Whose Name Is It:
Names, Ownership and Databases

Kerry Dematteis, Richard Lutz and Heather McCallum-Bayliss

2214 Rock Hill Road, Suite 201
Herndon, VA 20170
www.las-inc.com

1.1.1 Introduction

Personal names are important pointers to individuals in a society. Whereas in small, tribal societies, the context between name as label and its referent is transparent and direct, in modern technological societies, there is often great distance between the name as label and the person to whom it refers. This is especially true in cases where names are stored within large databases. These include government, medical, educational and even commercial records that are kept about individuals. Problems arise when attempting to retrieve records from those databases. How a name is stored within data records may, and often does, deviate in form from the way it is entered at the time of query. Indeed, personal names pose special problems in terms of data retrieval because names exhibit much more variation in form than do other lexical items. The word *chair* can refer to any members of the set of chairs, but its written form is fixed by standard English orthographic conventions. Names such as "Leigh" or "Johansen", "Stephen" or "Jeffrey" have a number of common spellings, and probably a number of uncommon ones as well.

Furthermore, compared to other labels, names, particularly names of persons, organizations and locations, exhibit much wider variation. Dates, for example, can be entered into a database according to a variety of formats, but those formats form a more-or-less closed set: month-day-year, day-month-year, full name of month, ordinal value of month, and so on.

In the case of personal names, some variation is predictable, and even acceptable. Nicknames, use of initials, use of maiden names are just a few of the more obvious ways that data entry of a name might vary. In addition, names that are 'similar" historically or simply phonetically or orthographically might be substituted for one another by those who do not know how the referent refers to himself. Variation in names is a source of concern, particularly in societies as culturally diverse as ours, where different naming conventions, different languages and writing systems and creative individual preferences come into contact with one another.

Managing databases of names is hardly a new problem, of course. As far back as the written record extends, governments have kept records for purposes of taxation, military

inscription and population census. And a variety of approaches have been tried as well. The Soundex system of filing names, first used for the U.S. census in the early part of this century, is still widely used, but limited in the kinds of variation it can handle and the level of recall and precision that it can achieve. And as databases increase in size, the need to automate the retrieval process becomes more imperative. Automation, in turn, requires that names be handled in ways that are sensitive to the ways they vary.

In this paper, we will discuss the problems with personal names as they relate to databases and automatic matching and retrieval systems. We will outline some of the sources for the variation in those databases, and propose an approach to responsible stewardship of proper names, especially as relates to an automated computer environment.

1.1.2 Names in Databases

It is reasonable to wonder why personal names seem to challenge automated matching and retrieval systems. It is unusual to think that a person's name poses any sort of general difficulty. After all, a name belongs to a person; that person "knows" his/her name and uses that name for personal identity. There is an assumed association of a name and a single person and therefore personal names are viewed as fixed items, much like numbers.

However, the apparently inseparable link between the name and the person can be broken when a name is entered into a database. There is now a dissociation of the name and the individual, so the ability of the name to discriminate uniquely is reduced, if not eliminated. The name in the database may refer to *many* people with the same name or a similar name; the name now selects *a group of individuals*.

Verification of identity is very difficult, then, because the pool of candidates can be quite large, even candidates with the same name. And, if the amount of information available is limited or abbreviated (e.g., ROBT for ROBERT), the chances of a successful match are even fewer.

Retrieval of records with reduced name information (such as initials) may be improved if additional personal information is available. Information, such as one's Social Security Number or date of birth or address, can be used to reduce the size of the group and increase the likelihood of identity significantly. But an error in any one of these adjunct data elements can mean that the name will not provide enough information to make a match successful.

It is important, therefore, to overcome a name's inability to identify uniquely when it occurs in a database by providing as much information as possible for a match.

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

1.1.3 Variation in Names

Personal names can not only refer to more than one individual but they can also pose another difficulty for establishing identity: personal names are not fixed data elements. They are fuzzy elements for which principles of similarity need to be determined. That is, personal names require that there be an understanding of when two names can be considered similar and to what degree.

In this regard, personal names do not function like other data elements with which we are more familiar – numbers, for example. The Social Security Number (SSN), for instance, is a fixed string of digits which has been issued to one and only one individual. Fraud and error are the only ways in which the SSN can vary but such variation produces a different SSN, which presumably belongs to someone else (or to no one, if it is an impossible string). Although some of the errors in an SSN may be predicted and can therefore be accommodated with special manipulation techniques, the fact remains that the incorrect SSN is a different one until identity can be reestablished.

Personal names, on the other hand, can show significant, sanctioned variation without losing the ability of that name to refer to the same person. Take the following name, for example; all the forms given could appropriately and correctly be used by Dr. Morgan with no intent to defraud and with no obvious error:

(1) ARNOLD BLEDSOE MORGAN
A. BLEDSOE MORGAN
A. B. MORGAN, JR.
ARNIE MORGAN
DR. A. MORGAN

None of the above variations is seen as unusual, which indicates that the kinds of variations noted are recognized as permitted differences in the way a single individual could represent his/her name. A full name, a nickname, a title (DR.), a qualifier (JR.) and initials are some of the ways in which one person might choose to present his/her name on different occasions. When these names occur in a database, however, it may not be as clear that all these names belong to the same person. DR. A. MORGAN could refer to DR. ANNABELLE MORGAN; A. B. MORGAN, JR. could refer to ARTHUR BRANDT MORGAN, JR. That is, although variations are permitted, these variations can interfere with identifying a small set of individuals in a database: The less specific the information in the name, the larger the set of individuals to which the name can refer.

The issue becomes more problematic when dealing with names from other cultures because the sorts of variation which are permitted in names may not be the same as those permitted in American names. For example, which of the names in each of the following sets are variations of one another?

(2) PARK DOE REE
PAG TO NI

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

TO NI PAG
(Korean)

MOHAMMAD ALI ABD EL NADIR NUR EL DIN
IMHEMED ABDUNADEER NOOREDDINE
MHMD NUR ABD AL NADER
(Arabic)

ENRIQUE CESAR VELEZ ARGUETA
ENRIQUE BELES
QUIQUE VELEZ A.
E. C. ARGUETA
(Hispanic)

Within each of these cultures, all the names given are permitted variants of the same name *except* the last one. In each case, the final name would be considered an unacceptable variation of the name under consideration; it would be another name. In the Korean naming system, the family name appears in the leftmost position and cannot move to the rightmost position. In Arabic names, name order is crucial; although the spelling variants of the name elements in the final name are acceptable, their order is not. In Hispanic names, the family name is the *next to last* name (VELEZ); the rightmost name (ARGUETA) may be dropped, but not the family name. ARGUETA would therefore refer to another family, if it occurs alone.

1.1.4 Categories of Name Variation

Each culture has a set of conventions which govern the appearance and function of personal names and has a range of permitted variation in its naming system. The categories of variation are generally the same across all cultures, however; it is how the variations are realized that differs. Names can vary in the following ways.

1.1.4.1 Spelling

To a greater degree than is observed in other textual material (i.e., words in text), spelling variation occurs pervasively in personal names:

(3) GODDARD / GOTTHARDT / GOEDHART

Because of its role as label, individualized ways of writing names abound in many societies. Japanese names, for instance, exploit the fact that multiple *kanji* characters typically share one pronunciation to create new and evocative ways of writing given names.

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

English orthography, with its many-to-many sound/letter correspondences, contributes to the problem in English-speaking countries, whether or not the name is English in origin. Dialectal differences, historical spellings and phonetic spellings all make the spellings of English names unpredictable. Thus, examples such as the following abound:

(4) BEAUCHAMP/BEECHAM
LEE/LEIGH
CONNOLLY/CONNALLY/CONALLY/CONLEY
WORCESTER/WOOSTER/WORSTER
THOMSON/THOMPSON/TOMSEN

Anglicized pronunciations of names of non-Anglo origin are especially likely candidates for misspellings once the link between the "owner" of the name and the data being written has been broken:

(5) GOEAS
RZEHAK
DJORDJEVIC

Spelling variation is especially conspicuous in names from cultures which use non-roman writing systems when such names have been subjected to romanization, e.g., the romanized form of an Arabic name:

(6) NOOR EL DIN / NURELDIN / NUREDDINE

Transcription systems often serve as a standard for formulating a romanized version of a name. For example, one transcription system for Korean Hangul will specify that the Hangul symbol for the Korean sound [p][1] (unaspirated voiceless bilabial plosive) is written as the roman letter P. Another Hangul transcription system may indicate that the Hangul symbol for Korean [p] should be written as the roman letter B. The name PARK will therefore vary with BARK:

(7) PARK/BARK

Because both transcription systems are standards, the roman transcriptions which they prescribe can be used to develop rules which will predict the differences in spelling that are generally found in the romanized forms of Korean names. A Korean spelling variation rule would, therefore, state that P and B may be substituted for one another in the romanized forms of Korean names.

In many cultures, available standard transcription systems are not used or are used inconsistently. The range of variation found in distinct instances of the same name is

[1] Square brackets, [ ], represent the standard linguistic convention for representing sounds rather than letters.

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

therefore not fully predictable from such systems. In Arabic, for example, although there are transcription systems used by libraries and other official agencies, transcription tends to be far less predictable and highly inconsistent, even with a single individual. For example, an individual whose name is "ABD EL NADIR" may romanize the name on one occasion as ABDUL NADEER and on another as ABDUNNADIR:

(8) ABD EL NADIR
ABDUL NADEER
ABDUNNADIR

Both name representations are "correct" and can be said to be accurate romanizations of the same Arabic name.

Even in cultures in which transcription systems provide a reliable standard, personal interpretation, accommodation to the spelling of another culture or perceptual confusion can cause the spelling to deviate from the standard. So, for example, the Korean name GO will vary with KO, because G and K are romanization alternatives from different transcriptions systems. An observed variant of GO, however, is GOUGH, showing the influence of English spelling:

(9) GO
KO
GOUGH

Spelling variation is one of the more challenging problems for automated multicultural name matching systems. One primary reason is that systems today use leveling techniques (whether they are keys or name matching rules) which are based on spelling variation appropriate for Anglo names. For example, the I and Y of SMITH and SMYTH would produce expected variations in English. R and RR (in HARRIS and HARIS) would also be found in variations of the same name in English. The same is not true of Hispanic names, however. MORO and MORRO are two different names because R and RR in Spanish do not generally produce variants:

(10) SMITH/SMYTH
HARRIS/HARIS
MORO/MORRO

It cannot be assumed that spelling variation associated with Anglo names is relevant to the names of other cultures.

Additional difficulties result from keys (such as the widely used Soundex key) which keep stable the first character of the name. This practice may result in match failures in Anglo names (FILIP/PHILIP) but even more readily in names from other cultures: VELEZ/BELES (Hispanic); GO/KO (Korean); MOHAMMAD/IMHEMED (Arabic):

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

(11) FILIP/PHILIP
VELEZ/BELES (Hispanic)
GO/KO (Korean)
MOHAMMED/IMHEMED (Arabic)

In addition, because such keys are derived from the consonants in a name, variation in the consonant inventory of a name will prevent a match. Note the presence of R in one variant of this Korean name and not in the other:

(12) PARK/PAK (Korean)

1.1.4.2 Morphology

Variation can also take place within the internal structure of a single name as well as to the "peripheral" name element(s) within that name. These include name affixes, such as the O' in the name O'LEARY; MAC in the name MACDONALD; DE LA in the name DE LA FUENTE and the SON in DONALDSON:

(13) O'LEARY
MACDONALD
DE LA FUENTE
DONALDSON

While name elements such as these had historical significance, most of them have lost their meaning and have become a part of the name itself; they are not used independently of the name stem. Because they have lost their meaning, at least within the morphology of personal names, they may be written in a variety of different ways: They may be conjoined to the base name; they may have a following space; they may have following punctuation; they may show variation in capitalization:

(14) OLEARY/O'LEARY
MAC DONALD/MACDONALD/MacDonald/Macdonald
DE LA FUENTE/DELAFUENTE

All present challenges for name match and retrieval systems.

In some naming systems, especially those which are written in non-roman alphabets, the realization of prefixes and suffixes may be even more varied. In Arabic, for example, prefixal elements may be conjoined to the base, change their spelling, undergo assimilation or other morphophonemic processes and may even be deleted. For example, the Arabic prefix ABD EL, meaning 'servant of the' may be represented in a variety of different ways, all acceptable:

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

(15) ABD EL RAHMAN
ABDUL RAHMAN
ABDURRAHMAN
ABD ELRAHMAN
ABDULRAHMAN
RAHMAN

Name affixes (both prefixes and suffixes) manifest a wide range of variation in the way they can be written. This diversity can pose significant problems for name retrieval and automated identification. If the presentation of the name is inconsistent from event to event, the chances of retrieval are significantly reduced. Such inconsistency in representing the affixal elements of the name may arise from any number of sources, including

- an individual's presentation of his/her own name due to a decision to change the representation or to be inconsistent in the way the name is represented;
- an employer's interpretation of an employee's name;
- an interpretation by a data entry person unfamiliar with the name format of another culture;
- constraints of the data entry format provided for the name (e.g., the name field is not long enough); or
- inadequate or non-existent instructions for name entry.

Whatever the source, inconsistency in the affix format can defeat a match. Even data edit rules which attempt to reduce the variations in such formats must be highly sophisticated and general enough to anticipate significant diversity:

(16) DE-LA-FUENTE
DELA FUENTE
DE LA FUENTE
DE LAFUENTE
DELAFUENTE

1.1.4.3 Syntax

By syntax of names, we are referring to the order of the name elements, to the permitted segment substitutions and deletions and to the value assigned to the name segments. Name segments have functions in the name format of the culture. The standard Anglo naming convention, of course, allots a role for given, middle and last name:

(17) Given Name: WILLIAM
Middle Name: BRAYSON
Last Name: INGLES

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

In addition to recognizing the role and order of each of the name segments, a person familiar with Anglo naming conventions would recognize that particular values are attributed to the various segments. We organize our official records by these values. The Last Name is the most valuable, the First Name is second in value and the Middle Name plays a much less valuable role.

(18) Anglo Naming Convention

FIRST NAME	MIDDLE NAME	LAST NAME
WILLIAM	BRAYSON	INGLES

Different conventions may govern the role, order and value of the name segments in other cultures. For example, the Chinese name CHUNG BO LI has the structure:

(19) Family Name: CHUNG
Given Name: BO
Given Name: LI

The values attributed to each of these is:

(20) Chinese Naming Convention

FAMILY NAME	GIVEN NAME	GIVEN NAME
CHUNG	BO	LI

Noteworthy is the placement of the surname within the Chinese surname in the leftmost position. Note too that the two Given Names are of equal value; that is, they readily occur together and the second one is rarely reduced to an initial, as it would be in an Anglo name.

The consequence of these differences in name order, name role and name value is that a name matching system or any matching rules and techniques must be sensitive to the matching of appropriate name elements, must not assume that the structure of a name from another culture is the same as that of the Anglo structure, must be able to anticipate variation within the conventions of another culture and must be aware of the potential inconsistency which cross-cultural representations can generate.

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

1.1.4.4 Social Influences

There are many name external factors which influence how names are reported or recorded. These factors can be classified as "social" conditioners; among them, the level of formality; the degree of assimilation into a new culture; the perceived markedness of the name. As these factors change, so do the name representations. For example, the level of formality may influence how an individual will represent his/her name. Under formal circumstances, PEGGY MCLAIN may be DR. MARGARET WOODLEY MCLAIN. PACO HERNANDEZ may be SR. FRANCISCO ESTEBAN HERNANDEZ LOPEZ. Under slightly less formal circumstances, PEGGY MCLAIN may be MARGARET W. MCLAIN and PACO HERNANDEZ could be FRANCISCO HERNANDEZ L. Notice that degrees of formality prescribe different representations of the names from different cultures.

(21) PEGGY MCLAIN/DR. MARGARET WOODLEY MCLAIN
PACO HERNANDEZ/SR. FRANCISCO ESTEBAN HERNANDEZ LOPEZ
PEGGY MCLAIN/MARGARET W. MCLAIN
PACO HERNANDEZ/FRANCISCO HERNANDEZ L.

Another factor that can influence the representation of a name is *acculturation*; that is, the degree to which a person from another culture takes on the conventions of the culture in which he/she finds him/herself. The adage "When in Rome, do as the Romans do" still governs many assumptions about expected behavior from persons from other cultures. However, there are circumstances in which this adage does not readily apply. An American in Japan, for example, is expected to present his or her name according to American syntax, not Japanese, both in speech and on the ubiquitous *meishi* (business card).

Immigrants, on arrival to the United States, may not speak the language, let alone have assumed the norms of the new culture. This is especially true with respect to their names, which label them uniquely within their own culture. Their passports may contain their name in its original form; their alien cards may have been based partially or in full on the passport; and the name on their Social Security card must agree with the documentation provided, often the passport and alien card. The naming conventions of the native culture are therefore perpetuated even if it is recognized that the name is "incorrect" according to the American naming conventions. A non-English speaking Somalian woman was observed applying for a Social Security card. Her sister, who spoke some English, accompanied her. Her documentation listed her surname as MAHAMUD, although her application was in the name MOHAMUD. Neither sister seemed to be concerned about the misspelling, which, by rule, was the way that the name would be placed on the Social Security Card. Two sisters residing at the same address would now have Social Security cards with different spellings of the same name. If the new arrival obtains a job, it would not be surprising to find that an employer might interpret the name in yet another way: MUHAMAD:

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

(22) MAHAMUD/ MOHAMUD/ MUHAMAD

Other examples of acculturation introduce other sorts of variation into the process of reporting a name. If NO AHN TOK, a new Korean immigrant, acculturates and assumes the American naming system, he may become AHN TOK RHO, moving his surname from the leftmost position to the rightmost and changing the spelling of his surname to a less marked form (cf., DR. NO) of the name (NO to RHO, which are both acceptable spellings of the same Korean name):

(23) NO AHN TOK/AHN TOK RHO

Another technique which signals acculturation, and which is often motivated by a desire to reduce the markedness of the name, is translation of a name. JUAN might become JOHN; GRUN may be translated to GREEN. Similarly, a name may be changed to conform to spelling conventions which govern Anglo names: the Croatian name DJORDJEVIC may be changed to GEORGEVICH to encourage proper pronunciation:

(24) JUAN/JOHN
GRUN/GREEN (cf. Non-nativized GRÜN/GRUEN)
DJORDJEVIC/ GEORGEVICH (cf. Non-nativized DJORDJEVI④)

If such adjustments of the name are not formalized, there may well be a mismatch between how an individual name is entered at the time of recordation and at the time of retrieval. And it may be that changes such as these are not understood as true name changes but rather accommodations to the new culture, acceptable variants or clarifications. A formal change in the name may, therefore, not be considered necessary.

1.1.5 Consequences of Name Variation

Clearly, an understanding and recognition of the permitted range of variation in names from other cultures is central to adequate name matching and retrieval. A single-faceted approach to dealing with personal names has proven generally ineffective. Because names vary in different ways, approaches to dealing with names must be flexible enough to accommodate these differences. An approach that works for one culture (e.g., reduction of the name which occurs in the Middle Name position to an initial) may hamper comparison with names from another culture because vital name information has been abbreviated. On the other hand, the variations in names from other cultures are generally predictable; their governing principles can therefore be included in a match system with the felicitous consequence that the likelihood of matching is increased.

1.1.6 Conclusions

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA

There is no single solution to the problem of name handling, storage and retrieval. In fact, the optimal solution lies in the ability of a system to find a balance among the three aspects of name management: the user, the data and the system or algorithm used to retrieve the name. When any organization is confronted with the task of setting up databases of names, updating the system with additional names, or retrieving database names, proper data stewardship of names is essential in order to accommodate the special characteristics of names as data. Data stewardship includes the following:

- The *user* must be schooled in the responsibility which he/she has for maintaining the quality of the data and be provided with the appropriate resources so that he/she can make motivated interpretations about the role of the name elements presented, especially for names from cultures with which he/she is not familiar.
- The *data* need to be collected in a standardized format, which allows for sufficient flexibility to accommodate names from a variety of cultures and does not prejudice the interpretation of the name parts. The data need to be manipulated in ways which are appropriate for recognizing the scope of variation which can occur in names – from this or other cultures and to be stored in ways which do not impair irreparably the chances of identification.
- The *algorithms* – the match criteria – need to be designed to recognize and allow for similarity which is relevant to the conventions which govern other cultures and to be able to address the issues which occur in the data (such as conjoined and unconjoined multipart names).

Too much burden on any one of these aspects of the name management process will produce a less than optimal system. For example, it is unreasonable to expect that a user could learn information about the naming conventions of all cultures of the world. Although such knowledge might result in a reduced need for sophisticated algorithms, the outcome is likely to be less than hoped for. On the other hand, if all responsibility is placed on the algorithms to anticipate every potential variant of names from all cultures, it may reduce the need for the user to be highly skilled but the user will also not recognize his/her responsibility for data stewardship. The required degree of complexity in the algorithms will also be unmanageable.

The goal of any systemic approach to these problems is therefore balance. The key is to maintain a balance between the competing forces and provide them all with the resources necessary to permit careful and proper handling of the names they encounter.

Originally written for:
1998 Annual Meeting
American Name Society
San Francisco, CA