Name or Number – Which Shall it Be?*

ROBERT M. LANDAU (357-03-6623)

ARE YOU A NAME OR A NUMBER? This presentation attempts to answer the question and to explore the traditional role of names identifying places, things, and persons. Name origins, grouping problems, and transliteration are considered; a short description of the Board on Geographic Names is given; non-personal names are covered briefly. Major emphasis is placed, however, on personal names: types, structure, methods of handling large volumes, and recent automation activity in this field.

Let us ask ourselves, "In a symbolic sense, are we today a name or a number?" If we could gaze into a crystal ball and see how many times each of us is being referred to in some transaction or other – a charge account posting, an income tax return audit, the recording of a deed, the posting of a personal check in a bank, a social security retirement entry and so on – I think first we would be amazed at the large number of transactions, undoubtedly dozens and perhaps hundreds of times each day for each of us. Thus we are speaking of billions of data transformation actions each day. Moreover, it is my estimate that the majority of these transactions referring to us use a *number*, not our names as the key, for there has been a strong trend in this direction in recent years. Why is this so? It is mainly because numbers can with ease and fewer characters be discretely assigned to each person, and thus are easier to work with than names, both manually and by machine methods. Extrapolating this trend, can we then conclude that it is only a matter of time before we all truly become a mere number in our symbolic world? There are those who would so conjecture. It is suggested that we have our social security number tattooed on our bodies at birth, put on our license plates, be our bank account number, employee number, etc., all in the name of efficiency. Have you noticed an increased use of your Social Security number lately? Many of us have.

* Presented at the Fourth Names Institute held May 1, 1965, at Fairleigh Dickinson University, Madison, N. J.

Let us back away from this thought for a moment and consider, "what is a name and why is it useful?" Names have been used since prehistoric times for the purpose of tagging or identifying places, things or people. These are often characterized as place names, non-personal names and personal names. Each of these categories of names has had a wealth of history in its development.

The origins of place names have been based on many things, such as natural or man-made features, battles, individuals, and local activity. Place names have been grouped many ways. Throughout history, the major grouping of place names has resulted from political divisions mainly in the form of countries. Often the grouping is based on ethnic or language divisions. This is the demographic side of the picture, from the descriptional viewpoint. However, place names can often be grouped in other ways. This may be by organization: for example, the places where a given organization has branch offices or places where a military organization has bases. Yet a third way of grouping place names employs certain methods for accounting or is justified by statistical reasons, such as a division for the purposes of levying taxes or for taking a census.

But grouping has not been the major problem. Standardization of place names and transliteration systems have been the big challenge in this field. Because of these difficult problems, some have suggested that each place name be given a grid coordinate designation to the degree of specificity required in each case. This unquestionably would in theory be an unambiguous method to designate the position of any place on earth. But, alas, the argument would only shift from the arena of place names to the question of which of the bewildering number of presently available grid systems and scales be used. Even more important is the fact that no present grid system, or locating technique, can tell us precisely enough where most places on earth really are. We are much further from this than we are from the acceptance of standard weights and measures, which is taking us generations to achieve.

One organization in the United States has been wrestling with these problems for many years. This is the Board on Geographic Names (BGN), established by executive order in 1890, representing various interested departments of the United States Government, but under the jurisdiction of the Department of the Interior. It had 25,000 names on file in 1943 (10 per cent foreign). Now there are

over 3 million names filed alphabetically by geographic areas. There is an annual increase of $^3/_4$ of a million foreign names and a few hundred domestic names. The Board has estimated that there are over three billion place names in the world, enough to keep its 75 people busy cataloging them for 14,000 years at the present rate. That is a monumental backlog – a real challenge to our Society. The Board has rendered tens of thousands of decisions on disputed place names. Often, in the interest of efficiency and uniformity, the Board's decisions have run contrary to the interests of other departments, such as that of the Post Office, or to local usage (though in principle that is honored by the Board). The Board was in conflict, for example, with Pittsburgh, Pa., for its edict that all *burghs* be spelled without the *h*, until 1911 when the Board reversed itself and agreed that *that Pittsburgh* could have its *h* back.

A more recent conflict in which the Board finds itself enmeshed is the Russian-to-English transliteration problems, which of course transcend *place* naming. This problem centers on how to Romanize the Russian Cyrillic characters. In the 1940's, various interested groups were debating the relative merits of four systems: the BGN, LC (Library of Congress), ACLS (American Council of Learned Societies), and PCGN (Permanent Committee on Geographic Names for British Use). In 1947 the first and last were merged. Now, in the 1960's interested groups are debating the relative merits of five systems (BGN, NSA, BSI, LC, ISO) with the machine people arguing for a sixth one which is completely reversible, character for character (as none of the above are) and thus more amenable to machine processing. There is no settlement in sight.

Non-personal names can be put in several classes. The first includes names of non-living things, plants, and animals. The second class includes those words used to describe various types of institutions: educational, corporate, govermental, religious, and so forth. These types of non-personal names become a significant problem when they are commingled with large personal listings, files or indexes. Indeed, many of these utilize what might normally be considered a personal name. The third class includes a highly stylized system of trade names, including trademarks. Here we could enter a tangle of legal and business practices.

Let me now turn to a consideration of *personal* names. Reflect a moment on how the natural languages have developed in our world.

There are more than 2,000 languages divided into ar least seventeen major groups. Four of these groups include languages that approximately 90 per cent of the world's population speak. These four include the Indo-European (about 50 languages), Hamite-Semitic (14 languages), Sino-Tibetan (four languages), and the Japanese-Korean (two languages). Untold thousands, perhaps millions of people throughout history, have spent their lives studying, organizing, translating, transcribing, explaining, writing about, or teaching these languages. The complexity staggers the imagination. But to keep from staggering ours too much, let us confine our attention essentially to a consideration of the handling of large numbers of personal names found in the English language. I will speak briefly to four major topics: name structure, the manipulation of name data, the impact and methodology of automation in this field and, lastly, some predictions of our coming name handling systems.

I mean by "name structure" the name itself, its historical development, its elements and such factors as frequency of occurrence of individual names among large files of names. Let us look for a moment at the history of names. In prehistoric times, a single name element came into common usage for the purpose of individual identification among small groups of people. By the time the Romans were in charge of the world, it had become the custom for the Roman citizens to be given three element names: a prenomen, the person's individual name; the nomen, the individual's family name; and the agnomen, which was the name element related to the individual's achievements of character. This practice fell into disuse during the middle ages, but became prevalent again in the last several centuries in the Western world and it is now quite normal for the average individual to have a three-element name. Of course, we still often have trouble identifying ourselves; witness the number of John Smiths in any telephone book.

It has been found that most name elements, particularly the surname, can be spelled or pronounced in many different ways. For example, the name Burk can be spelled at least ten different ways. These *variants* have developed by historic evolutionary changes: by individuals changing the method of spelling their name, or through a very complex process of transliteration from language to language. Scholars have found that there are indeed etymological relationships between large numbers of surnames. This means that they can

be put into logically related groups. This is now being done by several groups, in preparation for test in machine retrieval systems.

In addition to the name variants, three-element names can be expressed four different ways with a given spelling of the surname, for example: R. A. Jones, Robert A. Jones, R. Alfred Jones, and Robert Alfred Jones. These have been characterized as name variations in contrast to variants, described above. To compound the situation further, surname elements have such vexing properties as prefixes, hyphens, apostrophes, and articles.

It has been determined that in a typical large number of American names such as might be found in any large city's telephone directory, the most common names like Smith, Jones and Williams each will be approximately one per cent of the total, but never over two per cent. The same general percentages are found to hold for most given names. Tests have also been conducted to determine how many unique surnames would be found in a large number of surnames. One such test revealed that there were about 300,000 *unique* surnames in a file of 3½ million names. About 100,000 of these formed unique single name *groups* and the other 200,000 formed into approximately 50,000 name *groups* averaging about four name variants per group. It has been estimated that in a typical very large sample of American names, e.g., 10 or 20 million, we would probably find about 400,000 to 500,000 unique surnames which could further be reduced to probably fewer than 200,000 discrete surname *groups*. The importance of this point becomes apparent when one considers the automation of name searching activity in large indexes.

An interesting case recently occurred in Sweden. There were many complaints from the Swedish populace about so many people having the same last name. The Swedish Government set up a commission which decided to utilize a computer to generate additional surnames. They took approximately 2,000 typical first and second or multiple syllables, and combined these by a computer program and produced 900,000 new unique names. These have been provided to the Swedish people who now can, by a simple legal action, obtain a new and unique name. Even discounting such spurious generation of surnames, there are indeed hundreds of thousands of basic unique surnames in each of the many languages used in the world today.

But the main point is that there is an asymptotically *finite* number which is nowhere near the number of the world's population. Thus, in view of today's technology, automation of searching in large name indexes is a tractable problem.

Let me now discuss briefly some classification, storage and retrieval methods. The most popular way that names are classified today is alphabetically. It is rather astonishing to discover, however, that because of such problems alluded to earlier, caused by oriental as well as other transliterated names, prefixes, double elements, hyphens, numbers, apostrophes, abbreviations, articles, and so forth, that there is little standardization in this field. I have found no major groups in either government or industry which have agreed to a standard set of rules on how to alphabetize large listings of names. Of course this lack of standardization has not been a particularly critical problem until recent years, since most systems were "closed," in that they did not relate to other systems sufficiently to require standardization of name rules.

A second way to classify names, and one which has become popular in recent years is the use of alphanumeric codes. This, in effect, is a character compression technique and a popular example of this is known as "soundex." This system is presently being used by the Social Security System, Immigration & Naturalization Service, the Maryland and Illinois Driver Registration Programs, as well as numerous private organizations. The soundex technique retains the first letter of the surname as the initial code character, assigns numeric codes to certain consonants, drops all vowels and certain consonants, and assigns numeric codes to the first three numerically codable characters. Thus, the soundex code puts all surnames into a four-character alphanumeric code. For example, Burk spelled any way comes out as B620. This provides a 26,000 code system into which any and all surnames are placed. However, long names with more than four elements to be coded, but with the same first four coded elements, are indistinguishable. This is a natural result, since as explained earlier, there are close to 200,000 discrete English surname groups. For this and other reasons caused by the arbitrary non-semantic nature of this type name coding system, some unrelated names are put in the same groups and some related names are put in different groups. Thus the system does produce a significant number of file drops and nonretrievals.

Another scheme utilizes a purely phonetic criterion to group names with similar sounds. When a system is oriented toward legal requirements, this technique becomes useful because of the legal principle known as *idem sonans* which states essentially that any two names sounding alike are legally the same name. This system provides woefully insufficient criteria for the easy, positive identification of a specific name in large name groupings as they are found in the real world.

A more sophisticated approach is one which pragmatically places surnames into etymologically related groups. This is a difficult, expensive and time consuming operation, particularly in view of the transliteration problems. However, when this is done by competent linguists and if the results are made available to all interested groups, a dependable name grouping system with minimum drop-outs or look-up errors will be available. Some preliminary work is presently being done in this area by several groups in the United States.

Let us turn a moment now to the various methods utilized today to store large numbers of names. The two most common by far are alphabetical book listings (for example, telephone books) and index records. Some holdings are found in the form of dossiers or on microfilm. One interesting storage technique is utilized by the New York Telephone Company to answer quickly a telephone request for the telephone number of a subscriber. This device is a large console at which an operator sits. Within arm's reach of the operator are many small pigeon holes, each with a stick containing a strip of microfilm of several pages of the telephone directory. Each stick can be removed and placed rapidly into a rack which will allow the operator to choose rapidly one of the pages to be displayed on a microfilm viewer directly in front of the operator. Thus the operator can pull down the BAB to the BAS stick, insert the stick into the viewer, locate on the microfilm viewer all the Robert Barkers, select from those the one being asked about and provide the telephone number all in less than five seconds. This has proved to be much faster than using the telephone book.

There are, of course, millions of retrievals being made by humans in book listings every day and very little can be done to increase the efficiency of the average look-up of this type. The millions of look-ups being performed daily in large *index card* holdings, however, are

a different matter. Studies have revealed that it costs an organization between 25 cents to $1 for each index search performed. The variation here is great because the kind of look-up to be made can vary from a simple single identical name search to one involving other associated data, like date of birth, as a checking tool. It is this type of index look-up operation that is being subjected to critical analysis and testing, using computer techniques and equipment.

Some work has already been done in testing the automatic look-up of names among large name files and major problems in this area have been identified. First, it is evident that there is a tremendous cost that must be borne by the organization to convert the existing records to machine language. In a large index, say one of ten million index cards, this would amount to about a million dollars, requiring about 200 man years of editing, keypunching and verifying. Print readers may be used to convert some of the "cleaner" indexes, thereby reducing the cost somewhat. The second big problem to be faced is the choice of the best search strategy. The two major types are letter-by-letter lookup against exact names and the utilization of grouping techniques as described earlier. A third problem is the choice of the method of file storage and look-up techniques. The two major methods are serial versus a random-access type look-up. Each of these presents economic problems as well as other advantages and disadvantages. Enough has been done in this area to indicate, however, that existing off-the-shelf computer hardware is capable of performing a large number of name look-ups against large holdings of names with significant increases in accuracy, flexibility, speed, and a lowering of the total cost.

An example is the Drivers License Registration Service in the Department of Commerce. The purpose of the system is to prevent an individual who has been convicted of manslaughter or drunken driving in one state from successfully obtaining a new driver's license in some other state after his original license has been revoked. All of the states are now contributing to or using the system. The data base consists of that information provided by those states that have submitted data on appropriate individuals. The system started in 1962 with an IBM 1410-1401 but is now using an IBM 7010 and two 1401's. The total number of people involved in this program including operators, programmers, system design people, supervisors, and executives is 25. At the present time there

are roughly 560,000 individuals' names in the magnetic tape file. Including "generated" names (this means name variants, name variations, pseudonyms, etc.), the actual total number of names on the tape is about 1.5 million. The system is growing rather rapidly with a present rate of about 15,000 inquiries a day. The system is now designated to handle up to approximately 160,000 inquiries a day. On some days, over 100,000 inquiries have been processed. The normal schedule is that all inquiries received during a weekday are batched and run during the night shift. The previous day's requests are then ready for distribution in the morning mail to the various states that submitted the inquiries.

It is certain that in the next few years, many of the existing large personal name systems will be redesigned to take advantage of the new powerful tool of automation. I foresee networks of name (or number) data banks in such systems as credit organizations, hospitals, land title companies, driver registration records, and government records. Much work remains to be done, however, in the area of standardization, articulation and design of the best search strategies, file structures, equipment configuration, and conversion methods. Because I am confident that this work will be done with the encouragement and understanding of those who are interested in names, I have no hestitation to predict that our descendants indeed always will have names.

American University
Center for Technology and Administration

CORRECTIONS

Please make the following corrections in the December, 1966, membership list:

Harry L. Levy, 345 E. 69th St., New York, N.Y. 10021
(Our metathesis was showing. We listed the address as 96th St.)

Maurice A. Mook, P.O. Box 25, Boalsburg, Pa. 16827
(The zip code has now become a major problem. We listed it 16837)