



Customer Data Quality

Building the Foundation
for a One-to-One
Customer Relationship

A White Paper

Revised, 1999

Firstlogic, Inc.

Copyright © 1999 Firstlogic, Inc. All rights reserved. No part of this publication may be stored in a retrieval system, transmitted or reproduced in any way, including but not limited to photocopy, photograph, magnetic or other record, without prior written agreement and permission of the author, except for such limited purposes as may be authorized by the Copyright Act of 1976. Printed in the USA.

Customer Data Quality: Building the foundation for a one-to-one customer relationship

In today's competitive business environment, every day brings new challenges for customer acquisition and retention. To stay ahead, companies are turning to decision support solutions to help identify and manage their customer relationships. Solutions such as data warehouses or data marts, provide a solid foundation of accurate information upon which they can base their decisions.

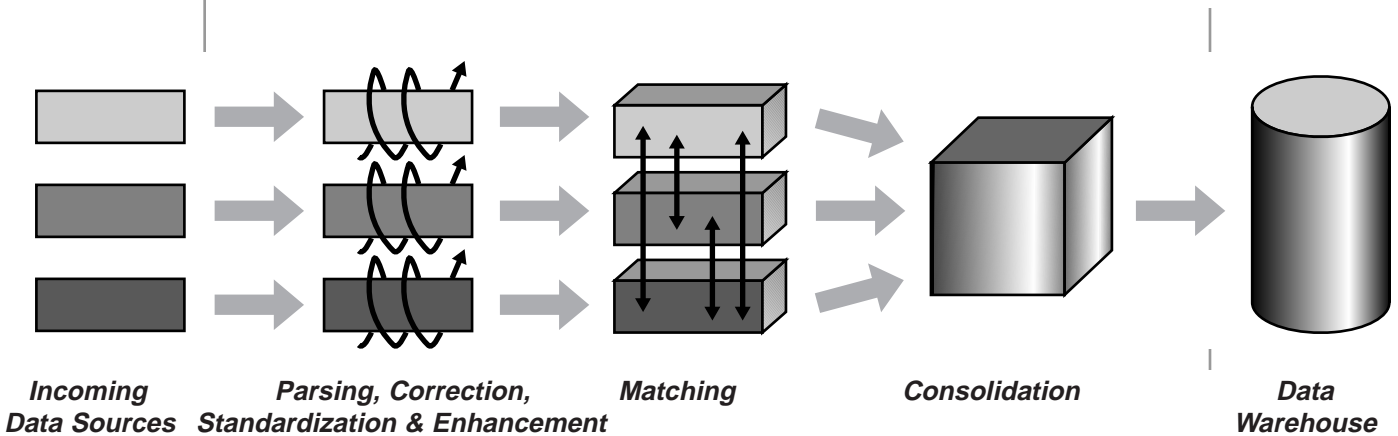
Accuracy is one of the biggest obstacles blocking the success of many data warehousing projects. In fact, according to numerous META Group surveys, data quality remains the number one challenge facing companies as they implement their data warehouses. META Group Program Director John Ladley estimates that, when building a data warehouse, 10 to 20 percent of the raw data used is corrupt or incomplete in some way. It is not unusual to discover that as many as half the records in a database contain some type of information that needs to be corrected.

One of the most crucial areas of data quality is customer information. For the most accurate information to support your business, you will need to incorporate data quality into each critical step — extraction, transformation, consolidation, and maintenance. Data quality is especially important to accurate consolidation because it allows you to recognize and understand customer relationships. As a result, you can gain a clear picture of your customers, analyze their buying patterns, and predict future sales.

Which data is most important to your company?

Before you evaluate data quality solutions, you need to determine the type of data that is most important to your company. Will you focus primarily on customer information, or are you most concerned with data such as part numbers, prices, and product descriptions? Once you make this decision, you can choose the data quality products most appropriate for your applications.

In either case, you will find that most data quality and consolidation products fall into one of two categories: *data-referenced solutions*, which combine reference tables with sophisticated algorithms, and *non-data-referenced solutions*, which rely on algorithms alone. In data warehouses where customer data is essential, data-referenced software is more effective. This type of software features an extensive knowledge base of empirical data, which allows you to enhance and improve the quality of your information.



Data quality is achieved in three stages: cleansing, matching, and consolidation. In the data cleansing stage, the data is parsed, corrected, standardized, and enhanced for accurate matching. In the matching stage, comparisons are made within and across your data sources to locate similar information. Finally, the matching data elements are consolidated and placed into a data warehouse, data mart, or other data storage area.

This paper will help you understand the foundation for building data quality into a data warehouse: parsing, correction, standardization, enhancement matching, and consolidation. You'll discover the principles and inherent problems of data quality and data consolidation, and the available solutions to help you manage them.

Parsing: Do you know what your data is?

Parsing is the first critical component in data cleansing. This process locates, identifies, and isolates individual data elements in your customer files. These components may include such data as a customer's first name, last name, title, company name, street address, city, state, or ZIP Code. Parsing makes it easier to correct, standardize, and match data because it allows you to compare individual components, rather than long strings of data.

Parsing the elements

Input Data

BETH CHRISTINE PARKER, SLS MGR
REGIONAL PORT AUTHORITY
FEDERAL BUILDING
12800 LAKE CALUMET
HEGEWISCH IL



Parsed data

| | |
|----------------|-------------------------|
| First Name: | BETH |
| Middle Name: | CHRISTINE |
| Last Name: | PARKER |
| Title: | SLS MGR |
| Firm: | REGIONAL PORT AUTHORITY |
| Firm location: | FEDERAL BUILDING |
| Range: | 12800 |
| Street: | LAKE CALUMET |
| City: | HEGEWISCH |
| State: | IL |

Parsing is a vital step for the cleansing and matching stages. This example shows how parsing identifies and isolates the individual elements of an input record.

There are several obstacles to parsing that may later hinder successful matching. Perhaps the most pervasive problem is discrepancies in the metadata – information about the data in your database. For example, the information in a field may not match its metadata profile. Inconsistent definitions and multiple data sources make it difficult to determine if fields possess the same characteristics from one source to the next.

Other obstacles include:

- *Misfielded* data – Data that is placed in the wrong field, such as name data in an address field.
- *Floating data* – Customer data that may be contained in different fields from record to record, resulting in data “floating” between fields.
- *Extraneous information* – The data may contain irrelevant or blank fields.
- *Atypical words* – Records may include ethnic, multicultural, and hyphenated names; unusual titles; abbreviated business names; industry-specific acronyms; etc.
- *Inconsistent structures and formats* – Operational, purchased, and exchanged data sources may be formatted differently from each other or from the data warehouse.

Inconsistent field formats

| <i>Field format</i> | | <i>Field data</i> |
|-------------------------------|--------|-------------------------------|
| First-middle-last-title |▶ | Juanita E. Alvarez President |
| Last-title-comma-first-middle |▶ | Alvarez President, Juanita E. |
| Last-comma-first-middle-title |▶ | Alvarez, Juanita E. President |

Inconsistent field formats, such as the ones indicated above, present an obstacle for unparsed data, making it difficult to combine all data sources into one data warehouse.

Correction: How do you know your data is accurate?

Once you’ve parsed your data, you are ready to begin the next phase of the data cleansing process — correction. Customer information is the most difficult type of data to cleanse and validate. If your data comes from a variety of sources, you may encounter:

- Variations in abbreviations, formats, etc., because of individual preferences of the person entering the information
- Misspellings caused by phonetic similarities during telephone data entry
- Outdated information due to name and address changes
- Transpositions resulting from keying errors

Correcting the elements

| Parsed data | | <i>The record on the left contains an alias street name and a vanity city name. The record below shows how those elements were corrected for matching.</i> |
|--------------------|-------------------------|--|
| First Name: | BETH | |
| Middle Name: | CHRISTINE | |
| Last Name: | PARKER | |
| Title: | SLS MGR | |
| Firm: | REGIONAL PORT AUTHORITY | |
| Firm location: | FEDERAL BUILDING | |
| Range: | 12800 | |
| Street: | LAKE CALUMET | |
| City: | HEGEWISCH | |
| State: | IL | |

| Corrected data | |
|-----------------------|--------------------------|
| First Name: | BETH |
| Middle Name: | CHRISTINE |
| Last Name: | PARKER |
| Title: | SLS MGR |
| Firm: | REGIONAL PORT AUTHORITY |
| Firm location: | FEDERAL BUILDING |
| Address: | 12800 SOUTH BUTLER DRIVE |
| City: | CHICAGO |
| State: | IL |
| ZIP: | 60633 |
| ZIP4: | 2398 |

The only way to intelligently correct and verify your data is to use software that references a reliable secondary data source. In many instances, correction is used only to prepare data for matching – the original records remain unchanged.

For instance, some neighborhoods have “vanity” city names (i.e., “Hollywood” versus “Los Angeles”) or “alias” street names (i.e., “Valley View Mall” versus “Highway 16”). Some residents living there may prefer to use one city or street name over the other, and therefore may be more likely to

respond to an offer bearing the name they prefer. However, by recognizing that both names refer to the same location, you have a better chance of identifying matches. By correcting the data, you can locate matches while still respecting clients' preferences.

Standardization: Is your data ready for matching?

Standardization, the next process in data cleansing, allows you to arrange customer information into a preferred and consistent format. Some of the biggest challenges for accurate standardization of customer data include:

- *Inconsistent abbreviations* – such as International Harvester, Intl. Harvester, Interntl. Harvester, Internatl. Harvester
- *Unusual titles* – for example Graduate, Realtor Institute (GRI); Specialist in Residential Appraisal (SRA); and Member, Appraisal Institute (MAI)
- *Misspellings and variant spellings* – i.e., Kwik, Quik, Quick

Software solutions that integrate secondary data sources perform more effective standardization, and allow you to make business decisions with confidence. For example, some software is certified to standardize addresses using the most widely accepted data source – the U.S. Postal Service's (USPS) National Directory. "Cleansed information across all database records ensures accuracy, and gives users a clear, consistent view of each customer," says Lou Guercia, vice president of business development for Tessera Enterprise Systems.

Standardizing the elements

Corrected data

First Name: BETH
Middle Name: CHRISTINE
Last Name: PARKER
Title: SLS MGR
Firm: REGIONAL PORT AUTHORITY
Firm location: FEDERAL BUILDING
Address: 12800 SOUTH BUTLER DRIVE
City: CHICAGO
State: IL
Zip: 60633
Zip4: 2398

Standardized data

Pre-name: Ms.
First name: Beth
First name match standards: Elizabeth, Bethany, Bethel
Middle name: Christine
Middle name match standard: Kristine
Last name: Parker
Title: Sales Mgr.
Firm name: Regional Port Authority
Firm location: Federal Bldg.
Address: 12800 S Butler Dr
City: Chicago
State: IL
Zip: 60633
Zip4: 2398

Once data has been corrected, as in the record above, the elements undergo further cleansing. The data is standardized according to the criteria that you indicate. These same criteria will come into play again later in the matching stage.

When cleansing certain types of data (names, business names, professional titles, etc.), “match standards” will facilitate more successful matching.

Match standards – typical representations of a data element – can only be assigned by sophisticated standardization software.

Match standards

General Electric> GE
Gen. Electric> GE

Al> Albert
.....> Alfred
.....> Alan
.....> Alphonse
.....> Almon
.....> Alexander

Sophisticated standardization software can assign match standards to facilitate more successful matching for elements such as personal and business names.

Some software can also standardize other customer information, such as pre-names, post-names, titles, and business locations (“Doctor” to “Dr.,” “Junior” to “Jr.,” “Floor” to “Flr.,” etc.). It may also identify genders, based on empirical name data, to give you a better understanding of your customers for one-to-one marketing.

Enhancement: Do you have all the data you need?

Enhancement, the final step in data cleansing, appends new data and completes missing information. The new data helps you better understand your customers and predict future buying behavior. The type of information appended may include:

- *Demographic data* – such as age, presence of children, income, education level, and marital status for individuals; and sales volume, number of employees, and SIC code for businesses
- *Geographic data* – for instance apartment or suite numbers, missing address elements, telephone numbers, latitude/longitude information, county codes, and political districts
- *Behavioristic data* – i.e., credit worthiness, preferred communication channels, and purchase recency, frequency, and monetary value (RFM)
- *Psychographic data* – such as hobbies, interests, and political affiliation
- *Event driven data* – including marriage, birth of a child, recent move, and corporate merger
- *Computed data* – for example credit ratings, lifestyle clusters, lifetime value calculations, and responsiveness scores

Enhancement may be accomplished by sending your customer data to another firm for processing, purchasing an outside source of customer data, or surveying your customers and manually updating their information.

These alternatives, however, may present some additional challenges:

- External processing may be time and cost-prohibitive.
- Your customer records may contain data too sensitive to send to an outside firm for processing.
- Internal resources and expertise may be insufficient to integrate an external data source.
- Customers may be difficult to reach or unwilling to provide additional information.

Enhancing the record

Standardized Data:

| | |
|-----------------------------|---------------------------|
| Pre-name: | Ms. |
| First name: | Beth |
| First name match standards: | Elizabeth Bethany, Bethel |
| Middle name: | Christine |
| Middle name match standard: | Kristine, Cristine |
| Last name: | Parker |
| Title: | Sales Mgr. |
| Firm name: | Regional Port Authority |
| Firm location: | Federal Bldg. |
| Address: | 12800 S Butler Dr |
| City: | Chicago |
| State: | IL |
| Zip: | 60633 |
| Zip4: | 2398 |



Cleansed and Enhanced Data (data source 2):

Ms. Beth Christine Parker, Sales Mgr.
Match standards: FN: Elizabeth, Bethany, Bethel;
MN: Kristine, Cristine
Reg Prt Auth
12800 S. Butler Dr.
Chicago IL 60633-2398
Phone: 708-555-1234
Fax: 708-555-5678
Gender: Strong, female
Account number; 00192744
SSN: 123-45-6789
Latitude: 41.662364
Longitude: -087.576274
MSA Code: 1600
Census block: 5104009
Cart: C004
DPBC: 02
LACS: F
County Code: 031
Lot: 0031
Lot order:A

Many data enhancement sources and techniques are available. It is important to select a solution that contains the type of data best suited for predicting the behavior of your customers and a method that can be effectively managed by your firm.

Matching: Can you find the duplicates in your data?

Matching allows you to identify similar data within and across your data sources. This is the ‘heart’ of data warehousing. Using cleansed information and match standards, you can eliminate duplicate representations and consolidate all information about each individual customer or an entire household. This will help you to:

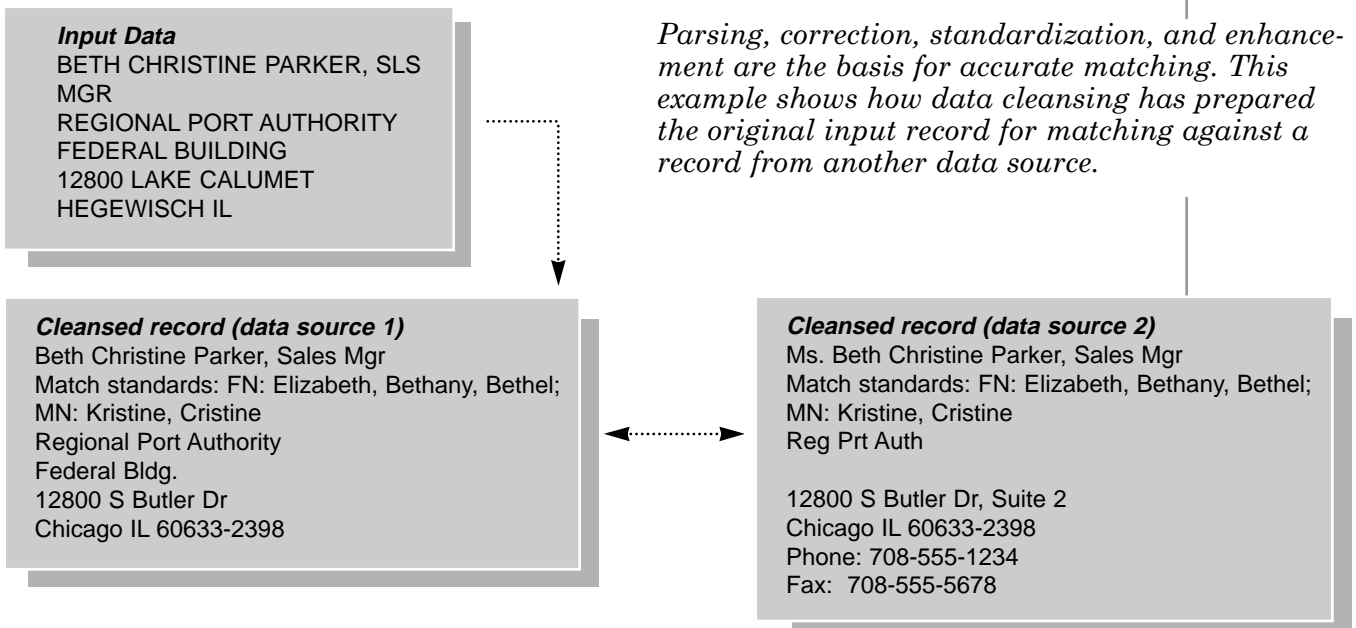
- Truly “see” each customer or household, and generate accurate data about them
- Enhance response rates of marketing promotions
- Reduce the risk of offending customers with repeat offers
- Identify trends and patterns to accurately target new prospects

One of the greatest challenges in matching is creating a system that incorporates your “business rules” – criteria for determining what constitutes a match. These business rules will vary from one company to another, and from one application to another. In one instance, you may require that name and address information match exactly.

In another, you may accept wider address variations, as long as the name and phone number match closely. Some additional challenges to matching business-to-business data include:

- Company mergers, acquisitions, or corporate name changes
- Relationships between divisions, subsidiaries, and parent corporations
- Business acronyms (i.e., NASDAQ or NYNEX)
- Initialisms – the first letter of one or more words in a title or phrase that are sounded one by one (i.e., AT&T or CIA)

Matching records



Consolidation: You’ve matched your data – now what?

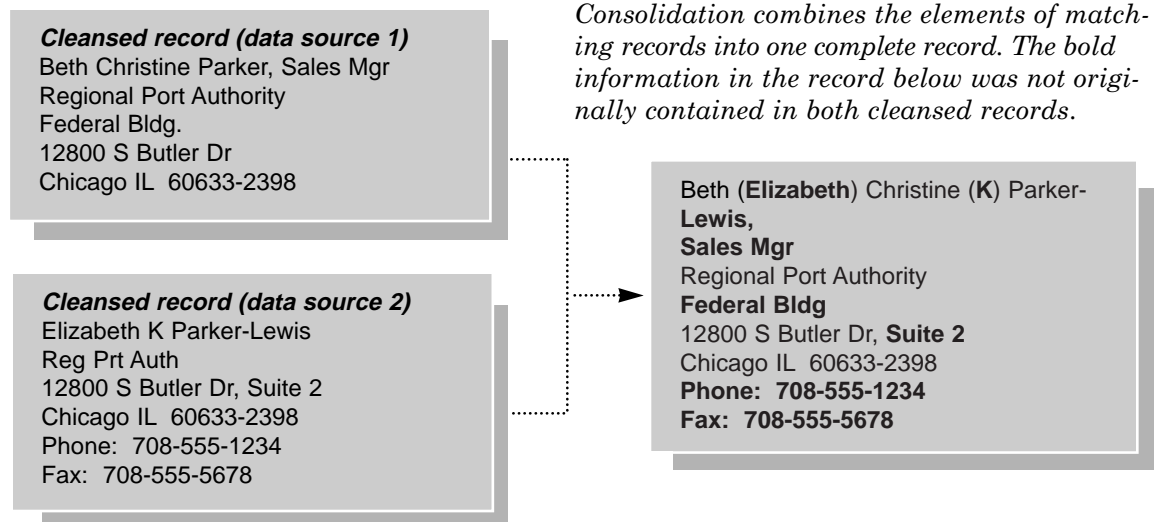
Once you’ve located the matching records in your data, you can identify relationships between customers and build a consolidated view of each. This critical component of successful one-to-one marketing allows you to gain a clearer understanding of your customers. “One-to-one marketing

allows organizations to better serve the customer at every point of contact,” according to Greg Holzwarth, Epsilon’s executive vice president. When you base marketing, telemarketing, sales, customer service, and accounting decisions on clean and accurate data, you can more easily retain customers by anticipating their needs.

There are two methods for consolidation, both of which are essential for most data warehousing and one-to-one marketing applications. The first consolidation process combines all of the data on any given customer using all of the available data sources. The second process – customer relationship identification — reveals links between your customers.

There are two common types of customer relationship identification: householding and business grouping. Typically, householding links consumer records that contain the same address and last name. Business grouping combines business records that share such information as company name, address, department, or title.

Consolidating records



By identifying the characteristics and buying habits of a group or household, you can create special offers and better target direct marketing efforts.

As you combine your data on each customer, you will need to determine priorities between data sources and specific data fields. For example, some records may contain more complete information, but may be from an unreliable source (see the ‘Consolidation Systems’ section on page 19).

Data cleansing solutions: What are you going to do about it?

As previously mentioned, most data quality tools fall into one of two categories: data-referenced and non-data-referenced. Both systems allow you to reformat your data, but data-referenced systems go a step further.

Data-referenced solutions combine reference tables with sophisticated algorithms. In most cases, this type of solution deals better with the challenges posed by customer-centric data warehousing applications. By accessing the empirical data in their reference tables, these solutions can intelligently parse, correct, standardize, and enhance your critical customer information, rather than simply format it.

Non-data-referenced solutions rely on algorithms alone. These solutions can handle less-challenging data, such as account balances, dates, product numbers, or test scores.

In either case, you should choose a data quality solution that includes:

- Customizable standardization options

- The ability to retain original data and save corrected information
- Flexible output options

Data-referenced systems: Best for customer data

For customer-centric data warehouses, the most effective systems use empirical information. They provide the foundation for logically consolidating multiple, diverse databases.

Non-data-based software vs data-based software

| <i>Processing with non-data-based software</i> | |
|---|-----------------------------|
| JERRY BENSON | Jerry Benson |
| 704 MARTIN LUTHER KING BOULEVARD | 704 Martin Luther King Blvd |
| PRINCETON FLORIDA 33031 | Princeton FL 33031 |

| <i>Processing with data-based software</i> | |
|---|-------------------------------------|
| Jerry Benson | Mr. Jerry Benson |
| | First Name Match Stds: |
| | Jerome, Jerald, |
| | Jeremiah, Geraldine, Jerilyn |
| 704 MARTIN LUTHER KING BOULEVARD | 704 SW 4th St. |
| PRINCETON FLORIDA 33031 | Homestead FL 33030-6914 |

Non-data-based software offers limited functionality and flexibility compared to data-based software. The items in bold show the information that data-based software appends and standardizes over non-data-based software.

The best systems for address cleansing use the National ZIP+4 Directory developed by the (USPS). They are designed with high-speed engines that are certified as being 99 percent or more accurate. By leveraging the USPS' multi-million dollar, multi-year investment, they offer the most comprehensive source for address parsing, correction, standardization, and enhancement.

The most effective data-quality software can also parse names, titles, business locations, business names, and financial terms such as trustee, retired, and deceased. By using empirical data and user-modifiable tables, these systems can more accurately locate ‘floating’, unfielded, or incorrectly parsed data. Because they are founded on actual data, you can properly process ethnic, hyphenated, or atypical names; and recognize business name and location data.

Data-referenced software can also perform enhancement for a more complete customer view. When choosing data-referenced data-cleansing systems, look for the ability to:

- Recognize formal and informal street names, multiple addresses (i.e., Post Office Box and street address in the same record), and vanity city names
- Minimize users’ learning curve using a significant base of pre-defined data – you should not need to train the product
- Perform enhancement with data already integrated into the reference files
- Flag outdated or invalid address data (i.e., rural-route address converted to ‘9-1-1’ address, or a nondeliverable address)
- Access and modify reference tables
- Assign match standards

Each of these cleansing components is essential for successful matching and consolidation.

Matching and consolidation solutions: How will it all come together?

If your data consists of unparsed data strings, you are in danger of missing many matches or of erroneously consolidating different customers. However, corrected and standardized data in discrete fields allows most matching systems to successfully detect:

- Transposed, missing, or extra characters
- Transposed words
- Phonetic errors
- Acronyms and abbreviations

Matching Systems

There are several varieties of matching systems available, each offering a different way to arrive at a match:

- *Key-code matching* performs identical comparisons using the first few characters in one or more fields. This primitive method is rarely practiced because it uses only a small sub-set of the data, which can result in many false matches.

- *Soundexing* detects phonetic similarities, such as ‘f’ and ‘ph’ or Quick and Kwik. These errors often result from data received over the telephone, particularly with data that can’t be standardized. However, soundexing is inadequate as a sole solution because it can only detect phonetic errors.
- *Similarity matching* – also referred to as ‘fuzzy matching’ – can identify matches by computing a degree of likeness between two discrete components. Because identical matches are not required, it can adjust for spelling, phonetic, typographical, and transpositional errors.

Similarity matching is widely considered the best matching method. It is especially valuable for data that cannot be standardized, such as last names, business names, and house numbers.

- *Weighted matching* can be used in conjunction with soundexing or similarity matching. It allows you to indicate the relative importance of fields that determine a match.

Similarity matching

In the address 1001 Rose St.:

1001 and 101 have a similarity score of 85%
1001 and 1010 have a similarity score of 75%
1001 and 1025 have a similarity score of 50%

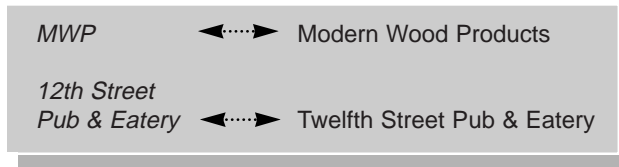
Similarity matching considers all characters in a field and their position to determine the degree of a match. In this example, the degree of the match is indicated by the similarity score.

Special purpose algorithms, which are extensions of similarity matching, apply exception logic ('if/then' rules) to traditional match rules. There are four categories of special purpose algorithms:

- *Special-case field logic* customizes matching techniques for specific fields. For example, these algorithms are used to identify matches between acronyms or initialisms and their full business names, or numeric components within names.
- *General-case field logic* applies additional match logic when it encounters certain anomalies, such as blank fields. In comparing discrete components, it allows you to specify when blank fields should be considered matches against fields that contain data.
- *Special-case multi-field logic* adjusts weighting, which depends on the data found in specific sets of fields. When householding, for instance, this method would assign a higher value to the 'name' field when the address is an apartment complex and the 'unit number' field is blank.
- *General-case multi-field logic* performs a second match when it encounters specific anomalies, regardless of the fields in which they are found. For example, it could search for elements with low parse-confidence scores, concatenate them, and compare the resulting data strings.

Special-case field logic

With special-case field logic, you can identify a match between a full name and its initialism, or between numbers in a name and their full spelling.



A combined approach – incorporating similarity, weighting, and special purpose algorithms – is usually best. As many as six to ten levels of matching may be necessary for complete data consolidation or maintenance.

When using a combined approach, you need the flexibility to modify business match rules. This is important as the reliability of certain data may vary between records and sources. For instance, people who move often have unreliable contact data. When this happens, your combined solution should provide the flexibility to use alternate fields as match criteria, such as social security number, credit card numbers, account numbers, and date of birth.

Consolidation Systems

The data cleansing and matching processes lead to one end result: accurate data consolidation. To build on this foundation, you'll need a flexible consolidation solution to combine existing operational data and maintain incoming data feeds.

Some key components of proven consolidation solutions allow you to:

- *Prioritize incoming data sources.* In-house databases are usually more reliable than purchased or rented data because they tend to be more up-to-date.

- *Prioritize fields.* Fields that have been cleansed and verified tend to be more reliable than those that have not. Again, source and recency are important considerations.
- *Maintain sources of original data.* Complete metadata allows you to trace data errors or discrepancies back to the source.
- *Identify unreliable or missing data.* Once identified, if it is economically feasible, you can request information directly from the customer or from a valid outside source.

Depending on the scope of your project, you may wish to approach consolidation one step at a time. If you don't have the resources to build an enterprise data warehouse, you may choose to start with a data mart. With each small success, you will gain expertise, confidence, and continued support for completing your data warehouse.

Don't overlook the foundation

As you implement your data warehouse, don't overlook its foundation – data quality – which is essential for quality matching and consolidation. It is critical to first determine which data is important, and to choose the most appropriate cleansing tools. When working with customer data, a data-referenced solution is best.

With a variety of matching solutions available, look for one that combines different types of matching algorithms. This will allow you to more accurately locate, understand, and target your customers.

Notes

By building parsing, correction, standardization, enhancement, matching, and consolidation into your system, you can meet your goals, which may enable you to:

- Make better and more immediate marketing decisions
- Increase market share
- Increase revenue
- Increase profitability
- Detect fraud more accurately
- Improve customer service
- Ensure clean feeds to on-line analytical processing (OLAP) and data-mining tools
- Implement your solution faster than building your own
- Understand your customers
- Guarantee the success of your project

The best course of action is to choose tools that offer the flexibility and accuracy your project demands. By implementing these tools, you will secure a solid foundation for building one-to-one customer relationships.

Firstlogic offers solutions that enhance the value of customer data and customer communications. The company's i.d.Centric brand offers customer data quality and consolidation software and services that make data warehousing applications more effective. The Postalsoft brand provides postal automation, mail preparation, and document generation software and services that make print and mail operations more effective and profitable. For more information call toll-free 1-888-215-6442, or visit our Web site at www.firstlogic.com.