Mechanizing a Large Index

By M. A. Wright*

This paper discusses the problems arising from the mechanization of a very large index of names, with particular emphasis on the difficulties arising from incorrect details in inquiries.

Introduction

A typical index contains a list of headings annoted with directions to enable the inquirer to obtain the information he wants. Unfortunately, the headings used in indexes are often contentious; consequently the inquirer often does not know how to translate his requirements into the headings used in the index. This difficulty arises particularly in libraries which use conventional tree-classification indexing systems, and, to a much less extent when "joint attribute" indexing systems are used (see Shera, Kent and Perry, 1957). The difficulty also arises in office indexes because the inquiries often contain errors or use alternative names.

In all types of index the human inquirer has to attempt to translate his inquiry into the index headings under which the information he wants is filed. This paper describes a method of automating a large office index which uses what is essentially a tree-type classification, and in which the inquiries never require more than a relatively simple translation. The example discussed is the index of names of insured persons held by the Ministry of Pensions and National Insurance (MPNI).

The Index of Names at the Ministry of Pensions and National Insurance

Our system of national insurance provides among other benefits, pensions for old and disabled persons and payments for sick persons. Every employed person pays a contribution to the MPNI fund and is given a NI number. Records of the contributions paid by each of the 34 million insured persons are kept at a central office; they are filed in order of NI number. Benefits are only paid if the claimant fulfills certain conditions and has paid sufficient contributions. Consequently, claims for benefit are sent to the central office to check that the requirements are satisfied. The contribution documents and a large number of other inquiries are also sent to the central office.

A large number of documents and inquiries arrive at the central office without a NI number or with a wrong number, but they each contain the name of the person making the inquiry and some other information. Every person's NI number can be obtained, from the name of the insured person, by the use of an index. Such an index is kept at the central office, and it is operated by over 300 staff.

A typical index record contains christian names and surname, together with NI number and sufficient other information to identify the person. This other information is of the kind which, it is known, is likely to be quoted in inquiries. We define the potential contents of each record as

$$(a, b, c, d \ldots J, N)$$

where a is a characteristic representing surname and, for example, a_r is a particular surname, b is the christian name characteristic, c is the birthdate characteristic, $d \ldots J$ are other characteristics and N is the NI number. There are 10 potential characteristics, but only the a characteristic is contained in all records.

There are a number of records containing any specified characteristics; for example, there are over 500,000 persons with surname SMITH. Table 1 shows the population of groups of surnames.

Table 1

NUMBER OF RECORDS WITH THE SAME SURNAME	NUMBER OF SURNAMES	POPULATION IN MILLIONS
20,000–500,000	150	10
5,000– 20,000	1,000	8
500– 5,000	6,000	8
1– 500	300,000	8

All inquiries contain the a characteristic, but only a few of the others. Thus a typical inquiry may contain

$$(a_p, b_q, d_s, e_t),$$

and its counterpart record may contain $(a_p, b_q, c_r, e_t, f_u, N_x)$. In this example, the information contained in (a_p, b_q, e_t) must be sufficient to identify the record.

A proportion of the inquiries are not traced (i.e. their counterpart record cannot be found), because the person is not insured. Others are not traced, because insufficient information has been supplied or because there are errors in the records or inquiries. However, by their diligent efforts, the clerks are able to trace many inquiries, despite errors. Table 2 gives some details of errors and omissions in a sample of inquiries.

Errors may occur due to three main causes.

(*a*) A characteristic may be inaccurate because one isolated character is wrong, e.g. one letter is wrong in an otherwise correct word. Such an inaccuracy may be completely irrational (i.e. a transcription error in which no relationship can be traced) or the result of misunderstanding or misreading. In these instances a relationship, perhaps obscure, can be traced between the error and the original

* Now with N.R.D.C., London, W.1.

Table 2

Inquiries traced	
with birthdate correct	50% of total
with birthdate wrong	2% of total
without birthdate	20% of total
Inquiries not traced	
those which contained insufficient information	2% of total
those which it was expected would not have an NI number	11% of total
others	15% of total

character. For example, if the letter A is badly written it can be confused with H, R or B.

(*b*) A characteristic may be inaccurate because a whole word is wrong. Again these inaccuracies may be either completely irrational, or due to a misunderstanding. Examples of such "errors" have arisen through the use in an inquiry of a shortened or pet name instead of the name on the index record; or through the use of a different spelling. For example, alternatives for the christian name *Alexander* are *Alec*, *Alex*, *Aleck* and *Sandy*; and alternatives for the surname *Shepherd* include *Sheperd*, *Shepard*, *Shepperd*, and *Shepheard*.

(*c*) Finally, a characteristic may be wrong because the correct significance may not be given to a part or parts of the description. For example, there may be confusion between surname and forename, so that *Thomas James* may be quoted as *James Thomas*.

Particularly difficult problems may arise where all three types of error co-exist. This problem arises frequently in inquiries from foreign seamen. For example, *Abdul Abdulla* may sometimes write his name as *Abdula Abdul*. Either name may be the surname and either or both may be spelt in a variety of ways according to the whim of the person writing them. This is especially likely to occur if the man himself is unable to write.

The average number of all inquiries is about 25,000 per day and the peak number is about 50,000 per day. Inquiries are of three kinds: to find NI numbers, to modify records, and to add records. The first operation on all types of inquiry is to find the counterpart record or to check that there is no counterpart record. The number of inquiries which cause a change of the index records is about 40% of the total.

A Survey of Techniques

At first we shall assume that there are 35 million records in the index, that each has ten potential characteristics, and that inquiries quote any few characteristics. It would appear at first thought, that the ideal, although at present impractical, form of storage for such an index would be a ten dimensional store, with records stored in order of one characteristic in each dimension. With this type of store, the characteristics quoted in any reference could select the appropriate dimensions and then trace the record (or records) at the intersection. But each record must be potentially capable of being identified by only a few of the ten characteristics. Therefore, it is essential that there is redundancy in the record.

A multidimensional store has to provide space for this redundancy. For example, if the characteristics could potentially represent n different records, then a multidimensional store has to provide a one bit store for all n records: it is estimated that in the MPNI index $n \approx 10^{40}$. If there are only m records, each described by a number of characteristics then only $m \log n$ bits need be stored. In the MPNI index $m \log n \approx 10^{10}$ bits. Thus a multidimensional store would be grossly uneconomic of storage.

There is a method of realizing many of the effects of a multidimensional storage system. This method is applied in various systems, one of which is known as the *Peek-a-boo* system (Shera, et al., 1957). The Peek-a-boo, and other similar systems, provide cheap storage on punched cards, a method of comparing combinations of characteristics, and a method of presenting the result of each comparison. A scanning system has to be provided to scan the results of the comparisons and to select the matching record. The result of the selection is a position in a sequence which provides a straightforward clue to the remainder of the record, e.g. the NI number. Human beings can scan the results, read the sequential position and find the NI number from a file: it would be expensive to perform these operations automatically. Furthermore, it would be impractical to add, delete and modify the characteristics of records.

An alternative would be to store the index records in some fixed order and search systematically to find records which matched inquiries. This systematic searching is only practical, in the present context, if the inquiries can be more or less directed to their counterpart records. This would involve storing like records together. One method of doing this would be to attempt to arrange the records in a hierarchy. For example, the records could be stored in groups according to the a characteristic, the records in each group could be arranged in subgroups according to a second characteristic, and within the subgroups the records could be arranged in sub-subgroups according to a third characteristic, and so on. It would be possible to do this only if sufficient characteristics were quoted.

The record matching an inquiry could be found by selecting the appropriate group, subgroup, sub-subgroup and so on until the number of records in the sub-subgroup was sufficiently small that it was practical to search them all. Alternatively, if the number of records in the sub-subgroup could be reduced to one, the record would be pinpointed. This method requires that inquiries, as well as records, should quote the characteristics used in grouping.

If a parallel store is used in this fashion, any individual inquiry could be answered "immediately," and inquiries

could be searched for in any sequence of arrival. However, the main disadvantage of parallel stores for the purposes of the MPNI index is that the cost of such stores with sufficiently fast access is high, or the cost of deleting and adding records is high.

It should be noted that access to the store need not be made in random order. It is well known that a serial store can be used efficiently for this type of work, provided that a group of inquiries are processed at one time. If we had to trace an inquiry in a serial store we would have to search serially through the store to find the appropriate sub-subgroup. But if we wanted to trace a number of inquiries they could first be sorted to the same sequence as the sub-subgroups. The first inquiry could then search the store in sequence until its counterpart record was found. While that search was progressing, none of the records required for the other inquiries could be missed: a search could be made in a similar manner for all inquiries of the batch. Thus a number of inquiries could be answered in a single processing of the index records.

However, the full information to select the groups, subgroups, etc., may not be available. If there is lack of information to enable, for example, the sub-subgroup to be selected then the inquiry could be traced by searching all records within the subgroup. This involves more searching. Also, if some information is wrong, it may not be possible correctly to sequence the inquiries. Therefore, some records, required by some inquiries, might be missed while the search for others was progressing. The records which were missed could be found by processing the index again, but this processing is expensive and time-consuming. To minimize processing, it is desirable to arrange inquiries in a sequence which is the same or very similar to the index sequence. This could be done by (*a*) using several indexes each stored in a different sequence, (*b*) classifying characteristics, (*c*) using so-called "heuristic" methods to make corrections to the inquiry.

It can be deduced from Table 3 that a combination of only a few characteristics is needed to identify a record, and it has been found that only a few combinations of characteristics are ever quoted by inquiries. Consequently, if system (*a*) were employed, only a few of the very large number of potential sequences of records need be kept.

Half of the inquiries correctly quote one particular combination of characteristics (surname, christian name and birthdate) which is usually sufficient to identify records. An index kept in this sequence would enable half the inquiries to be answered in a single processing.

Indexes kept in other sequences would be successful in answering a smaller proportion of the inquiries in a single processing. But some inquiries quote characteristics only barely sufficient to identify a record and the information in one characteristic may contain a small error. For example, an inquiry may quote surname, christian name and birthdate: there may be a small error in any of the characteristics. It would be impossible to find the counterpart record unless a method which ignored the error were used. Furthermore, there would be uncertainty of finding such a record, unless the inquiry could be directed to a group or subgroup of the index which contained the record. If the characteristics used in the prime grouping contained an error, the inquiry could not be directed to the group which contained the counterpart record, unless the characteristic used for the prime grouping was classified so that characteristics which may be confused one with the other were all grouped together. It is fortunate that in MPNI index, the characteristic most suitable for use in the prime grouping (surname) is also amenable to classification.

Table 3

CHARACTERISTIC	NUMBER OF VALUES USED IN INDEX	POTENTIAL NUMBER OF VALUES	NUMBER OF BITS
(*a*) Surname	300,000	over 10^7	25*
Classified surname	3,000	3,000	(16)
(*b*) 1st christian name (classified)	1,000	1,000	10
(*c*) Birthdate	30,000	30,000	16
(*d*) 2nd christian name (classified)	1,000	1,000	10
(*e*) Address	10^7	over 10^{10}	108*
(*f*) H.M. Forces number	10^6	10^8	32
(*g*) Married women's maiden name	300,000	over 10^7	25*
(*h*) Title	3	3	2
(*j*) Pension number	2×10^6	10^8	32
(*k*) N.I. number	30×10^6	10^{10}	42
Total			302

* Allowing only 5 characters for names.

If surnames were classified according to the *Soundex* code (Appendix I) and four symbols were used, 99% of misgroupings would be corrected. This would appear to be a very satisfactory practical proposition. But this aspect will be investigated further, because it is desired to show how machines could provide a service which is at least equal to the present service.

The Soundex coding system is a classification system based on phonetic spelling. Its object is to give the same code number to names which sound similar. It is thought, however, that most of the errors in names in the MPNI index are due to bad writing. If this were true, it would seem desirable to use a coding system based on the similarities of written letters; however it has not been possible to devise a suitable code.

An alternative system involves the use of a surname index or dictionary which shows the interrelationships of names. Such an index could be arranged by classifying names according to a rigid system and annotating a suitable code number to each group of names. All

known classification systems have anomalies, some of which are known to exist (others are unknown at any given time). Information about the known anomalies could be incorporated in the index. For example, surname HUMPHREY could be annotated with its classification code number, and also with the code numbers of other names with which the name may be confused. The system could also be applied to other proper names. If it were applied to christian names, then pairs of names, like Sandy and Alexander, which have the same meaning, could be given the same code number.

The anomalies referred to above were due to misspelling or the use of alternative names which have similar meaning. It may be practical to use a dictionary system to take account of relatively frequently occurring anomalies, but it would be impractical to incorporate all possible errors in a dictionary. A clerk is able to trace records, despite errors in the inquiry, by a series of "well judged" searches. Methods for conducting these searches are sometimes referred to as *heuristic* methods. They could be used as an adjunct to, rather than a replacement for, classification methods.

A method of applying heuristic methods would be to formulate probabilistic rules for construction of the words used in characteristics, and to manufacture "secondary" inquiries based on the original inquiry, but quoting new characteristics. A disadvantage of this method is that such a large number of "secondary" inquiries could be manufactured that the probabilistic rules would, in practice, be almost useless (except in certain special circumstances). Consequently, in general, some clue as to the kind of error that may have been made in an inquiry would have to be found. Such clues are sometimes available. For example, a badly formed letter on the inquiry would be more likely to be misread than a well formed one; and a birthdate, if it is wrong, is known to be probably incorrect in only one or two figures. Furthermore, certain *ad hoc* correlations must be satisfied by the inquiry. For example, there are correlations between age and marital status, and between age and a prefix of the NI number. It has been estimated that it would be relatively easy to formulate these correlations and to list the contexts in which they apply. If these correlations were not satisfied, they would identify characteristics which might contain errors. Even so, each characteristic has such a large number of potential values that it may be impractical to manufacture sufficient inquiries. For example, it may be suspected that the birthdate is wrong in a particular inquiry. Birthdate is usually quoted to show day, month and year of birth, e.g. 27.10.1933. Any of the figures may be wrong, and there are 45 dates which are different from the one quoted in only one figure (ignoring errors in the first two figures of the year). Alternatively, the error may be due to interchange of two or more letters; there are 28 ways of interchanging pairs of letters.

In some inquiries the significance of characteristics may be confused, e.g. surname and christian names interchanged or first and second initials interchanged. Inquiries with the former type of confusion may never be directed to the right group of the index unless a secondary inquiry is manufactured with the names interchanged. However, it is certain that some surnames are never used as christian names, and so the manufacture of secondary inquiries to cope with this deficiency is practicable.

If a figure or letter of a characteristic was obliterated, the significance of the succeeding letters or figures would be wrong. Thus, if a record quoted 17 Park Lane, Kempton, as address, and the inquiry quoted the same address, but with the K obliterated, then the two would compare identically up to the K, but thereafter they would disagree. A very large number of alternative addresses could be manufactured from empirical rules. Alternatively, it could be argued that a classification system could be devised to provide both KEMPTON and EMPTON with the same code number. Such a system would have to be very complex to take all such errors into account, but the absence of a town called EMPTON in a dictionary of towns could give a clue to the error. A method, in which sequence of characteristics was adjusted, could be employed, but there are a large number of ways of adjusting the sequence. One way of finding how to adjust it would be to attempt to align one letter, chosen at random, and to arrange the remainder according to this. However, the error quoted involves a change of sequence and a single-letter error. Even if the former error were corrected, allowance would have to be made for the latter. This leads to a method which makes use of the redundant information in characteristics. Arrangements could be made to accept a characteristic as a match, if only one letter (or perhaps two letters) in the characteristic quoted in an inquiry differed from the record. The record thus found could be subjected to scrutiny (automatically) to test whether the letter which differed had special significance. The test would require the use of either a set of rules or a table of known types of confusion, but such a table might be comparatively small. This method could be extended to allow selection of a record which partially matched the inquiry in sufficient characteristics, so that only one record was selected. The number of records in each surname group or surname-christian-name group could be stored in the index and used to indicate how much information, from other characteristics, was needed to select one record.

Some construction rules and tables of confusion are applicable only in some contexts. The contexts must thus be specified. If the rules and tables were applied when the contexts were similar as well as identical, the relative similarity of contexts would have to be measured. The method used to make classification systems could be used for this purpose, but this, and any other method, would be complex to apply to measurement of context. Information on confusions could be applied only when contexts were identical. This could be simply arranged if the differences between an inquiry and its counterpart record were stored in the index. This would enable an inquiry to be matched against a previous quotation of

inquiry, which may have contained further information, e.g. an address which matched the record. Thus a first inquiry quoting *John Shepherd* with given birthdate and address might match a record quoting *John Sheperd* with the same address but a slightly different birthdate. A subsequent inquiry might quote *John Sheperd* and the same birthdate as the previous reference, but no address.

If the results, obtained from searching and from the manufacturing of secondary inquiries, were analysed, it would be possible to modify the probabilities of success of the various methods of manufacturing secondary inquiries. This would result in a learning process, but, in general, it would be difficult to find rules because of the difficulties of defining contexts.

Practical Aspects

Table 3 shows that the information potentially available on each record amounts to about 300 bits. Thus the total potential information in the index is about 10^{10} bits. The only storage medium which satisfies the main requirements of the index is a serial store, and the most practical serial store is magnetic tape. At present it is practical to store 500 bits per linear inch of magnetic tape. It is also practical to arrange 6 information tracks on tape 0·5 in wide. Therefore 10^{10} bits could be accommodated on about 120 reels of tape each 3,600 feet long, assuming no blank spaces. It is practical to run this tape at 160 in/sec. Thus the 120 reels could be read in about 10 hours including an allowance for reel changing. It is probable that a reading time of 10 hours would provide a rather longer service time for answering inquiries than is desirable. Consequently two or three reels would need to be read concurrently. The cost of two or three reading mechanisms and 120 reels of tape would be comparatively small, and thus the storage system is economically feasible; it is, therefore, worth investigating a complete system employing magnetic tape.

Various techniques have been described to show how the record matching an inquiry could be found. It will be noted that

(1) a hierarchical system using sub-subgroups would enable the record which matches an inquiry to be pinpointed provided that the record and inquiry contain the characteristics used for sub-sub-grouping;
(2) the above system (1), combined with a classification system which classified like names together, would enable the record which matches an inquiry to be pinpointed when the names contain certain known types of error;
(3) the above system (2) would enable the location of the record which matched an inquiry to be reduced to a group or subgroup, provided that the inquiry contained the characteristic used for grouping or subgrouping. The record could be found by searching this group or subgroup for records which were identical to the inquiry in other characteristics;
(4) the system (3) above combined with a heuristic system would enable the records which matched an inquiry to be found by searching, when there are errors in any of the characteristics.

The amount of data processing that has to be done to find the records matching a large batch of inquiries is least in systems (1) and (2) and is most in (4). All inquiries could be answered by system (4) after sufficient processing provided sufficient information was supplied; but half of the inquiries could be answered by system (1) (which is more efficient). The four systems could be combined by attempting to answer all inquiries by system (2) and resorting to the more inefficient systems if it is not successful, provided that an indication can be obtained when a system has failed.

The mechanism for obtaining a sub-subgroup in a serial storage system is by searching for it and, if an inquiry quotes sub-subgroup (a_p, b_q, c_s) which does not exist, then there is a possibility of searching too hard for one inquiry and consequently missing the next. The sub-subgroups $(a, b, c)_r$ of the index would be stored in sequence. When the record, or the place where the record would be if it existed, is passed, the sign of the difference $(a, b, c)_r - (a_p, b_q, c_s)$ changes. This can be used to terminate the search for (a_p, b_q, c_s) using system (1) or (2). Therefore it is possible to attempt to answer a batch of inquiries by comparing inquiries with the index and changing the inquiry when the sign of $(a, b, c)_r - (a_p, b_q, c_s)$ changes, i.e. using system (1) or (2). Half of the MPNI inquiries would be answered by this system. The remaining inquiries could be reprocessed, but inquiries changed when the sign of $(a, b)_r - (a_p, b_q)$ changed. When two or more inquiries were made to one subgroup, the records in the subgroup would be processed once to answer each inquiry. Similar arrangements could be made to search the a groups for the inquiries not answered by searching subgroups. However, 28% of the inquiries are not traced by present methods.

The average density of inquiries would be about 10^{-2} per subgroup per day at peak times, but the peak density would be about 12 per subgroup per day to the largest subgroup. We shall now assume that the index can be represented by a simple model with 150 p subgroups of records with 4,000 records in each, 1,000 q subgroups with 1,300 records, 10,000 r subgroups with 500 records, and a large number of other smaller subgroups. The average density of inquiries to the p subgroups is 3 per day (assuming 50,000 inquiries per day of which 50% quote correct sub-subgroup characteristics). The average density of inquiries is 1 to the q subgroups and 0·35 to the r subgroups. If we assume that, in order to process a second inquiry to a subgroup, the magnetic tape has to be wound back and then played forward again, that the forward and backward speeds are the same and constant, and that the reversal times are negligible, then

the processing time is $(2n - 1)t$, where n is the number of inquiries per batch and t is the time taken to process the records once. Thus the total processing time $T = \Sigma(2n - 1)t + pt$, where p is the number of batches which receive no inquiries. The efficiency of the system can be measured by the ratio t/T. This ratio is unity when the index does not have to be run back and is about $1/(2n - 1)$ when n is large. It has been calculated that with a Poisson distribution of inquiries the ratio

$t/T = 0{\cdot}2$ for densities of 3 inquiries per subgroup
$= 0{\cdot}6$ for densities of 1 inquiry per subgroup
$= 0{\cdot}9$ for densities of $0{\cdot}35$ inquiry per subgroup.

Thus the total efficiency in our model is $0{\cdot}92$ allowing for searching subgroups. After this searching, the remaining unanswered inquiries would presumably comprise the 26% which are thought to have no counterpart record, plus a few other inquiries. It would be possible to segregate the 11% of inquiries for which counterpart records are not expected to exist, but it would not be possible to segregate the 15% of "others" (see Table 2) from the remaining inquiries for which records exist. Thus more than 15% of inquiries would have to be subjected to subsequent group searching, if this further processing were desired. To show the efficiency of group searching, we use the previous model, except that there are 10 times as many records in each group and densities of inquiries to groups are three times those to subgroups. The efficiency of group searching would be $0{\cdot}2$. However, if searching of the p groups was not included, the efficiency of group searching the remainder would be $0{\cdot}7$.

If the procedure of manufacturing secondary inquiries is adopted, the subgroup searching density would be increased. Thus, it is undesirable to manufacture more secondary inquiries than the number of subgroups per group. It must be noted also that the manufacture of secondary inquiries requires additional data processing, as does the sorting of secondary inquiries and editing of them to produce the best answers. The number of data-processing operations involved in the two sorting stages alone is $2qn \log_q n$ where q is the number of secondary inquiries manufactured for each inquiry and n is the number of original inquiries. It is evident that this should not be much bigger than the number m of operations involved in inspecting the index, i.e. the number of records in the index. Now $n_{max} \approx 10^4$ (allowing for 20% of inquiries) and $m/n \approx 3 \times 10^3$. Therefore $q_{max} \approx 70$ at times of peak activity, assuming the sorting is done by 2-way merging. No attempt has been made to estimate the probability of success of the secondary-inquiry method in a prescribed number of attempts, and it is therefore impossible to estimate the desirability of using this method instead of group searching.

It should be noted that 10^{10} bits of information on the index allows for considerable redundancy. There is redundancy, even if records have to be identified by surname and only one other characteristic. For example, classified surname divides the index into 3,000 groups assuming 4-symbol Soundex code. Thus, if the distribution of characteristics was uniform, information in the other characteristic could be restricted to 14 bits: the index would then contain only 6×10^9 bits. Furthermore, if surnames and christian names are recorded only at the head of each subgroup, the information stored could be reduced to 5×10^9 bits.

One of the alternative methods selects a record as a match even when some letters of characteristics are different, and then checks to see whether any such differences are significant. If this method is employed, there must be redundancy in characteristics. On the basis that surname plus one other characteristic can identify one record, only two characteristics contain much redundancy (see Table 3). If the redundancy in these were reduced, so that each characteristic contained only 18 bits, then the index would be 8% bigger than it would be if the maximum size of characteristic were 14 bits.

A complete index system involves several stages of processing.

For example, one system contains the following steps.

(1) Prepare inquiries in a form suitable for the machine, and number them.
(2) Sort inquiries to order of surname.
(3) Compare inquiries with surname dictionary, and record classification code numbers with inquiries.
(4) Sort inquiries to order of first christian name.
(5) Compare inquiries with christian name dictionary, and record classification code numbers with inquiries.
(6) Repeat 4 and 5 for second christian name.
(7)
(8) Outsort inquiries quoting married woman's maiden name, and repeat steps 2 and 3 for those quoting married woman's maiden name.
(9) Prepare inquiries in standard layout, and sort to order of classified surname, classified christian name and birthdate.
(10) Compare with index (see next paragraph). (Index kept in order of classified surname, classified 1st christian name and birthdate.)
(11) Extract inquiries answered in step 10.
(12) Compare remaining inquiries with index (see below).
(13) Edit results.
(14) Extract inquiries answered in steps 11 and 12, and also extract inquiries for which no trace is expected.
(15) Generate up to 50 secondary inquiries.
(16) Compare with index (see below).
(17) Edit results.
(18) Sort all answers to order of numbering used in step 1.
(19) Print results.

A sequence of inquiries might be

$$(a_p\ b_f\ c_m),\quad (a_p\ b_f\ c_n),\quad (a_p\ b_g\ c_p),\quad (a_p\ b_n\ c_q),$$

with two inquiries to the subgroup $a_p\ b_f$. When this subgroup was found the index tape would be searched for sub-subgroups. The search for $(a_p\ b_f\ c_m)$ would be discontinued as soon as the record characteristics $a\ b\ c_r > (a_p\ b_f\ c_m)$ (assuming that records were arranged in ascending order), and the search for $(a_p\ b_f\ c_n)$ would begin as soon as it was available. If it was not available until time t_r and the average time interval between finding adjacent inquiries is t_i then the chances of missing a record are t_r/t_i. It is estimated that this would cause about 1% of inquiries to miss their records. The search for the second record would be terminated when $(a\ b)_r > (a_p\ b_f)$. The records in the next subgroup would then be processed.

The above procedure would be followed for all inquiries: the index tapes would be processed once and about half the inquiries would be answered.

The remaining inquiries would be reprocessed, but during this processing (step 12) every inquiry would search at least all the records in the appropriate subgroup. Records which differed in only one character of a characteristic would be accepted as a match. If there were more than one inquiry per subgroup, the index tape would be run back to the beginning of the subgroup, after completing the search set up by the first inquiry. The procedure would be repeated until the subgroup had been searched by all the other inquiries. Processing of the next subgroups would then proceed. If there was only one record in a subgroup this fact would be recorded on the index. Any inquiry to such a subgroup would select, as a match, the only record, even if no other characteristics matched. The records obtained during this processing would be edited. If there was a difference between characters in the inquiry and the record, it would be necessary to ascertain whether or not the characters were significant. The information on types of confusion and, for example, a list of pairs of towns which differed in only one letter, would form the basis of this check. Single-character errors in parts of the address would be allowed, and errors of one or ten years, months or days in birthdate would also be allowed. Also, any error in birthdate would be allowed, if the age given is over 70 years, since it is known that there are often errors in the age quoted by old people, and there are fewer old people.

About 20% of the inquiries would be answered by the above method. It should be noted that steps 10 and 12 could be combined.

The remaining 30% of inquiries would be segregated into those for which a trace was not expected to be found, and the others. The former would be assumed to have no counterpart record. The latter would be used as the basis for secondary inquiries. Secondary inquiries would be generated with the initials and christian names of the original inquiry interchanged. Also, the surname would be tested against a special index to see if it was of the type that may be a christian name. If it was, the surname and christian name would be interchanged. Furthermore, the month and day numbers of birthdate would be interchanged if both numbers were less than 12. The secondary inquiries would be processed in same way as step 12.

A record which matched an inquiry in some particulars, but mismatched in others, may or may not be the counterpart record. For example, a record which matched any inquiry in surname, christian name, and address, but mismatched in birthdate, may give a wrong birthdate, or, alternatively, the record may refer to the father, son, brother or other relative of the person for whom the inquiry was made. In the present system such a record would be used as a "probable answer." This procedure could be continued in an automatic system.

Inquiries, together with their counterpart records, would be sorted back to the same order as the number on the original document (step 1) and the results attached to them.

The equipment needed to do this work includes machines for transcription, sorting, editing, printing, and also for reading and comparing inquiries with the records, and updating records. Existing equipment could do this work, but it may not be very efficient. However, the amount of work involved is not large, except in reading, comparing and copying index records, and in updating records. Computers are efficient for record extraction and updating, if the density of inquiries to records is high. But the density in MPNI index is never likely to be above 0·001. Consequently, it is desirable to use special equipment for this purpose. Some record-extracting devices are at present available, but they do not contain the equipment needed to perform all the operations needed for index work; furthermore, they are expensive. Therefore, it is desirable to develop, if possible, an inexpensive record-extracting device and an updating device.

The devices at present available employ complex magnetic-tape units with high-speed starting and stopping facilities, and large expensive buffer stores. It is proposed that the index magnetic-tape units should be similar to existing audio units, but faster, and with start, stop, and reversal times of about 0·2 sec. If the index data were recorded in large blocks (i.e. about 100 in. each) on six tracks at 500 bits/in., then 60 reels of tape would be needed. If the tapes were read at 100 in./sec the index could be read in about 6 hours on one reading device. It may be economic to use more units, operating at a lower tape speed, with longer start, stop, and reversal times.

The inquiries would be recorded on two or three magnetic tapes, each about 600 feet long, with data stored at the same density as on the index tape, but with gaps between each inquiry. The inquiry tape would be driven by a capstan assembly with high-speed starting and stopping facilities but no reeling facilities. The inquiries would be read, one at a time, into a high-

speed buffer store, and the inquiry in the buffer store would be compared with the index. Answers could be copied from the index tape by a second reading head placed a suitable distance apart from the head from which the data is taken to the comparison units. They would be copied on to a tape similar to the inquiry tape (or perhaps on to the inquiry tape itself).

The equipment would include circuits for

(*a*) detecting the end of a group or subgroup, or detecting when a sub-subgroup is passed,
(*b*) detecting the difference between the characteristics of record and inquiry, and also for ignoring a difference of only one character in a characteristic,
(*c*) detecting a symbol on the inquiry tape calling for a reversal of the tape, and for reversing the tape again when a specified symbol was detected on the index tape.

Conclusions

(1) A magnetic-tape system would provide a practical method of storage for MPNI index.

(2) Development of cheap record-extracting and record-updating equipment is required.

(3) Any number of inquiries, quoting three specified characteristics correctly, could be answered in a single sequential processing of the records, provided that the records were arranged in a hierarchy according to the three characteristics. A few inquiries which did not quote the three characteristics correctly would be answered, and the remainder could be segregated for further processing.

(4) The remaining inquiries which quoted sufficient characteristics correctly could be answered by systematically searching the records. It would only be practical to do this if each search could be directed to the part of the index which contains the counterpart record.

(5) The effect of certain types of error can be eliminated by classifying characteristics. The effect of other types of error could be eliminated by listing anomalies of a classification system. It would be impractical to extend this list to include all possible types of error.

(6) Methods could be employed to guess corrections of inquiries which may contain errors. It would be impractical to make random guesses, but it is possible to predict the possibility of some types of error, and it may be practical to guess corrections of those parts of inquiries which are most likely to contain an error. It would be impractical to make extensive use of this method, and the probability of success of it has not been estimated.

(7) It would be possible to improve the above methods by analysing the answers obtained by searching methods, but the existing knowledge of types of errors in MPNI index is extensive, and, consequently, improvement of this knowledge would involve very complicated analysis. In addition it would be difficult to achieve further improvement because of the difficulty of measuring contexts.

Acknowledgements

This work has been carried out as part of the research programme of the National Physical Laboratory, and is published by permission of the Director.

The help provided by colleagues in the Ministry of Pensions and National Insurance is acknowledged.

Reference

SHERA, J. H., KENT, A., and PERRY, J. W. (1957). *Information Systems in Documentation*, New York: Interscience Publishers Inc., p. 209.

Appendix: The Soundex Code

The Soundex Code is a system of coding names according to their phonetic sounds. The origin of the code is not known and despite extensive search no description of it can be found in scientific literature. However, it is widely used in the U.S.A. and has been reported as being devised by the Remington Rand Corporation.

The method of coding is as follows:

1. The first letter of the name is retained as the first letter of the code.
2. Vowels, W, H, and Y are deleted.
3. The second consonant of a double consonant pair is deleted.
4. The following letters are replaced by numbers:

Letters	*Code*
B, P, F, V	1
C, G, J, K, Q, S, X, Z, SC, CH, SCH, CK	2
D, T	3
L	4
M, N	5
R	6

It is usual to limit the number of symbols to three or four; zeros are added to the code if there are insufficient phonetic sounds.