DUPLICATE ENTRY DETECTION IN MAILING AND PARTICIPATION LISTS

by

HEBER REGAL NORCKAUER, JR.

B.S. Physics, Louisiana State University (1972)
M.S. Physics, Louisiana State University (1974)

SUBMITTED TO THE

SLOAN SCHOOL OF MANAGEMENT

IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS OF THE DEGREE OF

MASTER OF SCIENCE IN MANAGEMENT

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 1990

Signature of Author ____
Sloan School of Management
May 4, 1990

Certified by ____
Stuart E. Madnick, Ph.D.
Professor, Management
Thesis Supervisor

Approved by ____
Alan F. White, S.M.
Associate Dean for Executive Education

DUPLICATE ENTRY DETECTION
IN
MAILING AND PARTICIPATION LISTS

by

HEBER REGAL NORCKAUER, JR.

Submitted to the Sloan School of Management on May 4, 1990
in partial fulfillment of the requirements for the
Degree of Master of Science in Management

ABSTRACT

An investigation of the state-of-the-art in duplicate detection as performed in the mailing list/participation list industry and review of future technology applicable to this industry were performed. A prediction of the direction the industry will take in performing duplicate detection in the future was made.

Following a literature search, the principal players in the industries which manage mailing and participation lists were contacted and interviewed. These included the United States Postal Service and a number of members of the Direct Marketing Industry. Current literature on artificial intelligence and other ideas were reviewed for their applicability. Comparison of the requirements of the industry and the emerging technology was made and conclusions were drawn.

The conclusion of the activity is that the algorithms currently in use are mature rule based expert systems and will only become more efficient through further gradual (i.e., evolutionary) maturation. Three improvements are forecasted. First, actions by the Postal Service to improve the database against which addresses are compared (i.e., issuance of an authoritative list of addresses or compilation of National Change of Address data) will cause continued improvement in performance. Second, the evolution of automated transactions (e.g., on-line services and funds transfer) will significantly reduce input discrepancies and improve performance. Last, well into the next century with the evolution of large neural network systems a revolutionary improvement in duplicate detection might result. The ability of the neural network system to compete with the then state-of-the-art expert system is questioned.

Thesis Supervisor: Dr. Stuart E. Madnick
Professor, Management

ACKNOWLEDGMENT

In the course of this research activity numerous individuals and organizations lent assistance. The author wishes to acknowledge this assistance with the meager thanks that one can present here. Without their assistance completion of this activity surely would not have been possible.

Specifically, thanks to Professors Madnick, Wang and Siegel for their expert help, insight and review. Thanks to Almudena Arcelus, my colleague, whose insight and assistance and sharing of background information lent much assistance. Thanks to the United States Postal Service and the Direct Marketing Industry companies which took time out of their busy schedules and provided a significant amount of material to the project. Thanks to my wife, Sheila, for taking a year away from her career to come with me to Boston. Lastly, thanks to the United States Army for making attendance at Massachusetts Institute of Technology possible.

THANK YOU ONE AND ALL!

To my wife, Sheila.

TABLE OF CONTENTS

Chapter	Topic	Page
	Abstract	2
	Acknowledgment	3
	Dedication	4
	Table of Contents	5
	Preface	6
I	Introduction	8
II	Methodology	10
III	Semantic Confusion	13
IV	The Industry	17
V	Current Solutions	34
VI	Futuristic Solutions	63
VII	Conclusions	75
	Appendices	79
	Bibliography	90
	Biographical Sketch	94

PREFACE

The detection of duplicate entries in mailing and participation lists is an interesting and complex subject. At first thought this is a simple problem, but when one recognizes the huge quantities of data and alternatives often required to be analyzed to look for a duplicate, one quickly realizes the enormity of the task.

The author became interested in the study of the state-of-the-art in duplicate detection through an inquiry to Ducks Unlimited, a non-profit organization dedicated to the preservation of waterfowl habitat, about what management problems they were encountering. Ducks Unlimited indicated, because of the large number of alternative ways of joining the organization, they often were duplicating membership entries and looking foolish to their members when they mailed out duplicate magazines, renewals and other solicitations. They were interested in furthering their capability of determining duplicates both within the membership list and also in their solicitation of new members.

Inquiry about this topic quickly brought the author in contact with Professors Stuart Madnick and Richard Wang who have been actively conducting the Composite Information Systems Laboratory (CISL) project at the Massachusetts Institute of Technology's Sloan School of Management. One of the principal problems the CISL project has encountered relates directly to the detection of duplicates within its accessible databases. Hence, the CISL project team had an interest in knowing how others were solving the duplicate detection problem.

As it turns out, while the CISL project's problem was principally identification of duplicates so the information about them could be combined with other data or fed back to a user who might not have input a semantically correct request, the principal problem encountered in mailing list/participation list manipulation is identification of duplicate entries for purposes of purging duplicate records from such lists.

With this in mind the author undertook a project to investigate the state-of-the-art in duplicate detection as performed in the mailing list/participation list industry and review future technology in an attempt to predict the direction the industry will take in performing duplicate detection in the future. Hopefully, the knowledge of the CISL project team and others (e.g., Ducks Unlimited) will be furthered by this activity.

CHAPTER I - INTRODUCTION

This thesis documents research to investigate the state-of-the-art in duplicate detection as performed in the mailing list/participation list industry, documents review of future technology for its applicability to the industry and presents a prediction of the direction the industry will take in performing duplicate detection in the future.

The documentation of this activity is organized as follows:

A "Methodology" of how the activity was accomplished is contained in Chapter II.

Chapter III, "Semantic Confusion" contains a description of the problem of duplicate detection in general. This chapter presents a rather academic view of the problem.

Chapter IV presents an overview of "The Industry". This chapter describes some of the principal players and describes the size of the problem. Specific details of this general chapter for the principal players are reserved for the Appendices.

"Current Solutions" to the problem of duplicate detection as they exist today are presented in Chapter V. This is a composite of the findings of the industry survey of Chapter IV presented in a common and tutorial format.

"Futuristic Solutions" are presented and analyzed in Chapter VI.

Lastly, Chapter VII presents the conclusions of the analysis with a projection of where the industry will go in the future in terms of technology to solve the problem of duplicate detection.

CHAPTER II - METHODOLOGY

The methodology, or approach, to this research project was as follows:

A literature search on the subject of duplicate detection was performed. The trail of literature led to the Direct Marketing Industry as the principal user of the duplicate detection in their management of mailing and participation lists.

A number of the Direct Mailing Organizations were identified from the _Encyclopedia of Associations_ and contacted (e.g., Direct Marketing Association and the Association of Independent Mailing Equipment Dealers). Discussion with personnel from each of these organizations identified key contacts within the industry. One of the principal contacts identified was the United States Postal Service.

The United States Postal Service Customer Service Representative was contacted and provided a list of commercial firms "certified" as vendors of the Coding Accuracy Support System (CASS) and "licensed" to operate the National Change of Address System (NCOA). CASS is the Postal Service's system for certification of vendors for proper mailing list coding. NCOA is the Postal Service's system for changing mailing list addresses based on individual change of address notices. He also provided information from a number of the CASS certified and NCOA licensee companies which had visited with him.

Working a separate path through the technical applications side of the literature, the computer hardware, software and communications products and companies involved in duplicate detection were identified from Data Sources.

Additional contacts were established through the CISL project, contact with CitiBank and discussion with Professor Lovelock of the Marketing Department at the Sloan School.

After some preliminary contacts were made, it became apparent that there are two principal types of players in the manipulation of mailing and participation lists. They are the United States Postal Service and the software developers. The software developers can further be broken down into large entities including mailing list companies, mail order houses and organizations which develop their own software and manage their lists internally, and software developers who provide software and support services to users. Some software developers also were found to offer mailing list services. A review of the identified contacts was made and a number of the entities were selected for further contact and, if willing, study. Selection was made based on unique attributes identified during the research process, duplication of identification as a player from different research paths, known availability for further contact, size and the type of player.

Contacts were established via telephone and in most cases literature was provided by the organization. After review of the literature an interview was requested to answer questions not covered in the literature and get a feeling for the organization. A good feeling for the capability of each organization was established from this activity, as was the state-of-the-art within the industry.

Knowing the state-of-the-art within the industry, possibilities for improvements in the future were identified through review of literature, discussions with contacts and brainstorming with others in the CISL project. Each of the reasonable ideas for improvement were analyzed with an eye on the cost effectiveness of each.

Lastly, conclusions about where the industry might go in the future were drawn.

CHAPTER III - SEMANTIC CONFUSION

Semantic confusion results when deciding if two records are the same or different and against what criteria one wishes to determine them to be different. Since everyone does not have a single, unique name given and adhered to from birth and an address unique to himself/herself for all time, this problem is not solved by a trivial comparison to determine if an exact match exists. In the real world people have nicknames or change their name, people have their name and address misspelled, human and automated input devises make mistakes, numerous people of similar and differing name reside at the same address, and some people even maintain two or more addresses. Likewise, the person attempting to identify duplicates may not be interested in the duplicate residents at an address but just assurance that he only has each address recorded once in his database. Each of these problems is different and the costs/benefits of recognizing each of these also differ.

Fortunately, definition of the problem nicely sorts itself into a number of second and third order problems. Once these problems are solved the list manager can then apply logic to solve his specific problem. These problems are described below:

Attribute Naming - Attribute name problems occur when two like entities are confused because of differences in their entry. In general the average human can easily interpret the difference between such records and with varying certainty declare them duplicates. Attribute naming problems separate themselves into two subsets. They are entry spelling uncertainty/error and nicknaming/abbreviation error.

EXAMPLE III-1

Joan Smythe
123 Boothfield Road
Lacey Şprings, AL 35754

Jaon Smith
1234 Booth Field Road
Laceys Spring, AL 35754

Example III-1 presents five examples of spelling uncertainty/error. "Joan" is simply misspelled by the transposition of the "o" and "a". "Smythe" is confused with the homophone "Smith". One (or both) street address either has too many or too few digits. "Boothfield" and "Booth Field" are two different spellings for the same road. And lastly, two different colloquialisms for "Laceys Spring" are used. From this example one can easily see how error or misunderstanding can corrupt even a simple address. Yet, one can also see that an average human would declare these to be the same entry.

Example III-2 presents five examples of nickname/abbreviation error. "Robert" and his initial "S." have been replaced by the common nickname for Robert, "Bob". "One" has been replaced by its digital equivalent, "1". And, "Place", "Suite" and "Massachusetts" have been replaced by their respective abbreviations "Pl", "Ste" and "MA". While these two addresses have exactly the same meaning to a human a computer sees them entirely different.

EXAMPLE III-2

Robert S. Jones
Suite 2356
One Longfellow Place
Boston, Massachusetts 02114

Bob Jones
1 Longfellow Pl, Ste 2356
Boston, MA 02114

Inference Matching - For many list entries the name and address or supplementary data might be used to infer duplication. For instance in Example III-1 one might bias their decision about whether or not to declare a duplicate based on the address being rural. There may be a higher probability that people with common last names live in a small area. One might fix the city address based on the ZIP Code or vice versa.

EXAMPLE III-3

Billy Ray Inglis
Route 1, Box 356
Elora, TN 37328

Billy Ray Inglis
P.O. Box 104
Huntland, TN 37345

Example III-3 shows two addresses for the same individual. Noting the rural address one might infer that there was only one Billy Ray Inglis within the three digit ZIP Code zone "373".

Lastly, additional data might be used to determine a duplicate. In Example III-3 if both Billy Ray Inglis records showed a telephone number of (615) 469-7780, that would be a significant indicator that the records were for the same individual.

While all of the above are easily recognized as duplicates by the average human, an algorithm which eliminates all such duplicates with little error can quickly be seen to be very complex. Even the knowledge of how many people reside within a ZIP Code zone taxes the capability of human recall much less the problem of comparing literally millions of records to identify duplicates.

CHAPTER IV - THE INDUSTRY

This chapter presents an overview of the mailing list/participation list industry. The order of presentation of this information is as follows: Following a brief historical section a review of the United States Postal Service and its pivotal role in the industry is provided. Then a discussion of the other players and their role is wrapped together in a general discussion. This section segments the market, describes some of the principal players, discusses some of the service features and problems and alludes to the size of the problem in general. For specific details about various players in the industry the reader is directed to the appendices and the readings in the bibliography.

History

Not until the 1960's were mailing lists of significant size to be noteworthy known to exist. The organized list consisted of file cards and the operation of duplicate detection was performed by hand. "Address-O-Graphs" were a commonly used method of keeping records and printing them for organizational lists prior to that time. With the advent of practical and affordable computers in the mid to late 1960's the mailing list industry for direct marketing surged. Today a number of the leaders in the field's roots can be traced to entrepreneurs of that time.

The United States Postal Service

Because of their pivotal role in the handling of mail, the United States Postal Service is a major player in the Direct Marketing Industry. The Postal Service is very interested in the elimination of duplicate deliveries, the elimination of undeliverable mail, and pre-processing of the mail to facilitate delivery. The Postal Service is so interested in these that they offer financial incentives and service assistance to interested parties to obtain it. The financial incentives include a discount from the $0.25 per ounce or less per piece first class rate to $0.21 for pre-sorting by Five-Digit ZIP Code on down to $0.195 per piece for things like pre-sorting by Carrier Route, using ZIP+4 and Barcoding. Similarly, third class rates fall from $0.165 to $0.101 for profit and $0.084 to $0.053 for non-profit organizations. The services offered for free or at a nominal fee include conversion of lists to incorporate ZIP+4, address correction, and changes of address; evaluation of vendor services; and cross reference between ZIP+4 and Census Geographic Base File/Dual Independent Map Encoding Files to assist market researchers and demographers to relate ZIP+4 to Census Bureau demographic statistics. They also provide free bundling materials to assist in accomplishing bulk mailing activities.

While the Postal Service maintains a list of over 50 million address changes, they do not maintain a definitive list of all addresses. Hence, the Postal Service does not have a list of the occupants of or businesses at every address. They use bounded definitions of addresses to determine ZIP Codes (e.g., Even numbers 30 - 50 Memorial Drive, Cambridge, Massachusetts is ZIP+4 02142-1347. Hence, the Postal Service is not sure if 38 Memorial Drive exists until a carrier tries to deliver to that address.). This is further complicated by Carrier Routes not being assigned to consecutive ZIP+4 addresses.

It does, though, run a state-of-the-art operation using Optical Character Readers to interpret ZIP Codes and Bar Codes at rates as high as 36,000 letters an hour per machine and CD-ROMs for information retrieval.

Its Coding Accuracy Support System (CASS) which provides an evaluation of vendor services is a very useful service to the industry as a whole. Its principal purpose is to improve the quality of Five-Digit ZIP Code, ZIP+4 Code and Carrier Route Information System information. CASS employs two stages:

In Stage I, the Postal Service provides addresses written with correct codes on a computer tape. This information can be used internally by the customer to evaluate the accuracy of their code matching software which is either in-house or under consideration.

For Stage II, a test tape of addresses is supplied without correct code information. The service organization will then perform a list conversion using their software. That product will then be scored by the Postal Service for matching accuracy. Firms attaining the minimum acceptable score, 95 percent correct, are certified.

The Postal Service provides interested parties complete lists of names and addresses of these certified vendors. Certification is performed for a six-month period after which a firm must qualify again. While under constant revision (monthly) the CASS I tapes vary from 14,000 to 15,000 records for the three services and from 15,000 to 45,000 for CASS II. In addition to a raw score indicating what percentage of the records the vendor got correct, he is provided with feedback on the individual errors he got in evaluation of the CASS II tape. The CASS I and II tapes are almost letter perfect (i.e., The address components are all spelled consistent with the ZIP Code index). Because of this vendors often score 99 percent and above, mostly missing things resultant from erroneous input data (i.e., Misspelled addresses, addresses for which a ZIP Code is undefined or the ZIP Code index incorrect or inconsistent).

A second significant service managed by the Postal Service is the National Change of Address System. Under this activity the Postal Service licenses data processing organizations to use change of address information compiled and distributed solely to licensees by the Postal Service. The vendors use this information to standardize and change addresses on customer provided mailing lists for a nominal fee. Selection of licensees is competitive and based upon technical and management ability to meet the computer requirements, market the product and properly manage the service. There are 17 authorized licensees at this writing. The merge/purge software used for this task is recognized as the state-of-the-art in mailing list duplicate detection. While the detailed code implementation of the software is left to the vendor, the specific rules used to determine duplicates is closely controlled and checked by the Postal Service.

To supplement evaluation of these services the Postal Service has established the National Deliverability Index. The National Deliverability Index identifies and scores six factors deemed critical for optimum mail processing and delivery. These criteria provide valuable information concerning: Matching and standardization against the ZIP+4 File; Use of apartment numbers necessary for accurate delivery; Complete rural route and box number information; Use of correct Five-Digit ZIP Codes; National Change of Address up date frequency; and Removal of Moved, Left No Forwarding Address records. A prospective software purchaser or system manager should review the results on these activities and use them as a basis of performance and quality decisions.

In terms of duplicate detection the Postal Service offers three significant items:

First the Coding Accuracy Support System and National Deliverability Index offer a baseline for software development and evaluation. It is possible in the future these activities could grow to incorporate better evaluation of duplicate detection. They additionally could serve as a learning base for a neural network based duplicate detection system (See Chapter VI).

Second, the National Change of Address duplicate detection algorithms, further discussed under solutions, represent a good basis of the state-of-the-art in elements to consider in duplicate detection.

Lastly, the Postal Services standards for address writing, abbreviation and storage are de facto standards within the industry.

Other Players

Numerous industries are presented with the problem of detection of duplicate records on mailing/participation lists, but none more than the Direct Marketing Industry. Numerous segmentations of this industry can be made (e.g., profit/non-profit, mail order, subscriptions, etc.) but they all have one common thread, the maintenance of large quantities of name, address and other records. Even industries which do not consider themselves to be a part of this industry find themselves effectively members because of the quantity of records they keep (e.g., professional, hobby, fraternal, credit services, reservations). Most of these organizations realize the extent of their involvement and themselves sell to members/participants through use of their lists or sell the list or its use to others for this purpose. Because of the extent and competitiveness within this industry efforts to detect duplicates and perform what within the industry is referred to as "merge/purge" have become quite sophisticated.

One quickly can segment off the group of activities which develop duplicate detection software from the remainder of the industry. Review of the industry reveals that a further segmentation of the duplicate detection software developers into two groups is practical. These two groups are those which develop and use their software internally for their own exclusive use (even if that use is only to provide services) and those which develop and sell software to those who have use for it. As well as provide necessary service support to users of their software, many of the firms in the second category also have departments which provide list services.

A second segmentation of use in analyzing this industry is the size of the list targeted to be manipulated and the computer hardware to be used for this manipulation. This second segmentation is important because the number of records one can manage is controlled to a large extent by the size of the computer's storage and its operating speed. Hence, a small business or organization might be interested in and only able to afford a small personal computer based system while a large business, organization or service company could justify and afford a major system based on a mainframe computer. The difference in capability between the two types of systems leaves such a gap that most intermediate sized businesses or organizations use services. Additionally, the intermediate entity is often interested in growth. The services are his major source of addresses, hence his close relationship with them is enhanced by the service arrangement.

The 1989 edition of Data Sources contains listings for list management software targeting 17 different types of computers. For the IBM PC-MS/DOS alone there are 114 list management software suppliers offering 155 packages, only 14 of which claim to have some duplicate detection capability. For IBM-mainframe computers there are 12 list management software suppliers, half of which claim to offer duplicate detection. IBM computers were found to dominate the literature and were the only computer vendor claimed by the companies contacted.

Table IV-1 lists the vendors contacted segmented by size of computer they use or support and market segment they target. In addition to identifying algorithms which are discussed in Chapter V, the survey of the vendors revealed many interesting facts. The relevant information is summarized in the following paragraphs.

TABLE IV-1 Vendors Contacted

Company	Target System/User	
Type	Mainframe	Personal
Services	Acxiom	
	Creative Automation	
	Epsilon	
	First Data Resources	
	Group 1 *	
	Harte Hanks	
	LPC **	
	Wiland	
Software	CMD	Group 1 *
	Group 1 *	Flowsoft
	LPC **	

* Group 1 is only company to service all markets.
** LPC provides both services and software.

The mainframe computer operations are used primarily with large databases and batch operations. Many of the systems in place literally have every address and each individual at that address identified. While some of the application programs were found to be written in Assembly Language or C for stated reasons of efficiency, a surprising number were found to be written in COBOL. It is suspected but not confirmed that the COBOL algorithms are less complex than the others. Though with the high speed computer equipment of today, machine speed seems to allow the implementation of reasonable algorithms in a Higher Order Language. Since no correlation between the age of the programs or company and the use of machine language or COBOL could be found, a suspicion that the COBOL based programs represented early (i.e., 1960 and 1970) implementations appears incorrect. A clear trade between efficiency and complexity of programming and maintenance is being made.

The claims for number of records processed ranged as high as 3 million records per hour in a 30 million record database. To give an idea of the size of some databases and the complexity of the problem, when TRW, the United States' largest credit reporting service maintaining credit records on over 145 million individuals, acquired Executive Services and entered the Direct Marketing Industry they used a second party's software to merge the over 490 million consumer records of Executive Services down to 138 million records. The process took five days and many passes. These numbers correspond to a reduced record for over 75 percent of all people over the age of 18 in the United States. And, it also indicates the size of the problem in that 490 million records, over 2.5 records per person over the age of 18, existed in the database TRW acquired.

Another indicator of the size of the industry is that the major software vendors have subroutines to not only print labels or make labels for mail bundles but go on to producing labels and packing lists for pallets and truck loads of mail.

Only about half of the companies contacted concerned themselves with International mailings and most of those were confined to Canada. It would be expected that different addressing rules would come into play in the international market (e.g., EZ-6 Canada's ZIP Code equivalent clearly is different), but it is interesting to note that even with the internationality of names special rules to handle name combinations and alternative English spellings are used on Canadian duplicate detection routines. Other countries covered included the United Kingdom and West Germany. The literature on foreign applications may be limited by language barriers for clearly other developed countries such as France have sophisticated postal systems and Direct Marketing Industries.

Moving on to the Personal Computer applications, while many individuals use Personal Computers to manage organizational lists few of these applications require any sophisticated duplicate detection capability. Yet they advertise the use of Soundex algorithms with Match Coding as an option. The two systems reviewed offer on-line duplicate detection (i.e., the operator can input a file and its duplicate, hopefully only one, is identified and returned). It is important to note that human intervention is noted not because of the ability to operate better than the machine. It is mentioned as an added feature for dealing with customers. In some ways humans are better than the machine at identifying duplicates (e.g., spelling errors) but in others no where near as good or efficient (e.g., ZIP Code correction). The systems require between 512k and 640k of memory. Literature indicates using an AT machine and hard disk 20 million names can theoretically be managed but 50,000 names is all that can efficiently be managed.

Before finishing this chapter some discussion of the uses of duplicate detection, performance and definition of jargon is in order. The next chapter contains a number of algorithms with examples of their operation. These examples come from the open literature (i.e., literature available to the general public though often specifically prepared by the vendor to describe his system). Most every competitor spoken with felt he had proprietary capabilities in his software yet none even hinted of the use of any advanced techniques beyond those presented.

The specific algorithm used depends on the application at hand. For example, a mail order business may be interested in elimination of duplicate addresses from a mailing list which consists of a list of current customers believed not to contain duplicates and a list of members of an organization active in his business (e.g., a mail order hobby supply house merging his customer list with the membership list of the Academy of Model Aeronautics). As he merges the two files he must check for duplicates between the lists and duplicates on the organization's list because it is common for multiple members of a family to belong to the same organization. He further might want to identify the matches, and look at response rates to advertisements offered to members through the organization or by members of the organization and mail to the family and not individuals when forwarding to a multimember address.

Going back to the example, the organization in its regular membership mailings might want to reduce its expenses by only mailing one publication to each family. But since each member is entitled to a publication it might be better off identifying duplicate addresses and sending a response card asking if it could do this. For fund raising non-profits such a technique has other attractiveness because it makes the organization look efficient while reminding the family of its existence.

Later, the master list must be used again to solicit renewals, in this case duplicates are not of interest unless they have not been checked for during membership entry. An untapped possibility is the bundling of renewals to an address in one envelope. Hence the organization would be printing and mailing envelopes containing one, two, three, etc. renewal applications.

Lastly, suppose the organization wants to solicit contributions. it may want to mail only one solicitation to each address but may want to target a different mailing to addresses having multiple members (e.g., Introduce paragraphs with "your family" verses "you").

The design of the duplicate detection algorithm in each application is different and the penalty for error and payoff for success in each application is different. The point of all of this discussion is that one person's duplicate is not necessarily the same as another's. The benefit and expense of each type of error, Type 1 where a record that should have been declared a duplicate was not and Type 2 error where a record that should not have been declared a duplicate, are different. Nothing is more clear than in the banking industry where there is very little margin for error when distributing money and credit and a lot of margin for error in soliciting new accounts and deposits.

There are a number of difficulties in rating the performance of duplicate detection systems. First definition of the application to be solved must be clearly identified. Then a large, representative sample set needs to be defined and the system allowed to operate on it. Lastly, the resultant data set needs to be evaluated. Since there are no standards other than the Postal Service and, since to be meaningful sample sets must be large, evaluation is very difficult. The Postal Service rates many of the software packages using relatively large, pristine data sets and evaluates the better vendors performance at over 99 percent. Other numbers thrown about the industry for non-pristine files range from 90 to 97 percent for consumer mail to 40 to 80 percent for business mail. Business mail is significantly more difficult because of two factors. These arise principally because business mail consists of four or more lines by comparison to the consumer address that consists of three lines. The extra lines are used to identify individuals at the company and/or assign a title for the individual. The first factor is that because of there being more data there just is more chance for mismatch and hence not declare a duplicate (Type 1 error). The second is that even though the Postal Service requests titles not be used (Note how confusing titles can be. V.P., Vice Pres, Vice President, Executive Vice President, etc. can all be the same.) and the name always appear at the top this often is confused furthering the complexity of the task.

The Postal Service has defined a conservative set of duplicate detection rules for its National Change of Address activity (See Chapter V). It requires strict compliance by the licensees in implementing these rules. In addition to strictly implementing the Postal Services National Change of Address rules most licensees offer "Nixie" service. Nixie service is the individual licensee's application of his own rules and database matches to identify addresses highly suspected to be erroneous. The Postal Service allows this to encourage the extraction of non-deliverable mail before it is created.

The last point to be made is that beyond duplicate detection many other factors are used to purge lists to compile the final list for a mailing. There are very complex logical processes which are used to target mailings to specific market segments. An incredible amount of secondary information has been compiled on each individual and address by some services. This data is cross referenced, etc. to determine customer prospects. While many vendors offer this type of service, none were noted to be using it to identify duplicates. Most notably among the vendors performing these customer matching services was a company named Persoft, Inc. which had client claims that its expert system had successfully been used to reduce selected customers for solicitations by 50 percent while maintaining over all response at 80 percent with file sizes of over a million records.

CHAPTER V - CURRENT SOLUTIONS

This chapter presents the details of the findings of the industry survey presented in Chapter IV. It has been separated into three sections. The first section discusses algorithms or algorithm components which are in use for detecting duplicates. Following that is a discussion of how these algorithms are logically applied to accomplish specific tasks. Lastly, some commentary is presented describing some general observations such as algorithms which have been considered and are not used.

General Algorithms

A number of generic algorithms for matching list entries are presented in the following section. In some cases the algorithm itself is applicable to an entire list entry (e.g., Match Codes). In other cases the algorithm is only applicable for application to a line, token or sub-string of a list entry (e.g., Soundex). According to a United States Postal Service representative the algorithms discussed represent the algorithms used in over 98 percent of the mailing list software in use today. The trivial concept of exact string matching will be ignored though efficient implementation of exact string matching algorithms into any duplicate detection system is mandatory.

Before discussing any algorithms, an important concept, the concept of approximate matching, needs to be introduced. That concept is differentiating between exact matching (the trivial case), passing a matching algorithm (i.e., exactly matching after application of a rule) and receiving a score on a matching algorithm. Some algorithms afford themselves to immediate pass/fail criteria. Others allow for a scoring. For example, an algorithm might drop all vowels from a string and exactly compare the result giving a pass or fail output. An equally valid algorithm might indicate the percent of the characters of a string which match by location. The importance of this second type of algorithm is that pass/fail criteria can be adjusted to the application (e.g., in the simple example 3 of 5 or 60 percent). On the other hand when algorithms are combined to form a rule the pass/fail algorithm might be used as part of a logical or weighting function depending on the application. As the algorithms are discussed it will be clear to which category they belong.

Standardization - Unfortunately there are almost as many standards for storing list entries as there are lists. While the United States Postal Service has a best practice standard it will accept mail marked in many ways. The Postal Service's standard calls for all alphabetical characters to be in upper case. Even at its best mailing rate the Postal Service will accept non-standard addressing if Barcoding and Carrier Route Sorting are provided. Even when mail is to be addressed with non-standard formats, the manipulation of the entries seems always to be done in upper case (i.e., All alphabetical characters of the entry are converted to upper case). From this point on all examples will assume the use of all upper case characters.

A second standardization which is common is to convert all address entries to the standard abbreviations in Table V-1. Note that even this system is not without problems. For example, no abbreviation is defined for "Saint", while the two common abbreviations for "Saint", "ST" and "STE", are reserved for "Street" and "Suite", respectively. Also, the common abbreviation for "Place", "PL", is not defined. Other problems arise when name, address and city tokens consist of these reserved words. Example V-1 contains a smattering of these problems. Contextual rules (explained in the following section) which sort and identify proper abbreviations are used to solve these non-trivial problems. As with the use of upper case characters, these standard abbreviations will be used in all examples from this point forward.

EXAMPLE V-1

LANE WEST
5353 W NORTH ST
SAULT SAINTE MARIE MI 49785

TABLE V-1 USPS Standard Abbreviations (Other than States)

Apartment	APT	Expressway	EXPY	Room	RM
Avenue	AVE	Freeway	FWY	Square	SQ
Boulevard	BLVD	Lane	LN	Street	ST
Circle	CIR	Parkway	PKY	Suite	STE
Court	CT	Road	RD	Turnpike	TPKE
North	N	West	W	Southwest	SW
East	E	Northeast	NE	Northwest	NW
South	S	Southeast	SE		

The third standard is that the address record consists of no more than four lines with the top line containing the attention or person to receive the piece, the second line, which is optional, containing the company name, the next line containing the complete street or box address including the apartment or suite number and the last line containing the city, state abbreviation and ZIP Code.

The fourth standard of the format is that no punctuation is used except a dash, "-", between the fifth and sixth digits of the ZIP Code. This dash is not required when only a Five-Digit ZIP Code is used.

And, lastly, a single space is used as the delimiter.

Context - Though somewhat trivial the context of a tokens occurrence requires review before abbreviation or other rules are applied. Example V-2 illustrates how context might be confusing. In general, abbreviation other than reduction of a first and middle name to initials should not be performed on the first two lines of a record (i.e., name and optional company line) though when dealing with companies reduction of "and" to "&", all forms of "Incorporation" to "Inc.", etc. have proven to be helpful. Prefix titles such as "Mr.", "Ms.", "Dr." and suffixes such as "II', "III", "Jr." and "Esq" for names, and directions such as "NW", "NE", etc. for street addresses must be considered.

EXAMPLE V-2 Context Differences

```
     NW STREET                NORMAN W STREET
     375 N PARK WAY           375 NORTH PARKWAY
     KANSAS CITY MO 64120     KC MO 64120
```

ZIP Codes - The United States Postal Service's 1963 addition of the Zone Improvement Plan (ZIP) Codes to addresses greatly facilitated their delivery service. Later, to further facilitate their service they stretched the Five-Digit ZIP Code to nine digits (ZIP+4). In addition to receiving discounts for the use of ZIP Codes in bulk mailing, there are other great advantages to using ZIP Codes. ZIP Codes set the standard for address identification and probably offer the best single segmentation key available to the duplicate searcher. Automated routines which add and correct ZIP Codes are not trivial. To understand the problem one must first understand more about ZIP Code assignments.

Every address is assigned a ZIP Code based on its location. The nation is segmented into 10 National Areas. The first digit of the ZIP Code identifies this National Area assignment. The next two digits are assigned based on a Sectional Center Facility or Large Post Office based on population density within an area. The next two digits specify the Post Office, Delivery Area or Delivery Office.

For example, in a metropolitan setting, a "0" first digit indicates the post office is in the Northeastern United States (New England), adding "21" indicates the address is in the Boston Metropolitan area served by the Regional Post Office in Boston. The next two digits being "14" (e.g. 02114) indicate the address is serviced by the Charles Street Station.

In a rural setting, for example, a "3" first digit indicates the Southeastern United States, a "59" in the next two places indicates an address serviced by the Sectional Center Facility in Gadsden, Alabama. Lastly, "76" for the next two digits indicate service by a Post Office in Guntersville, Alabama.

A subtle difference here is that some addresses are handled by Post Offices, some by Postal Stations and some by Post Office Branches. To the customer all these locations look like a Post Office because they are marked as such. But, the assignment of ZIP Code varies from one to the other. As will be seen when the last four digits are assigned in the full ZIP+4 assignment, keeping track of the Five-Digit ZIP Code is more important than the city name.

Each Post Office, Delivery Area or Delivery Office delivery area is further separated into Sectors designated by the next two digits and Segments designated by the last two digits of the ZIP Code. A segment can be as small as an individual Post Office Box, mailbox or mail drop within a company but usually includes all addresses along one side of a city block or a range of floors in an multistory building.

One of the problems that arises in trying to assign ZIP Codes is illustrated by the data in Table V-2. If one were attempting to send a letter to an address of "30 Cambridge Street, Cambridge, MA" they would use a ZIP Code of "02141" (or "02141-1815"). They would not find this address indexed under Cambridge but as a Delivery Office in the Boston Region. Hence, an equally acceptable address would be "30 Cambridge Street, Boston, MA 02141". This is because "Boston" defines a Postal Region as well as a City.

On the other hand for an address of "30 Cambridge Street, Boston, MA" one would be unsure which of the five ZIP Codes defined in the table was correct since for an address of Boston with an even address below 40 on Cambridge Street any of the five Five-Digit ZIP Codes in the table are acceptable. Hence, while addressing to

Boston is acceptable without the proper Five-Digit ZIP Code the mail will possibly not go to the right address. In fact, there are 81 acceptable Five-Digit ZIP Codes for Boston. And, for five of these there is a possibility of an assignment to 30 Cambridge Street.

TABLE V-2 Cambridge Street ZIP Codes in Boston 021XX

Cambridge Street Address	ZIP Code	Delivery Office
Even 0-40	02129-1302	Charlestown
Even 0-66	02141-1815	Cambridge C
Even 0-98	02114-2909	Charles Street
Even 0-98	02151-5211	Revere
Even 0-98	02178-1301	Belmont

Thus if the piece were addressed to Cambridge it is understood to be a Boston delivery in a Delivery Area serviced by the Cambridge Station and omission of the fourth and fifth digits of the ZIP Code, or even the entire ZIP Code for that matter, could be corrected.

The ability to assign the correct ZIP Code, at least the Five-Digit ZIP Code, is very important to the ability to detect duplicates. The data on ZIP Codes is available from the Postal Service in printed form, on magnetic tape and on CD-ROM. Selection of the ZIP Code allows the searcher to segment the data into a reasonable geographic area and indicates with a high degree of assurance that a properly spelled existing street has been defined.

Match Codes - The concept of Match Coding is to assign a code which should be unique to each address based on its attributes. Later this code which is much simpler than the list entry is compared to candidate entries and if an exact duplicate is found a duplicate is declared. (Note that while exactly matching Match Codes are usually required to declare a duplicate there would be nothing prohibiting almost matching Match Codes to be declared as duplicates. Such an action would largely negate the usefulness of the concept and in most instances implementation of a simpler Match Code would probably be more effective.)

Example V-3 shows an example Match Code and its application to a list entry. Example V-4 shows the same match code applied to two similar list entries. Note that the Match Code for each entry would remain constant for that entry and could be stored with the entry to reduce future computation. Also, notice that in one case the match code would have declared the entries to be the same and in the other to be different while most humans would have declared both entries to obviously be the same.

The Match Code used in the example is used only for illustrative purposes and any code of the type shown will work, but some work better than others. And, some codes that work well for one application work poorly for another and vice versa. Because of this most software vendors and developers make proprietary claim to their particular Match Code for various applications. On the other hand, one can often find an easily decipherable printout of a business' Match Code along the border of a mailing label clearly showing a lack of concern about revealing it.

EXAMPLE V-3 Sample Match Code

Address	Match Code
FRANCES S IANACONE 3535 S WAKEFIELD ST ARLINGTON VA 22206	22206INA535WAKEF

In this example the match code consists of the string composed from the Five-Digit ZIP Code; first third and fourth letters of the last name; last 3 digits of the address; first four letters of the street address and first letter of the first name.

EXAMPLE V-4 Sample Match Code Application

Address	Match Code	Remarks
FRANCES S IANACONE 3535 S WAKEFIELD ST ARLINGTON VA 22206	22206INA535WAKEF	Candidate
FRANCIS S INNACONE 3535 S WAKFIELD ST ARLINGTON VA 22206	22206INA535WAKFF X	FAILS
FRAN INNACONN 3535 WAKEFIELD AVE ARLINGTON VA 22206	22206INA535WAKEF	PASSES

X - Does not match

Match Codes are not entirely different from the development of a Hashing function except that in a Hashing function the programmer is establishing a memory location for storage and not maintaining the contextual usefulness of the data. In its simplest form a Hashing function storage system that throws away collisions would be operating as a duplicate eliminating Match Code.

The Match Code is an elegantly simple, yet powerful technique for matching duplicates. One of the major difficulties with the use of Match Codes is that while they assure an exact match of selected attributes, they do not assure an absolute match of the entries and have a large margin for error both in missing a desired approximately equal entry and matching a dissimilar entry. Match Codes do work well when small entry lists are used and the list entry is performed to a uniform standard. Probably the major advantage of Match Codes is that because of their simplicity in handling an entire address entry at once and ability to be stored with the data, hence, not requiring processing each time records are compared, they execute very rapidly on the computer. This explains why many early and personal computer duplicate detection systems relied on Match Codes. Match Codes have even proven to be of considerable use in manual applications.

Attribute Matching - There are a wide variety of attribute matching algorithms. Principal among these are those which transform synonyms, homophones, etc. into consistent canonical forms. Example V-5 presents an example of both synonym and homophone differences in two similar records. In this case the nickname (synonym) for "Robert", "Bob", appears along with the homophones "Mohr" and "Moore". The address number two hundred two, or "twenty two" is confused with "twenty-two". And, "Greenewell" has received two different spellings.

EXAMPLE V-5 Synonym and Homophone Confusion

BOB MOHR
202 GREENEWELL STREET
SUMMERLAND CA 93067

ROBERT MOORE
22 GREENWELL AVE
SUMMERLAND CA 93067

Handling the synonym is among the most difficult problems in duplicate detection. The only known method is to compile from experience a list of common synonyms and convert all records containing them to a common canonical form. This can be a time consuming activity without explicitly known benefit. Another problem with this is that transformations are not always transitive. For example, one might agree that when one compares a "R" to "Robert" a match should be declared. This situation is likewise probably reversible (i.e., "Robert" is a good match to a "R"). On the other hand, while "Bob" and "Robert" may be good matches "Bob" and "R" may not. There certainly is more opportunity for error in the second situation. Hence, algorithms must address each instance of the name separately. Because of this complexity and learned effectiveness, the various software developers claim these lists to be proprietary. On the other hand the Postal Service has established a standard list for NCOA use.

EXAMPLE V-6 Soundex Algorithm

GALE DIXON
123 RIDGE ROAD
TWIN BRIDGES MT 59754

GAIL DICKSON
123 BRIDGE RD
TIN BRIDGE MT 59754

Name:	Left Hand Entry		Right Hand Entry		Action
Step 0	GALE	DIXON	GAIL	DICKSON	None
Step 1	G040	D0205	G004	D022205	Conversion
Step 2	G4	D25	G4	D2225	Drop "O"'s
Step 3	G4	D25	G4	D25	Drop Runs
Step 4	G4	D25	G4	D25	Truncate
	Pass	Pass	Pass	Pass	Pass

Address:	Left Hand Entry		Right Hand Entry		Action
Step 0	RIDGE	ROAD	BRIDGE	RD	Drop #
Step 1	R0320	R003	B60320	R3	Conversion
Step 2	R32	R3	B632	R3	Drop "O"'s
Step 3	R32	R3	B632	R3	Drop Runs
Step 4	R32	R3	B632	R3	Truncate
	Fail	Pass	Fail	Pass	Fail

City:	Left Hand Entry		Right Hand Entry		Action
Step 0	TWIN	BRIDGES	TIN	BRIDGE	Drop ZIP
Step 1	T005	B603202	T05	B60320	Conversion
Step 2	T5	B6322	T5	B632	Drop "O"'s
Step 3	T5	B632	T5	B632	Drop Runs
Step 4	T5	B632	T5	B632	Truncate
	Pass	Pass	Pass	Pass	Pass

The first character of the name is reserved and taken as the first character of the test string (The observation that the first letter of a word is usually not incorrect is an important assumption about this algorithm.). Thereafter numbers are assigned to the letters according to the following table:

#	Letter	#	Letter
0	A E I O U H W Y	1	B F P V
2	C G J K Q S X Z	3	D T
4	L	5	M N
6	R		

Next all "O"'s are removed from the string, then runs (consecutive occurrences of the same number) are reduced to a single digit. Finally the string is reduced to four characters, the first letter and up to three digits. The resultant strings are then compared.

One method of handling the homophone is exactly the same as the synonym (i.e., compile an acceptable list of matching forms). While this is done by many it is more often performed through use of the Soundex algorithm or a variation of it. The Soundex algorithm, based on the 1918 and 1922 Patents of Russell for a manual filing system, is exemplified in Example V-6. Basically the algorithm converts the characters and syllables of a word that are phonetically similar into common characters and then collapses the word to emphasize the more phonetically distinct portions. In the example the name and city would be found to match while the address would not. The same result would have emerged if each of the strings had not been parsed and the operation performed on the entire string assigning "0" to blanks, " ".

One of the shortcomings of the Soundex algorithm is the assumption that the first letter of the word is correct. While this is not a bad assumption it is responsible for many errors when names like "Tchaikovsky" or words like "Pneumatic" occur.

Over the years a number of people have devised similar algorithms which reduced words into abbreviations for comparison. Another strong attribute of the Soundex algorithm is that it fixes many transpositions.

Approximate matching - Approximate string matching algorithms have matured significantly over the last decade because of their application to spell checking in word processing. Many of these algorithms perform first step comparative checks which verify a high likelihood of misspelling.

In English only 66 percent of the possible two character combinations exist in words. This drops to twenty percent for three characters, two percent for four characters and less than one percent for five characters. Hence, one quick way to identify a misspelling, not correct it, is to check the string and all possible sub-strings to see if they contain an illegal combination. Unfortunately, because of the internationality and colloquialisms used in names these rules do not work well with names and work only slightly better for addresses, though it might be possible to find a set of combinations applicable to names and addresses.

A second form of algorithm is one that interrogates the string for possible errors by fixing or detecting the error and comparing it to possible solutions. Since 80 percent of typing mistakes are single character omissions, insertions and substitutions or adjacent character transpositions and errors from other input means (e.g., Optical Character Scanners) are largely substitutions, deletions and insertions (Not transpositions) similar but different algorithms are

---

EXAMPLE V-7 Single Error Scoring

1) Wrong Letter (Simple exact match comparison).

Candidate	Comparison	Score	Comparison	Score
	HODGE		RODGERS	
ROGERS	X XXXX	5	XXXXX	5

2) Additional Letter (Canidate has extra character?).

Candidate_____	Comparison	Score	Comparison	Score
	HODGE		RODGERS	
OGERS	XXXXX	5+1=6	XXXXXXX	7+1=8
RGERS	XXXXX	5+1=6	XXXXXX	6+1=7
ROERS	X XXX	4+1=5	XXXXX	5+1=6
ROGRS	X XXX	4+1=5	XXXXX	5+1=6
ROGES	X XXX	4+1=5	XX XX	4+1=5
ROGER	X XXX	4+1=5	XX XX	4+1=5

3) Transposition (Canidate contains transposition?).

Candidate_____	Comparison	Score	Comparison	Score
	HODGE		RODGERS	
ORGERS	XXXXXX	6+1=7	XXXXXXX	7+1=8
RGOERS	XXXXXX	6+1=7	XXXXXX	6+1=7
ROEGRS	X X XX	4+1=5	X XXX	4+1=5
ROGRES	X XX X	4+1=5	XX XX	4+1=5
ROGESR	X XXXX	5+1=6	XXX X	4+1=5

4) Left Out Letter (Canidate missing character?).

Candidate____	Comparison	Score	Comparison	Score
	HODGE		RODGERS	
_ROGERS	XX XX	4+1=5	XX	2+1=3
R_OGERS	X X XX	4+1=5	X	1+1=2
RO_GERS	X XX	3+1=4		0+1=1
ROG_ERS	X X XX	4+1=5	X	1+1=2
ROGE_RS	X XX XX	5+1=6	XX	2+1=3
ROGER_S	X XXXXX	6+1=7	XXX	3+1=4
ROGERS_	X XXXXX	6+1=7	XXXX	4+1=5

This example demonstrates the scoring of a selected set of one character substitution, omission and insertion and adjacent character transposition scoring tests for the candidate "ROGERS" against the names "HODGE" and "RODGERS". All possible single corrections are attempted. One point is scored for each "X", "+1" or underlined character. An "X" indicates a mismatch between characters. A "+1" indicates a correction was used. And, an underlined character indicates an inserted character allowed to match. Since the lowest score is selected, in this case the score would be four for "HODGE" and one for "RODGERS".

---

useful in ferreting out these mistakes.

One of the difficulties in evaluation using such an algorithm is the many possibilities available for substitution. To perform every possible combination with every possible match to identify a match is just time consuming. A method is demonstrated in Example V-7. In applying this algorithm each record is scored against the candidate. A low score is indicative of a good match hence once a predetermined threshold or previously lower outcome is passed comparison can be shifted to the next record. Each record only requires one pass by all the other records. In general the solution to this minimization problem can be solved using matrix algebra.

Algorithm Application

In this section the application of the general algorithms presented in the previous section is presented. As noted in the previous section some of the applications are trivial because the algorithm uses the entire record (e.g., Match Codes). In other cases the algorithm results from an execution of a number of the aforementioned algorithms. In almost all cases some pre-processing of the list entry is required to, for example, correct ZIP Code or bring the address to an appropriate canonical form.

Before one can begin to put together a set of rules or a software package to detect duplicate entries they must decide what they mean by duplicate. In some applications duplicate addresses are what is being attempted to be detected and eliminated to prevent duplicate mailings. In other cases duplicate names are what is trying to be detected to eliminate duplicate memberships or credit files. There are literally hundreds of applications one can consider. With every application both Type 1, where a record that should have been declared a duplicate was not, and Type 2, where a record that should not have been declared a duplicate was, errors occur. Hence, one must, in addition to defining their definition of duplicate, assess the effect of each type of failure and act accordingly. Because of the cost of failure of most high value automated transactions that rely on these techniques (e.g., Teletype money transfers between banks), their outputs are reviewed by a human prior to execution.

Table V-3 presents eight generic rule definitions for three line addresses. This set of definitions, used for illustrative purposes, is not comprehensive because the set of possible rules is nearly infinite. The set of possible rules grows exponentially when one adds the possibility of additional lines to an address to include, for example, company name. A second complexity in the definition of rules is found in defining the severity of discrepancy that will be accepted. Even bounding the rules into a workable set for most applications is difficult though a few service companies and most commercial software vendors do list a set of about 16 standard matching rules which apply their internal algorithms. These internal algorithms determine the severity of the approximate matching test (i.e., How big a discrepancy will be allowed and a match still declared).

Table V-3 Generic Duplicate Rules

Rule	Definition/(Comment)/Example	
A	Same Person - Same Address (No Contradictory Discrepancies)	
	JOHN SMITH 1234 BROADWAY AVE CHICAGO IL 60610	JOHN SMITH 1234 BROADWAY AVE CHICAGO IL 60610
B	Same Person - Same Address (Minor Name Discrepancy)	
	JOHN SMITH 1234 BROADWAY AVE CHICAGO IL 60610	J T SMITH 1234 BROADWAY AVE CHICAGO IL 60610
C	Same Person - Same Address (Address Discrepancy)	
	JOHN SMITH 1234 BROADWAY AVE CHICAGO IL 60610	JOHN SMITH 124 BROADWY AVE CHICAGO IL 60610
D	Same Person - Same Address (Apartment Discrepancy)	
	JOHN SMITH 1234 BROADWAY AVE, APT A CHICAGO IL 60610	JOHN SMITH 1234 BROADWAY AVE APT A1 CHICAGO IL 60610

TABLE V-3 Generic Duplicate Rules (Continued)

E Same Person - Same Address
(Discrepancy in Both Name and Address)

JOHN SMITH 1234 BROADWAY AVE CHICAGO IL 60610	J T SMITH 124 BROADWY AVE CHICAGO IL 60610

F Same Surname - Same Address
(Different Person - No Discrepancy Other Than Given Name (and/or Prefix))

JOHN SMITH 1234 BROADWAY AVE CHICAGO IL 60610	MARY SMITH 1234 BROADWAY AVE CHICAGO IL 60610

G Same Surname - Same Address
(Different Person - Discrepancy In Given Name (and/or Prefix), Address, and/or Surname)

JOHN SMITH 1234 BROADWAY AVE CHICAGO IL 60610	MARY SMYTHE 124 BROADWY AVE CHICAGO IL 60610

H Same Address - Different Surname
(Different Person)

JOHN SMITH 1234 BROADWAY AVE CHICAGO IL 60610	TOM JOHNSON 1234 BROADWAY AVE CHICAGO IL 60610

With a little thought most duplicate definitions can be handled by the logical application of these rules. For example, Rule C might be a better one to implement than Rule H when sorting a membership list for duplicates because quite often multiple members of a family will belong to the same organization. On the other hand Rule H might be a better choice for mailing a solicitation for a record club in a neighborhood predominantly composed of college students because individual students are being targeted and multiple students may occupy the same address due to apartment sharing.

Now consider how to implement a rule using the generic algorithms provided in the previous section. For this example Rule E has been chosen. Again there is no single right solution.

Figure V-1 illustrates the rules operation. The rule operates in two steps.

FIGURE V-1 Flow of Rule E's Implementation

```
+--------------------------------------------------------+
¦ STEP 1                                                 ¦
¦  +-------------+     +-------------+     +------------+¦
¦  ¦    Input    ¦     ¦Convert Alpha¦     ¦  Eliminate ¦¦
¦  ¦   Record    +---->¦     to      +---->¦ Punctuation¦¦
¦  ¦             ¦     ¦ Upper Case  ¦     ¦            ¦¦
¦  +-------------+     +-------------+     +-----+------+¦
¦                                                ¦       ¦
¦                                                ¦       ¦
¦  +-------------+     +-------------+           ¦       ¦
¦  ¦   Soundex   ¦<----+    Parse    +<----------+       ¦
¦  ¦ Each Token  ¦     ¦   Record    ¦                   ¦
¦  +------+------+     +-------------+                   ¦
¦         V                                              ¦
¦  +------+---------------------------------------------+¦
¦  ¦ By Context Compare to Standard Abbreviations       ¦¦
¦  ¦    -Exact Match to Abbreviation? --> Continue      ¦¦
¦  ¦    -Exact Match to Abbreviation Word? --> Change   ¦¦
¦  ¦    -Soundex Match to Abbreviation Word? --> Change ¦¦
¦  +------+---------------------------------------------+¦
¦         V                                              ¦
¦  +------+--------------------------+   +--------------+¦
¦  ¦ ZIP Code Correction             ¦   ¦              ¦¦
¦  ¦    -Exact Match City/Street?    +-->+  Sort/Store  ¦¦
¦  ¦    -Soundex Match City/Street?  ¦   ¦              ¦¦
¦  +---------------------------------+   +--------------+¦
+--------------------------------------------------------+
```

First, Step One which would be the same for any rule consists of standardization of the candidate input. The Alphabetical characters of the entry are converted into upper case characters. Then all punctuation is eliminated. Next the record is parsed into the tokens delimited by the remaining spaces and the lines of the entry. Each Alphabetical token is converted by the Soundex algorithm and said conversion along with the token are stored for comparison. Then appropriately placed tokens by context are exactly checked for a match to the standard abbreviations, and if not matching they are

FIGURE V-1 Flow of Rule E's Implementation (Continued)

```
+--------------------------------------------------------+
¦ STEP 2                                                 ¦
¦  +-------------------+          +--------------------+ ¦
¦  ¦  Next Record on   ¦          ¦      Candidate     ¦ ¦
¦  ¦    Master List    ¦    +-----+       Input        ¦ ¦
¦  ++------+-----------+    ¦     +--------------------+ ¦
¦   ^      +----------------+                            ¦
¦   ¦      v                                             ¦
¦   ¦     / \               / \               / \        ¦
¦   ¦   /Exact\           /Same \           /Same \      ¦
¦   ¦ / Same 3- \  Yes  /  Name   \  Yes  / Address \ Yes¦
¦   ¦ \Digit ZIP/------>\ Within  /------>\ Within  /--+ ¦
¦   ¦   \Code?/           \One? /           \One? /    ¦ ¦
¦   ¦     \ /               \ /               \ /      ¦ ¦
¦   ¦      ¦ No              ¦ No              ¦ No    ¦ ¦
¦   ¦      +---+-------------+-----------------+       ¦ ¦
¦   ¦          v                                       ¦ ¦
¦   ¦         / \                                      ¦ ¦
¦   ¦       /More \                         +----------+ ¦
¦   ¦ Yes / Records \                       ¦            ¦
¦   +-----\on Master/                       ¦            ¦
¦           \List?/                         ¦            ¦
¦             \ /                           ¦            ¦
¦              ¦                            ¦            ¦
¦              v                            v            ¦
¦  +-----------+------------+   +-----------+-----------+¦
¦  ¦     Not Duplicate      ¦   ¦       Duplicate       ¦¦
¦  ¦ ---------------------- ¦   ¦ --------------------- ¦¦
¦  ¦       Add Entry &      ¦   ¦      Merge Data       ¦¦
¦  ¦       Store Data       ¦   ¦  With Existing Entry  ¦¦
¦  +------------------------+   +-----------------------+¦
+--------------------------------------------------------+
```

compared using the Soundex data to the source word for each standard abbreviation and then each abbreviation. When matches occur the abbreviation is substituted for the entry unless it is already in that form. The list is then run through a Zip Code correction and completion routine using first exact matches and then the Soundex data as a basis for street and city name matches. All records are sorted into three digit ZIP Code sets and this new standardized record, consisting of a standardized record and its Soundex data is then stored for future use.

In Step 2 each candidate input record is compared with each record on the master list until either a duplicate is found or the record declared not to be a duplicate and added to the list. A master list would be established by starting with one record, comparing records to the list and adding those which do not match.

This process proceeds as follows. First comparison is made with the three-digit ZIP Code for an exact match (in a normal application candidates and the master list would be sorted by three-digit ZIP Code and the appropriate section of the Master list stored in high speed memory to facilitate rapid comparison). Then the name lines are compared and passed as a potential match if each token passes by exact match, exact Soundex data match or abbreviation/nickname list match. Similarly, the address lines are compared and passed if all alphabetical tokens exact or Soundex match and the numeric information exact matches or differs in only one character.

Once the match is declared the appropriate action to the situation is performed (i.e., the record deleted, the record brought to the attention of an operator, additional records pertaining to that entry recorded in a common file, the file marked as a match for that test with a pointer to the matching file, a counter incremented to indicate how many times that file came up as a duplicate, etc.). If records do not match exactly, the decision about which data to select as correct can often be difficult.

As noted in Chapter IV the Postal Service has established a strict set of rules for identifying NCOA duplicates but left the implementation details to be defined by the licensee. This set of rules is outlined in Table V-4.

General Comments

While each of the described algorithms works they all have limitations. There are few, and no meaningful, statistics on the application of nicknames. Hence, a decision to accept "Bob" and "R" as a match is made without benefit of knowledge of the probability of "R" representing Richard, Ronald, Romeo, Randall, etc. and of the probability "Bob" not representing Robert or some other "R" name. The Type 2 error is just too large for any rule to work correctly 100 percent of the time.

TABLE V-4 NCOA Matching Rules

Street Name Comparison using ZIP+4 Code match logic based on numerical weights and penalties.

Primary Number Comparison matched only if both have the same house number, post office box, or rural route number and box number.

Apartment Number Comparison matched if numbers and order are the same and if alpha/numeric information is the same but differs because of transposition (e.g., 7J equals J7).

Name Match Comparison

Last Name - Use ZIP+4 Code logic based on numerical weights and penalties. Parse input names. Test hyphenated last names to see if one is a title (e.g., Brown-Esq.). Last name prefix comparison nearly the same (e.g., MCARTHUR equals MACARTHUR).

Family Moves - do not match to first names. First names matched only for individual moves. Match only if both have same first name. No match if either file has first initial (e.g., E. JONES does not equal ED JONES).

Middle Name Comparison only if in both addresses. There is a match only if both are spelled out and equal. If one address has a middle initial there is a match only if the other equals the first letter of middle name (e.g., JOHN B. TYAN equals JOHN BYRON TYAN).

Nickname Comparison is made by comparing best name to nickname table (e.g., BOB equals ROBERT, etc.).

Multiple Response Selections are matched by comparing qualified individuals record to family record. There is no match if input female title matches with male individual title (e.g., MS E JONES does not equal MR E JONES).

Business Name Comparison uses the match logic based on numerical weights and penalties.

All of the following address components are to be checked for during parsing: street name, apartment number, state, P.O. Box number, suffix (house number, ZIP Code), pre-direction (building name/number), and post-direction (city name, box number).

Attempts to measure these probabilities along with demographic makeup would make an interesting project.

Inference from secondary data appears to be a possible solution. Two pieces of secondary data that are most valuable are Social Security Number and telephone number. This is because few people have two Social Security Numbers and few residential addresses or individuals at a business have more than one telephone. Such matches are trivial to make and very effective if the data is available. Service Merchandise, a large wholesale to the public distributor, and L.L. Bean, a large mail order house, both are known to use telephone number as a key to their customer database. Though, if the rest of the records do not match it is difficult to tell which of the records to use as correct. Also, because of the length of such strings once the data for comparison is limited by such things as ZIP Code zone, an error in one digit in these strings has little effect on declaration of duplicates.

The more sophisticated systems do not use simple yes and no decisions to the logical decisions but assign probabilities of a match based on the number of exact matches, Soundex matches and single error matches in an input line and uses these probabilities to determine if a match exists. The setting of the passing probabilities is often left to the system operator.

Table V-5 contains a partial listing of secondary data that is known to exist in various direct marketing databases (These were extracted from the attributes of

TABLE V-5 Secondary attributes of Name/Address Records

- Gender
- Age
- Household Income
- Occupation
- Marital Status
- Number/Gender/Age of Children
- Own/Rent
- Length of Residence
- Size/Type of Dwelling
- Political Profile
    - Affiliation
    - Activity
- Lifestyle/Hobbies
    - Art/Antiques
    - Astrology
    - Automotive Work
    - Book Reading
    - Bible Reading
    - Bicycling
    - Boating/Sailing
    - Bowling
    - Cable TV
    - Camping/Hiking
    - CB Radio
    - Collectibles
    - Civic Activities
    - Crafts
    - Crossword Puzzles
    - Cultural Events
    - Current Affairs
    - Electronics
    - Fashion Clothing
    - Fishing
    - Gardening
    - Grandchildren
    - Golf
    - Cooking
    - Health Foods
    - Home Decorating
    - Video Tapes/Recording
    - Wildlife/Environment

- Type of Vehicles
- Age of Vehicles
- Vehicle Purchase History
- Telephone Number
- Subscriptions/Clubs
- Census Code
- Congressional District
- Nielsen Code
- Mail Order History
    - Purchases
    - Returns
    - Advertisement Source
    - Type Products
- Lifestyle/Hobbies (Continued)
    - Home Workshop
    - Pets
    - House Plants
    - Hunting/Shooting
    - Money Making Opportunities
    - Motorcycling
    - Needlework/Knitting
    - History
    - Computers
    - Photography
    - Physical Fitness
    - Racquetball
    - Real Estate
    - Recreational Vehicle
    - Running/Jogging
    - Science Fiction
    - Science/Technology
    - Self-Improvement
    - Sewing
    - Snow Skiing
    - Stamp/Coin Collecting
    - Stereo/Records/Tapes
    - Stocks and Bonds
    - Sweepstakes/Lotteries
    - Tennis
    - TV Sports
    - Video Games
    - Wines

Acxion Corporation's "Infobase" including National Demographics and Lifestyles, R.L. Polk and Company, SmartNames, Inc. and Donnelly Marketing databases, and Wiland Services' "Ultrabank". Much of this information is compiled from information available in the public domain. For example, SmartNames, Inc.'s "Homes" database is primarily derived from drivers' licenses, voter registration records and city and county real estate records. The reader might be shocked by the apparent sensitivity of some of this data. Opportunity clearly exists to develop rules that would provide high likelihood of match or mismatch based on individual or household profile.

Gender is an interesting attribute because of the ability to use a first name to infer it correctly. The inference of gender using first name is right from 60 to 75 percent of the time. This has high potential when one has an authoritative gender of a record and only a first initial and is matching to a record which has a high probability of inferring gender (e.g., Record 1: First Name "M", Gender "Male"; Record 2: First Name "MARY"; Inference: Do Not Match).

But secondary data is not always available and when it does exist it is often in an obscure format, possibly even purposely hidden as in the case of a list renter. Because of this, secondary data is seldom used to make duplicate decisions. On the other hand, secondary data is often used exclusively to make selections of high potential customers for mailings. In some of the more advanced mailing solicitations, the covers and inserts in catalogues or even the entire selection of flyers in a package of flyers are selected on the basis of secondary data.

In conclusion, the algorithms to detect duplicates are mature. Their implementation and success are largely based on the application. The more sophisticated (i.e., State-of-the-Art) algorithm sets are implemented effectively as rule based routines tuned, or tunable, to the application and user's desires. The rules used to develop the systems have evolved over time from proven human techniques of detecting duplicates (e.g., Match Codes and the Soundex Algorithm) and hence the routines are expert systems. A significant amount of secondary data is available to assist in duplicate detection but in few cases used for that purpose. Noted as very important among these is the telephone number.

CHAPTER VI - FUTURISTIC SOLUTIONS

A two step process was used to investigate next and future generation solutions. First ideas were identified and then they were evaluated.

The identification process consisted primarily of literature review, questioning of contacts and brainstorming with members of the CISL project. A part of this analysis included definition of the stages or steps in the process in order to identify areas where or times when the process might be improved. This definition is presented in Figure VI-1. One important observation that can be gleaned from the figure is that as with most processes the earlier a problem or mistake can be eliminated the better.

FIGURE VI-1 Generic Flow of Duplicate Detection

```
+---------------+                       +----------------+
¦   New/Input   ¦                       ¦ Product File   ¦
¦     Data      ¦                       ¦(May be modified¦
¦(May be one or ¦    +--------------+   ¦ Input or Master¦
¦ many records) ¦--->¦ Merge/Purge  ¦-->¦ File)          ¦
+---------------+    ¦   Process    ¦   +----------------+
+---------------+    ¦ (Hardware &  ¦   +----------------+
¦Existing/Master¦--->¦  Software)   ¦-->¦ Possible Human ¦
¦   Data File   ¦    +--------------+<--¦   Interaction  ¦
¦ (Preprocessed ¦                       +----------------+
¦  if required) ¦
+---------------+
```

Once potential areas for improvement were identified they were analyzed for their potential utility and benefit.

The potential improvements identified in order of occurrence in the process are:

1) Improve or modify the input and master data set. This might include establishment of a unique name (or number) for each name and address or systematically developing a better method or standard for inputing data;

2) Increase speed of processing. This would allow use of large/more complex algorithms in the same amount of time; and

3) Develop a better algorithm. Such an algorithm might surpass the level of expert system used today by application of an objective code derived entirely from a learning set of addresses.

Each of these potential improvements are analyzed below.

Input Improvement

The idea of improving the data set can take on a number of aspects. They range from assignment of unique names or numbers to every entity to simple improvement of formats, abbreviations, etc..

Conceptually the simplest of these improvements is assignment of a unique name or number to each individual and address and use these to identify duplicates. While such a system might sound like George Orwell's 1984 in many regards it has positive potential. A typical name and address combination consists of about 45 characters (one might want to add a few to include country and planet codes) selected from a set of 36 (the author assumes the use of Arabic numbers (0-9) and upper case Roman letters (A-Z) though many codes including ASCII would suffice). Allowing random order such a system could address over 10 to the 69th power individual and location combinations. But this is unreasonable since an arguably orderly set of rules (languages) are normally used to control the set of possible outcomes, after all isn't this what allows duplicate detection to work as it does today. In fact, the system of address definitions is not too bad considering about 30 characters, with some duplication in address and ZIP Code, will get you to most any address when an orthogonal coordinate set of 15 digits is necessary to get one to any 10 meter by 10 meter location using a map.

Assignment of a randomly created number to each individual is not without its problems. To account for everyone in the United States alone would require a nine-digit number (not unlike the social security number). The biggest problem with such a number is that it often gets recorded wrong. Hence a more rational system would include a parity check of some type. A simple system such as "casting out nines" would only add one digit, and be somewhat effective though that system suffers from an inability to detect transpositions and accepts 10 percent of its errors. A better system would be to add two characters and cast out 99's. Thus an eleven digit number could be used to identify each individual in the United States quite readily. To go beyond the United States would require the addition of a couple of more decimal places but at 14 digits there is no problem in uniquely identifying each person on earth for a few generations with each number having a two digit check sum. Coupling the unique individual and efficient location codes would again create an individual and address combination of about 30 characters so you may not have gone far but you would have incorporated two important elements. They are unique identity and an easy error check.

How strong the error code would be at 2 digits is illustrated as follows. Assume you have a million people with a nine-digit identification number. Addition of two more digits creates an 11-digit number. If errors occur with the entry of every 900th digit entered (one per hundred entries is assumed very good) the nine-digit system would yield 10,000 errors, the 11-digit system would yield 12,100 known errors and 122 unknown errors. Running through the numbers for a four-digit check sum reveals less than two unknown errors per million entries at the one per 900 digit error rate. Of course there are more sophisticated schemes such as those used for digital communication but those are difficult for the layman to use.

The question now is "How and when does one go about checking on and correcting the input error?" Such a number has no immediately relevant pointer to the vicinity of its individual. It is not clear that there is any benefit unless errors are detected immediately. On the other hand, the telephone companies successfully append a four-digit code to telephone numbers to create an adequately unique calling card to be secure.

The effect of going to the full use of alpha-numerics (36 characters) would only be to reduce the length of the code one or two digits and certainly complicate a simple check sums operation. On the other hand the use of mnemonics to ease recall of long strings might be helpful, if care were applied to assure that it was not misused or easily misunderstood and thus causing further confusion in the problem it was trying to ease.

Since most learned people know their Social Security Number it is not unreasonable to believe that such a system could be made to work. But, when the civil liberties problems encountered (legal and otherwise) in getting people to reveal their Social Security Number is coupled with the fact that while most educated people know their Five-Digit ZIP Code few know their ZIP+4 Code (nine-digit), such a system is not viewed as realizable without an adequate incentive, like "You can't be caught without it!"

Returning to the telephone credit card example it is highly likely that the proper address is appended to each telephone number after about two months of installation or the telephone would be disconnected when the bill went unpaid. Again the cost of doing business assures a good check on the system. This demonstrates why telephone matching is such a strong attribute in duplicate detection.

The ability to successfully use telephone, and in some transactions other credit cards with passwords in on-line services, with very low error rates leads to another possible solution. That solution is positive identification of all transactions using a credit card or password on-line service. This seems to be a logical progression as we move to a cashless society. In fact today we are not far from being able to trace all financial transactions. Again, the credit card and on-line service much like the telephone assure a properly appended address after a couple of months.

To further enhance this paperless, or at least human errorless, premonition the quality of Optical Character Readers have realistic development goals of less than one error per million characters. Using check sums and coupled to an expert system with a complete consumer database error rates could approach zero.

Another method of improving database information to allow better future matches would be to give the consumer access to the database information with the intention of letting him correct it. Optimally this would be an on-line activity which asked the consumer a series of questions about himself. A fallout of this also would be a growth in his or her appreciation for the data it contains and the importance of consistently using the same name and address. A down side of this would be people purposefully confusing the input, (e.g., for fun many people have their cat or dog receiving junk mail, and for profit people enter confused but similar names and addresses to receive multiple rebates when they are limited). Also, many people would be sensitized about the quantity of information about them in databases and possibly try to confuse it in an effort to protect their privacy. Lastly, unless all this information was placed in the public domain, or each vendor offered compensation, the likelihood of differing inputs from the same person to different databases would be significant and confusing. It is possible a "National Database" could be compiled by the Government or private firm acting for the Government could be established on a voluntary basis. Such a system would allow individuals wishing to participate the opportunity to know and control the data on file for them. Again, because of the Orwellian ramifications and whimsical and malicious entry problems, the likelihood of such a database coming to fruition are not viewed as realistic.

Looking to the future the United States Postal Service has immediate plans to effectively expand the ZIP+4 Code to ZIP+6 in, and only in, their barcoding operations. The 11-Digit Barcode (Not referred to as ZIP+6) will consist of the ZIP+4 Code with the last two digits of the street address appended to it. The rational is that in cases where the ZIP+4 Code points to a city block they will be able to automate the sort into delivery order. Their goal is to reduce average manual sort time for each carrier from four hours a day to two and a half. This could create tremendous savings to the system. To encourage this they are proposing further financial incentives to mailers to include these extra characters in their bar codes and adding additional capability to their bar code readers to assist the mailer in placing the bar code in a location more convenient to print for bulk mailing (e.g., On the top of an address label or in an envelope window). This additional two digits will only appear on the barcoding and not in the written address thus being transparent to all but the bulk mailer. This coding will not affect apartments and multistory buildings where confusion within a ZIP+4 Code zone might occur.

Because of the Postal Services methodical approach to correcting problems, other improvements not known to be planned but sure to occur include completion of the abbreviation system to add standards for words such as "Place" and "Saint".

Another place the Postal Service could improve their operation is in improving the quality of the National Change of Address data. Currently the National Change of Address data is compiled from individual input on a Post Card supplied by the Postal Service. This card requests very limited input information (e.g., one name, the old address, the new address and if change is for firm, entire family or individual signer only). The Postal service is missing a golden opportunity to be comprehensive and identify each occupant and allow for the moving of multiple occupants or one occupant with multiple names (e.g., nickname). It is even possible that such a system could be cost effectively established, at least in well populated areas, in an on-line service eliminating some paperwork and Postal Service time in entering information and possibly allowing for a more comprehensive and correct change of address entry. One foreseen problem with such a system would be control of the terminal by the Postal Service to prevent unauthorized use by pranksters, etc..

Improved Processing Speed

While they might reap the benefits of speed as it matures to an affordable price, improved processing speed will occur and is not likely to be driven by the Direct Marketing Industry. Processing speed is not entirely controlled by computer operating speed. A large portion of duplicate detection time is spent shuffling data around for comparison. Hence, a major portion of processing speed is tied up in ratios of access time to the various components of memory (i.e., RAM, disks (virtual memory), tapes. etc.). As these items mature and become available at lower prices, they will also come into use.

It would appear that the algorithms in use today are relatively mature and that increased processing speed by even an order of magnitude or more will not significantly increase their efficiency. It would, though, allow the industry to become more competitive in that the bigger businesses should be able to process more for less and smaller businesses not able to invest the time and effort to make their code efficient will be able to afford computing power to overcome this shortfall.

Improved Algorithms

Everyone in the industry spoken with felt that no major breakthrough was possible because, in the statistical sense, of Type 2 error. There are just too many exceptions to any rules that might be defined to an expert system which is effectively what the state-of-the-art systems are. Emerging spell checking algorithms are indicative of being able to identify a larger range of near misses but the rules to use this data are undefined. It is quite possible, though, that these experts being so closely involved with their current solutions are looking at solution of the problem with a non-objective view.

In terms of improving algorithm performance it is possible that detailed studies of names by demographics could identify the probability of various expert system's rules being correct and reduce Type 2 error when demographics were known or implied (e.g., In a rural neighborhood the likelihood of "Billy" not being a match with "William" are possibly higher since a first name of "Billy" is a common in rural America and seldom found other than as a nickname in an urban setting). Unfortunately, many of the findings would not have much significance and to obtain sufficient data to be statistically significant would be difficult at best. Such a project would have to be a goal of the next census. The payoff is just not clear and certainly not clear enough to justify the expense. It is possible that as computer technology evolves the ability to compile such statistics will improve.

The thought that a neural network artificial intelligence program which was fed vast numbers of correctly matched records and non-matching records could develop its own objective functions for duplicate detection is a promising but very futuristic idea. Unfortunately, the technology to operate on such vast amounts of data and then have a function useful to the handling of large quantities of data is only a gleam in the eye of science today.

CHAPTER VII - CONCLUSIONS

There are a number of conclusions that can be drawn from this activity about both the state-of-the art in duplicate detection and the direction it will take in the future.

The state-of-the-art in duplicate detection is a mature rule based expert system tuned to the application. In the simpler applications systems detect consumer duplicates well in excess of 90 percent of the time. The algorithms in use are not founded on sound scientific principles but have evolved through trial and error implementation of logical rules based on the operation of the language (e.g., the Soundex algorithm) and postal system. The most significant problems remaining result from Type 2 errors which are impossible to completely overcome since there seem to be exceptions to every rule (We have all heard Johnny Cash's song about a boy named Sue). These algorithms and their performance are only likely to improve in small incremental amounts. Because of the competitiveness in the industry it is believed that small improvements will continue to evolve possibly decreasing both Type 1 and Type 2 errors by as much as a half over the next decade.

Beyond this, three improvements are foreseen. Two of these improvements are seen as evolutionary and will mature with the industry. The other is revolutionary and very futuristic.

The Postal Service action seen as significant would be the introduction of a comprehensive list of addresses, possibly to include addressees and telephone numbers. There is no known plan to do this, but over the next generation an exceptionally complete list of this type will surely emerge solely from the National Change of Address records. For this reason, and the high probability of such a list not being exceptionally accurate, at least its first generation, this change is viewed as evolutionary. Its effect on the industry's performance will be progressive as the list and its use mature. This is the only improvement that is seen to come as a direct result of actions within the Direct Marketing Industry and could be significantly enhanced by the Postal Service's improvement of their National Change of Address input to allow for multiple name and telephone number inclusion.

Recognizing that the principal source of mailing and participation list input are the result of some financial transaction (e.g., mail order, warranty, registration, rebate application), as electronic funds transaction matures and we become a cashless society, and more people use on-line computer systems to pay bills, place orders, etc., list entry errors should decrease. The reduction of these errors coupled with use of identification numbers not subject to ambiguity will clearly help in the detection of duplicates. This will require the industry to use this secondary data which it certainly will as its quality improves. And, even if they do not choose to make use of it, the primary data which is called up through the use of electronic means should effectively be error free and hence give exact matches during merge/purge operations. The evolutionary change is not seen as being driven by the Direct Marketing Industry, but by the Banking Industry. AT&T's recently announced intention to issue banking credit cards and New England Telephone's offering of the "Info-Look" data line are excellent examples of how accounts of various types may be identified together in the future.

The last change, and next true quantum step in improved operation, will occur when large scale neural network systems become affordable for the job. One must realize that in some applications a quantum leap might not be significant (i.e., Reducing error rates from one percent to one half a percent). Yet, in other applications (e.g., company sorting) it may be quite significant (i.e., Going from error rates of 30 percent to 15 percent). It would appear that this change will not take place until well into the next century because the systems needed for this application will be required to absorb monumental amounts of data and have operation speeds far in excess of those available today. Because of the high performance standards set by the expert systems of today and evolved into tomorrow, the performance of the first generation, or more, of these systems to be developed is likely to be disappointing. This will make initial application unprofitable and maturity of the follow-on generations difficult. Key to the development of such a system will be the base from which the system will learn. A second factor complicating this implementation is the fact that the technology will not mature as a direct result of a need for duplicate detection. Hence, it may be applied in a potentially suboptimal manner to the duplicate detection problem.

It should prove interesting to see if the neural network system can effectively compete with the rule based expert system. We won't know until the next century.

APPENDIX I

Company: Acxiom Corporation (CCX Network)

Business: Provide a communication network for direct marketing industry with access to business and consumer marketing information and fulfillment services for database, individual mailing lists and merge/purges.

Target Computer(s): IBM Mainframes

Source Code: Assembly Language

Duplicate Detection Algorithms: Proprietary set of rules that include NCOA.

Interactive Capability: Yes, but not to duplicate detection.

In Business Since: 1969

Gross Sales: $20 million

Employees: > 600

Database: National Demographics and Lifestyles; Smart Names, Inc.; R.L. Polk; Donnelly Marketing

USPS Certified/Licensee: ZIP+4, Carrier Route, NCOA

Point of Contact:

Regina Mickens/Tommy Walker
Acxiom Corporation
301 Industrial Boulevard
Conway, AR 72032-7103
(501) 450-1424/(501) 329-6836

APPENDIX II

Company: Consultants for Management Decisions, Inc.

Business: Management consultants specializing in computer solutions to management problems.

Application: System (CitiExpert) to process fund transfer telexes automatically.

Source Code: C and Assembly mix.

Duplicate Detection Algorithms: Internally developed expert system which parses, converts each token to standard form, compares each to list of standard elements and existing accounts.

Interactive Capability: Final output is confirmed by human.

In Business Since: 1982

Units in Service: One

Employees: Approximately 35

Point of Contact:

Kenan E. Sahin
Consultants for Management Decisions, Inc.
One Main Street
Cambridge MA 02142-1517
(617) 225-2220

APPENDIX III

Company: Creative Automation Company

Business: State-of-the-art computer services for the direct marketing industry including merge/purge; ZIP Code correction; list enhancements and overlays; nixie elimination; address correction; credit screening; carrier route and Five-Digit postal presorting; impact, laser and ink-jet personalization; continuous forms bursting, trimming and folding services; mailing list maintenance and rental fulfillment; and response analysis systems.

Target Computer(s): IBM Mainframe

Duplicate Detection Algorithms: Wide variety of options tuned to application.

Interactive Capability: No

In Business Since: 1969

Units in Service: (IBM Mainframe); (PC-MS/DOS)

Employees: 150

Database: Customers

USPS Certified/Licensee: ZIP+4, Carrier Route

Major Customers: American Family Publisher, CitiBank

Point of Contact:

Neil Sorensen
Creative Automation Company
220 Fencl Lane
Hillside, IL 60162
(312) 449-2800

APPENDIX IV

Company: Epsilon

Business: Full service database marketing company provides commercial and non-profit clients with database marketing services. Services include strategic planning and consulting database management, market research and analysis personalized direct mail; creative production and fulfillment services; telemarketing sales lead management and direct mail fund raising.

Target Computer(s): IBM Mainframe

Source Code: ALC (IBM Mainframe, Group 1)

Duplicate Detection Algorithms: Group 1

Interactive Capability: No

Gross Sales: $50 million

Employees: 500

Major Customers: Amtrak, The Chase Manhattan Bank, N.A., Texas Instruments, Smithsonian Institution, Bauch & Lomb

Point of Contact:

Eileen M. Sullivan
20 Cambridge Street
Burlington, MA 01803
(617) 273-0250

APPENDIX V

Company: First Data Resources, Inc.

Business: Postal presort services.

Target Computer(s): IBM Mainframe

Source Code: COBOL

Duplicate Detection Algorithms: Straight NCOA

Interactive Capability: No

In Business Since: 1971

Gross Sales: $300 million

USPS Certified/Licensee: Five-Digit ZIP, ZIP+4, NCOA
(Claim Carrier Route but Postal Service does not list)

Major Customers: Bankcard Services

Point of Contact:

Dave Ingwersen
First Data Resources, Inc. (An American
10825 Farnam Drive Express Company)
Omaha, NE 68154-3263
(402) 392-5203
(800) 643-2828

APPENDIX VI

Company: Flowsoft Custom Programming

Business: Develop and distribute personal computer based software package that helps businesses and organizations sort, label and assemble mailings.

Target Computer(s): PC-MS/DOS

Source Code: Assembly

Duplicate Detection Algorithms: Match Codes

Interactive Capability: 100%

USPS Certified/Licensee: Carrier Route

Point of Contact:

William A. Anderson
Flowsoft Custom Programming
1166 Franklin Road, Suite A-2
Marietta, GA 30067
(404) 955-5461

APPENDIX VII

Company: GROUP 1 SOFTWARE, INC.

Business: Develop and market comprehensive line of mail management, postal discount and laser-printing personalization software for IBM mainframe and PC.

Target Computer(s): IBM Mainframe; PC-MS/DOS

Source Code: ALC (IBM Mainframe); C & Assembly (PC-MS/DOS)

Duplicate Detection Algorithms: Weighted matching logic.

Interactive Capability: Yes, PC-MS/DOS product only.

In Business Since: 1973

Units in Service: 195+ (IBM Mainframe); 1000+ (PC-MS/DOS)

Gross Sales: $19 million

Database: Not Applicable.

USPS Certified/Licensee: Five-Digit ZIP, ZIP+4, Carrier Route, Users are NCOA certified.

Major Customers: TRW, Automated Image Management

Point of Contact:

Patti Cutchis
Group 1 Software (A Comnet Company)
Washington Capitol Park
6404 Ivy Lane, Suite 500
Greenbelt, Maryland 20770-1400
(301) 982-2000 Extention 336
(800) 368-5806

APPENDIX VIII

Company: Harte-Hanks

Business: Comprehensive consumer database management specializing in record keeping.

Target Computer(s): IBM Mainframe

Duplicate Detection Algorithms: Tables, weighted Match Codes

Interactive Capability: No

In Business Since: 1968

Database: Customers

USPS Certified/Licensee: ZIP+4, NCOA

Point of Contact:

Bill Maxfield
25 Linnell Circle
Billerica, MA 01821-3961
(508) 663-9955

APPENDIX IX

Company: LPC, Inc.

Business: Supplier of computer software and services for name and address applications software standardizes, verifies copies, presorts, merges/purges, highlights missing apartment numbers and makes up mail to maximize postal discounts and to reduce volume of undeliverables.

Target Computer(s): IBM Mainframe

Source Code: COBOL and Assembly

Duplicate Detection Algorithms: Logic rules with weighting matching and Match Code depending on application. Soundex and Phonetic algorithms.

Interactive Capability: Yes, full interactive capability with purchase of separate package.

In Business Since: 1972

Units in Service: 165+ IBM Mainframe

Employees: 80

USPS Certified/Licensee: Five-Digit ZIP, ZIP+4, Carrier Route

Point of Contact:

LPC, Inc. (A Pitney Bowes Co.)
1200 Roosevelt Road
Glen Ellyn, IL 60137-6098
(312) 932-7000/(800) MAI-LERS

APPENDIX X

Company: United States Postal Service (USPS)

Business: Provide mail delivery. Establish standards. Certify vendors and licensees.

Target Computer(s): IBM Mainframe

Source Code: COBOL

Duplicate Detection Algorithms: NCOA Standard

Interactive Capability: No

Gross Sales (FY-88): $38 billion (160 billion pieces of mail. 34 percent (54 billion) of which were sorted on automated equipment. Of this 54 billion pieces, 42 billion of the 46 billion flats (91 percent) were presorted. Third class bulk revenues were $7.3 billion).

Employees: 756,600

Database: NCOA and own

Point of Contact:

Ann Harrison/Maggie Jones (CASS)
Mike Murphy (NCOA)
National Address Information Center
United States Postal Service
6060 Primacy PKY
Memphis, TN 38188-0001
(800) 238-3150

APPENDIX XI

Company: WILAND SERVICES, INC.

Business: Comprehensive services for list maintenance and promotional programs including merge/purge and postal presort. Donor file maintenance for nonprofit mailers. File maintenance and database management for catalogers, financial companies, retailers and publishers. Specialize in database clean up/supplementation. Beginning software sales.

Target Computer(s): IBM Mainframes

Duplicate Detection Algorithms: 16 selectable rules

Interactive Capability: No

In Business Since: 1971

Gross Sales: $16 million

Employees: 310

Database: 88 million addresses; 215 million individuals; track at least 12 attributes including age, income, occupation, sex/marital status, number/gender/age of children, length of residence, own/rent, dwelling size, resale value of vehicles, type vehicle, new vehicle purchase history and telephone number.

USPS Certified/Licensee: ZIP+4, Carrier Route, NCOA

Major Customers: Sears, Bank of America, Condenast (Magazine subscriptions), Sharper Image

Point of Contact:

Leigh A. Lelivelt/Bill Kindelberger
6707 Winchester Circle
Boulder, Colorado 80301-3598
(303) 530-0606/(800) 869-LIST

BIBLIOGRAPHY

"Address Information Systems"; United States Postal Service; Publication 40; October, 1988.

"Annual Report of the United States Postal Service Fiscal Year 1988".

"Automation Plan for Business Mailers"; United States Postal Service; Publication 67; October 1989.

Berghel, H.L.; "A Logical Framework for the Correction of Spelling errors in Electronic Documents"; Information Processing & Management, Volume 23, Number 5, Pages 477-494, 1987.

Berghel, Hal; Roach, David; Talburt, John; "Applications of Approximate String Matching: The logic of Spelling"; PC AI; January/February 1990; Pages 24-27.

Berghel, Hal; Roach, David; Talburt, John; "New Directions in Approximate String Matching: Intelligent Problem Solving with Strings"; PC AI; September/October 1989; Pages 24-27.

Berghel, Hal; Roach, David; Talburt, John; "New Directions in Approximate String Matching: The Mechanical Cruciverbalist"; PC AI; November/December 1989; Pages 45-47.

Burnett, Ed; The Complete Direct Mail List Handbook; Prentice-Hall; 1989.

Campbell, J.A.; Cuena, J.; Perspectives in Artificial Intelligence, Volume 2: Machine Translations, NLP, Databases and Computer Aided Instruction; Ellis Harwood Limited; 1989.

Chait, Lawrence G.; "Direct Marketing in the New Artificial Intelligence Age"; Direct Marketing; January, 1985.

Churbuck, David; "Smart Mail"; Forbes; January 22, 1990; Pages 107-108.

"The Complete Address Guide"; United States Postal Service; 36 USC 380; 1990.

"Consumer Marketing Duplicate Detection"; Market/Customer Database Systems; June, 1988.

Data Sources (Software) 1st Edition 1989; Ziff-Davis Publishing Company; Pages J-553-J-564.

Dillingham, Susan; "Business: Cards Get Credit for Smartness"; Insight; October 24, 1988; Pages 44-45.

"Direct Marketing Expert System Serves IBM Shops"; Computerworld; December 31, 1984; Page 130.

The Direct Marketing Market Place, 1988 Edition; Hilary House Publishers; Pages 245-266.

Edwards, Paul L.; "American Express Testing New Software Aid", Advertising Age; July 8, 1985; Page 781.

Encyclopedia of Associations (24 Edition); Gale Research, Inc.; 1990.

"5 Creative Solutions for Your Business Needs"; United States Postal Service; 1988.

Hall, Patrick A.; Dowling, Geoff R.; "Approximate String Matching"; Computer Surveys, Volume 12, Number 4, Pages 381-402, December 1980.

Karr, Albert R.; "Labor Letter: A Special News Report on People and Their Jobs in Offices, Fields and Factories: Working the Mail"; The Wall Street Journal; March 20, 1990.

Keller, John J.; "AT&T Launches Its Credit Card for Consumers"; The Wall Street Journal; March 27, 1990.

Kupfer, Andrew; "Technology: Now, Live Experts on a Floppy Disk"; Fortune; Pages 69-79; October 12, 1987.

Lorin, Harold; Sorting and Sort Systems; Addison-Wesley; 1975.

Madnick, Stuart E.; et al; "CISL: Composing Answers from Disparate Information Systems"; Extended Abstract from IEEE Workshop on Heterogeneous Database Systems; September 1989.

Madnick, Stuart E.; Wang, Richard Y.; "Logical Connectivity: Applications, Requirements, and an Architecture", MIT Working Paper Number 2061-88; August, 1988.

Miller, Richard K.; Neural Network Implementing Associative Memory Modules in Neurocomputing, Volume II: Research Assessment; SEAI Technical Publications; 1987.

"National Change of Address"; United States Postal Service; Notice 47; March 1988.

"National Deliverability Index"; United States Postal Service; Notice 41; March 1989.

Nickel, Karen; "Policy and Politics: Can This Man Really Deliver"; Fortune; August 14, 1989.

"Operation Mail"; United States Postal Service; Notice 48; March, 1988.

Podems, Ruth; "What's in Store for the 1990s?; Target Marketing; September 1989; Pages 22-23.

Polilli, Steve; "Persoft Direct Mail Package Picks Profitable Buyers"; MIS Week; July 17, 1985; Page 78.

"Postage Rates, Fees, and Information"; United States Postal Service; Notice 59; April, 1988.

Postal Bulletin 21750; United States Postal Service; November 16, 1989; Pages 40-44.

Russell, Robert C.; Patent 1,261,167; "Index"; 1918.

Russell, Robert C.; Patent 1,435,663; "Index"; 1922.

Russell, Robert C.; Patent 1,435,664; "Index"; 1922.

Ryan, Alan J.; "Mail System Dials in on Targets"; Computerworld; May 11, 1987; Pages 23-27.

Sahin, Kenan; Sawyer, Keith; "The Intelligent Banking System: Natural Language Processing for Financial Communications"; Innovative Applications of Artificial Intelligence; Pages 43-50; AAAI Press/MIT Press.

Sales Literature, Acxiom Corporation.

Sales Literature, Creative Automation Company.

Sales Literature, Epsilon.

Sales Literature, First Data Resources, Inc..

Sales Literature, Flowsoft Custom Programing.

Sales Literature, Group 1 Software.

Sales Literature, Harte-Hanks Data Technologies.

Sales Literature, LPC, Inc..

Sales Literature, Wiland Services.

Shogase, Hiro; "The Very Smart Card: A Plastic Pocket Bank"; IEEE Spectrum; October, 1988; Pages 35-39.

Sroge, Maxwell; *Inside the Leading Mail Order Houses*; NTC Business Books; 1987.

Stone, Bob; "Direct Marketing: Then and Now"; Direct Marketing; May 1988.

"TRW Says It is Working on 'Largest Merge/Purge' of 490-Million Names"; DM News, December 15, 1987.

"ZIP+4 Code"; United States Postal Service; Notice 186; May 1988.

BIOGRAPHICAL SKETCH

Heber R. Norckauer, Jr. is attending Massachusetts Institute of Technology as a Alfred P. Sloan Fellow from his position as Chief of the Special Program Office's Technical Management Division at the United States Army Missile Command, Redstone Arsenal, Alabama. Prior to assuming this position, he served as an Engineering Manager in the Multiple Launch Rocket System Program Office and in the United States Army Missile Research, Development and Engineering Center, both located at the United States Army Missile Command. Upon graduation from college he received a commission in the United States Army Ordinance Corps and served four years active duty. He received a Bachelor of Science degree in 1972 and an Master of Science degree in 1974, both in Physics and from Louisiana State University. He is a 1986 graduate of the Defense Systems Management College's Program Management Course. He and his wife, Sheila, reside in Guntersville, Alabama, where they are regularly visited by his two sons. His hobbies include boating, shooting sports, radio control modeling, and bowling. He is a Lifetime Member of the National Rifle Association and a member of Ducks Unlimited.