examining output, checking each participant's naturally occurring duplicates and suppression hits for reasonability. In addition, each seed name was hand-checked in the output and duplicate listings. What will follow in subsequent articles is a summary of what was learned from that exercise.

There is a general belief among many direct marketers that merge/purge technology has evolved to such an advanced level that differences between vendors are insignificant. This results in

comparisons being made on price alone. Our study suggests that there is, in fact, a wide range in the quality of performance. It also indicates that reputation does not necessarily correlate with results.

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.