

Evaluating Merge/Purge Systems: Part Four

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

Original version of an article appeared in the October 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 Four, focuses on unduplication at the household versus individual level, additional issues for business-to-business direct marketers, and advanced duplicate recognition techniques.

Unduplication at the Household vs. Individual Level

Most consumer mailers choose to unduplicate at the household level rather than at the individual level, because it is generally believed that one mailing package to each address is optimal. A few do opt for unduplication at the individual level, because it is to their advantage to have more than one customer per household. These mailers are faced with a number of additional decisions concerning the definition of a duplicate. All of these focus on the first name:

For example, should these be duplicates if all other aspects of the record are identical?

- Robert Smith and R. Smith – One record contains a first name and the other an initial.
- Robert Smith and Bob Smith – The first record has a proper name and the second a commonly used nickname for that proper name.
- Bob Smith and R. Smith – One record has a nickname and the second an initial of the related proper name.
- Mrs. Robert Smith and Paula Smith – Here, first names indicate opposite sex, but the title indicates the same sex. Are there two people here, perhaps mother and daughter, or merely one individual?
- Robert Smith and Robert Smith Jr. – One record contains a suffix and the other does not.
- Roberto Smith and Roberta Smith – Here, a slight difference of one vowel in the first name could indicate either different people of different sexes, or merely a single input error. Many parents give names to their children that are similar to their own. Again, there is no right or wrong answer.
- Robert Smith and R. Smith (gender code) – The interaction of gender codes with first name permutations can make a difference. For example, the handling of Robert Smith and R. Smith in the first example will depend upon whether R. Smith has been coded “male,” “female” or “unknown.” And, the Roberto/Roberta decision in example six would also be affected by the sex code.

Issues for Business-to-Business Marketers

So far, we have discussed issues that apply to both business-to-business and consumer unduplication efforts. A number of additional, unique issues must be faced in a business merge/purge, making it far more complex than the consumer version:

Different Levels of Unduplication

First, unduplication can be at the company, location or individual level:

Record #1	Record #2	Record #3
Harold Abbott, Jr. CNL, Inc. 29 Woodsons Ave. Norwalk, CT 06850	John Holton CNL, Inc. 29 Woodsons Ave. Norwalk, CT 06850	Charles Dillinger CNL, Inc. 26 Henry Street Norwalk, CT 06850

Here are three records with three individuals at the same company: two at one address, and the third at a different address.

- With unduplication at the “company level,” all of the records are duplicates, because all are located at CNL, Inc.
- With unduplication at the “location level,” number one and number two are duplicates, because only these two both work at 29 Woodsons Avenue.
- And, with unduplication at the “individual” level, none match, because all refer to separate individuals.

Multiple Company Names

A second issue in business-to-business unduplication is that companies frequently have several names or abbreviations of names.

Imagine for a moment that you want to send only one mailing piece to each company you contact. Ask yourself how you would know whether or not these are separate companies within a single address, multi-office complex:

Record #1	Record #2	Record #3	Record #4
CNL, Inc. 29 Woodsons Ave. Norwalk, CT 06850	Software Sales Group 29 Woodsons Ave. Norwalk, CT 06850	S.S.G. 29 Woodsons Ave. Norwalk, CT 06850	Systems Design Group 29 Woodsons Ave. Norwalk, CT 06850

- The first record contains the parent company name: CNL, Inc.

- The second, The Software Sales Group, is a division of CNL, Inc.
- The third record contains a frequently used abbreviation of the division in record two.
- The fourth is a second division of CNL.

Multiple Addresses

A third issue is that many companies have multiple street, building name, and/or post office box addresses:

Record #1	Record #2	Record #3	Record #4
Software Sales Group 29 Woodsons Ave. Norwalk, CT 06850	Software Sales Group 91 Beeker St. Norwalk, CT 06851	Software Sales Group P.O. Box 106 Norwalk, CT 06850	Software Sales Group 26 Henry Street Norwalk, CT 06850

- The first record is the standard address for one division within the firm.
- The second is the warehouse address for this division. It is, however, at the same location, with the only difference being the street entrance.
- The third record is one of many post office boxes maintained by the company.
- The fourth is the research and development center for this division, located several miles away.

The business-to-business problems illustrated in these sets of examples are extremely difficult to handle by themselves. When several such problems appear within a single group of records, however, the challenge becomes almost overwhelming.

Additional Business-to-Business Complications

There are, in fact, still more complicating factors:

- Executives often hold several titles, spread among multiple divisions and geographic locations:

Record #1	Record #2
Harold Abbott, Jr. Vice President CNL, Inc. 29 Woodsons Ave. Norwalk, CT 06850	Harold Abbott, Jr. Director of R&D CNL, Inc. 26 Henry Street Norwalk, CT 06850

- Harold Abbott is listed as a vice president on the roster at the CNL corporate headquarters.
- Abbott, however, spends most of his time at S.S.G.’s research and development location on Henry Street.
- In addition, business records frequently contain non-address characters within the individual name field, such as “care of” and “attention.” Sometimes, only a title will accompany a non-address character (e.g., Attn. Warehouse Mgr.)
- Transposition of fields is a very common problem.

It is for these reasons that good consumer performance does not necessarily guarantee strong business capability. A business-to-business merge/purge is so different that there are merge/purge vendors who perform only this type of job.

Advanced Duplicate Recognition Techniques

Advanced duplicate recognition techniques are being developed to deal with the complex problems found in a business-to-business merge/purge. Central to these techniques is an attempt to catch additional duplicates by “linking” together multiple records that have various but different fields in common. Some vendors are successfully extending these advances to the consumer arena.

Assume, for example, that you want to send just one mailing piece to each company:

- Record #1 is the first to be examined. The mail piece will be addressed to this individual.

Record #1

James Carlson
Allen Services, Inc.
1426 Hope Street
Stamford, CT 06904

- Record #2 has the company name in common with number one. It is therefore a duplicate, and should not be mailed. In addition, we have established, for the first time, that Allen Services is in Suite 406.

Record #2

Allen Services, Inc.
1426 Hope Street, Suite 406
Stamford, CT 06904

- Record #3 has the suite number in common with record number two, which indicates that Greg Sterrett is also an Allen employee, and should therefore be suppressed. Notice that all three records have been indirectly linked together, which would not normally be possible.

Record #3

Greg Sterrett
1426 Hope Street, Suite 406
Stamford, CT 06904

- Record #4 has the suite number in common with number three, and should therefore also be suppressed as a duplicate. It also establishes, again through this linking process, that Johnson, Allen & Associates is an alternate company name.

Record #4
Janice Cuperstein Johnson, Allen & Assoc. 1426 Hope Street, Suite 406 Stamford, CT 06904

- Record #5 has this alternate company name in common with record number four. This indicates that Mike is also an Allen employee, who should therefore not be mailed. As the linking process continued, it is now known that Allen Services has a post office box.

Record #5
Mike Blumenstein Johnson, Allen & Assoc. P.O. Box 557 Stamford, CT 06904

- And finally, via a post office box match with Record #5, we can establish that M.T. Pepitone is an Allen employee, who should also not be mailed. The overall result of this linking process has been that only one of these six records has been retained.

Record #5
M.T. Pepitone P.O. Box 557 Stamford, CT 06904

Linking techniques are at the leading edge of merge/purge technology, and are still being perfected. We cannot promise that any currently existing software could handle these six records perfectly. It is clear, however, that the technique has impressive potential to uncover duplicates that would normally go undetected.

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.