Record identification using variable alphanumeric names

By D. G. W. Thomas*

This paper considers the problem of using a variable-length alphanumeric name, which may be read into a computer system in a variety of forms, as a key or file reference, and describes a data input program which was written for the English Electric-Leo KDP 10 which successfully employed this technique.

Statement of the problem

There are many potential computer applications where input data for the system can be produced as a by-product of a current typing operation. However, in many cases the information typed is insufficient to enable transactions to be applied to a file because of the absence of a file record number or other file reference. The type of information available to the typist to identify the relevant file record is the name of a customer as printed on a letter heading, bill or invoice, or some other form of printed description.

Such applications could occur in maintaining many sorts of files, particularly with a customer or supplier file, where the source documents originate outside the computer user's organization. It would rarely be satisfactory to rely solely on a reference number supplied by another organization as the computer file record number.

Manual encoding

The usual procedure is for a clerk or typist to look up and insert the relevant file reference, before the typing operation, on the document from which the data are to be prepared. The major disadvantages of this are:

(*a*) The time and cost in looking up a reference number.

(*b*) The considerable risk of looking up the reference number incorrectly or of mistyping it. Thus the look-up and typing operations have to be manually checked, involving further labour costs, or a check digit (such as a modulus eleven check) has to be incorporated in each reference number and checked by the computer.

(*c*) In certain cases it would be inconvenient for a file reference number to appear on the document being typed.

(*d*) It may not be convenient or even possible to have an up-to-date copy of the list of reference numbers available at each document preparation point.

In such circumstances there are many advantages in using a computer to encode the typed alphanumeric information in a way which would enable it to be used as the file reference, and it is this problem which is the subject of this paper.

In this paper, the word "name" means a string of characters including punctuation marks and spaces, which can be of any length, constituting a field of information; e.g. AUSTIN A30; THE KING OF SIAM; THE ENGLISH ELECTRIC CO. LTD.

The computer approach

There are two approaches to the problem. The first is to use the name, either as typed or after computer editing, to form an alphanumeric key and to use this as the file reference. Alternatively, the name, after any necessary editing, can be used to locate a file reference number in a table, which number can then be attached to the transaction record and form the reference key for subsequent computer processing. The latter approach has the advantage that more than one form of a name can be used to refer to the same file record (e.g. ENGLISH ELECTRIC or THE ENGLISH ELECTRIC CO. LTD.).

Typing variations

This brings up the most significant and troublesome feature in using typed names for identification: this is the great variety of ways in which a given name may be typed (including minor mistypings) and *still* be unmistakably recognized by a human reader as the name of the same customer or item. Abbreviations, conjunctions and other words which convey little information are particularly often mis-spelt or omitted.

For instance, consider the (hypothetical but by no means untypical) name:

LA GRANDE COMPAGNIE DE L'AMERIQUE S.A.

In this example a typist may omit the words LA, L', DE, S.A., either individually or in any combination without causing any real doubts about the identity of the company. Also the letters S.A. could be run into one "word" SA.

Just these permutations, ignoring any variation in punctuation and word spacing, give forty alternative forms, all of which convey to the human "processor" the same unique identification.

Although, in order to reduce the number of variations,

**Formerly with the English Electric Company Ltd., Kidsgrove, Staffs.*
*Now with the Rio Tinto Zinc Services Ltd., 6, St. James's Square, London, S.W.*1.

typists could be required to type only one standard form of each name, this has several disadvantages:

(*a*) It is almost impossible to eliminate variations and standardize punctuation and spacing.
(*b*) The error rate would increase and retyping would reduce the output of useful work.
(*c*) More than one form of a name may be in common use and hence equally legitimate.
(*d*) In very many cases the information is being typed from some source document bearing a title supplied from an outside organization, and it is very desirable that the typist is only required to copy such information and not to carry out an editing function as well, with increased risk of error and reduced rate of work.

Thus there are good reasons for considering in what way a computer can be programmed to reduce such variations; the ideal would be that whatever variations on a name were comprehensible to a human reader without ambiguity would be accepted and correctly identified by the computer, and whatever variations were ambiguous to a reader would be rejected by the computer program.

As a first step to reducing this variety a number of techniques for editing the names are considered. In these, manipulations are carried out on the name as typed to produce a string of characters which is called the "edited name". The process is referred to as "editing".

The design of an editing strategy

There are a number of techniques which can be applied to a name to reduce its length and range of variation; these are considered here in order of increasing sophistication. The objective in each case is to reduce the number of different forms of the edited name to a greater extent than the parallel reduction in significant information in the name. In other words to distil the information content of the name from the dross of punctuation, conjunctions and "noise" letters and symbols.

(*a*) Inevitably the first step is to delete all punctuation marks and variations in spacing, producing either a string of letters or words set out in some standard manner, e.g. DOMBEY & SONS, (LONDON), LTD. becomes DOMBEYSONSLONDONLTD or DOMBEY & SONS LONDON LTD
(*b*) Certain short words such as conjunctions and abbreviations usually convey little information and so can well be omitted from the edited name; e.g. LA GRANDE COMPAGNIE DE L'AMERIQUE S.A. becomes GRANDE COMPAGNIE AMERIQUE. The omission procedure can either be to omit all words of a given length, or those words which belong (or do not belong) to a list of selected words. The first alternative is logically simpler and easier to program.
A combination of these procedures can be used, all words of certain lengths being deleted, while words of other lengths are compared against lists of reserved words. Common candidates for omission are THE, AND, LTD, MESSRS, and almost all one- and two-letter words.
The actual choice of methods and of words to be deleted should be dependent on an analysis of all the titles in use in a particular file—and on the degree of security from the editing scheme.
(*c*) The next group of techniques involves manipulation of the characters within the words themselves. The first approach is to take the edited name (a string of characters) produced by one of the previous procedures and truncate it to the required length by dropping characters off the left or right end of the edited name. Whether left or right truncation is employed will depend on whether more information appears to be given by the initial or final letters.
(*d*) More efficient, if more drastic, are techniques involving selective omission of letters from the words themselves. Some letters, notably vowels, occur more frequently than others, and hence provide lower discrimination, i.e. carry less information. Certain letter positions in a word (e.g. the initial letter) have more significance than others; while in English the second letter supplies very little information, very often being a vowel. Thus to omit certain frequently-occurring letters, or letters in certain positions in a word, or every *n*th letter, may provide an efficient technique for abbreviating a name.
A further refinement, which gives a good protection against creating the same edited name from two distinct names, is to use the omitted letters to form a check digit (e.g. by accumulating the binary values of the omitted letters to form a hash total). If the omitted characters were wrongly typed, or if the original names differ only in omitted characters, the edited names will almost invariably have different check digits, and thus still provide a satisfactory degree of security.

For example, let our strategy be to omit all vowels and then to omit selectively every second letter: then "KENSINGTON HIGH STREET." becomes—firstly, by omitting spaces and punctuation:

KENSINGTONHIGHSTREET

secondly, by omitting vowels:

KNSNGTNHGHSTRT

and thirdly, by masking out every other letter in each word:

KSGNHHSR

Objectives of an editing strategy

The objectives of an editing strategy are:
(*a*) To produce from the vast number of acceptable typed names a much smaller number of edited

names which, with their file reference numbers, will be stored in the computer, i.e. a minimal list of minimal-length unique names. From the computer user's point of view the advantages of having fewer and shorter names are:

(1) The reduction in processing and table look-up time.
(2) The reduction in the amount of internal storage required.
(3) The reduction in the size of the file of reference numbers and edited names which has to be kept up to date.

(*b*) To reduce the risk of two different names producing the same edited name to an acceptable level, where the acceptable level of risk is determined by the nature of the application; e.g. a finance house could only accept zero risk of two accounts being confused, whereas a considerable risk of duplication might be acceptable for a stock file in a manufacturing organization.

(*c*) At the same time, to minimize the number of visually identifiable typed names which will lead to the identification of an incorrect or of no file reference.

Comments on editing strategies

It is clear that these objectives are to some extent mutually contradictory, and the aim in designing an editing strategy is, in the light of what is known about the characteristics of the set of names to be processed, to select a technique which achieves a good compromise between these objectives. This requires much study and experimentation on the data, and the strategy chosen will depend heavily on the data characteristics and on personal judgement, and on the particular characteristics of the computer to be used.

Where the input is produced on tape writing machines having facilities for inserting control characters between different fields of information, a better discrimination between edited names may be achieved by applying different editing strategies to different fields within one name. For instance, where the name consists of a title and an address, different strategies could be applied to the title and to the address fields.

The degree of editing which can be undertaken on any type of computer will depend on the machine's facilities for making many-way comparisons and for the rearrangement of characters. In such operations character-addressable machines have a considerable advantage.

A variable-length character machine such as the English Electric-Leo KDP 10, on which the author carried out much experimental work, is particularly suitable for such data manipulation.

Experimental work

This "computer editing" approach arose from a study of the data input problems associated with a large data-processing application, and computer programs were written for KDP 10 which successfully demonstrated the feasibility of this approach. The aim of the input run was to read in transaction records from paper tape, each record bearing a variable-length alphanumeric name from which the account concerned had to be identified, and to form an edited name which was then used to locate an account number in tables of account numbers and edited names held in the core memory.

Characteristics of the data

The alphanumeric names contained two subfields: a customer's title and his address. Both of these were variable-length fields and could exhibit typing variations. There were, for some customers, more than one edited form of their title, and also certain customers had a number of branches at different addresses and separate accounts were maintained for each branch. The number of accounts was around 1,500 and, to cope with alternative forms of the titles, around 2,000 titles and 2,500 addresses and account numbers had to be held in the table.

The editing strategy employed

In order to reduce the variety in the customer's title and address, editing strategies are employed to standardize, and incidentally shorten, the title and address. Because of the different characteristics of the data in the two subfields, different editing strategies are employed on each subfield.

The title field contains many one- and two-letter abbreviations and conjunctions which add little to the information contained in the title. The titles also tend to be relatively long. Thus the first stage of the strategy adopted deletes all punctuation and all one- and two-letter words. Then, from the resulting strings of words, an abbreviated string of characters is formed by deleting the letters in the even-numbered positions of each word. The binary values of the letters omitted are accumulated to form a hash total, which is used to create two check digits. The abbreviated string of letters plus the two check digits forms the edited title. If the edited title is longer than 18 characters it is left-truncated to that length.

In the address field, however, one- and two-letter words are usually highly significant, either referring to house numbers or postal districts. So, for this field, the editing strategy is the basic one of deleting all punctuation and compressing the words to form a string of letters. Since significant information, in the form of house numbers and town names, is present at both ends of the edited address no truncation is employed.

Thus if the customer were:

LA GRANDE COMPAGNIE DE L'AMERIQUE S.A.,
155 ANGEL PAVEMENT, LONDON, E.C.1.

then the edited title and address would be:

GADCMANAEIU36
155ANGELPAVEMENTLONDONEC1

(*via* GRANDE COMPAGNIE AMERIQUE
155 ANGEL PAVEMENT LONDON EC1)

Program performance

The program successfully read and processed a batch of several hundred test transactions; no transactions were given an incorrect account number, and a surprising variety of mistypings were successfully processed, only a small proportion having to be rejected for retyping. The time to edit a name of 20–80 characters was between 150 and 200 milliseconds, and to look up an account number was between 50 and 300 milliseconds, depending on its position in the table. Thus the processing speed is quite fast enough to justify considering this technique in a commercial application.

Conclusion

With the growth in the use of by-product paper tape from accounting machines and typewriters, and the introduction of optical character-reading machines in the near future, many applications will face this type of problem. In particular, to gain the full advantage of optical reading machines, any manual operations such as encoding reference numbers must be avoided, and thus it is anticipated that the field of application for the techniques described in this paper will enlarge significantly.

Acknowledgements

I am indebted to English Electric-Leo Computers Ltd. for permission to publish this paper, and to Messrs. D. J. Blackwell and I. Edmonds of English Electric-Leo Computers for their assistance in the execution of this work and the preparation of this paper.

References

BARRET, J. A., and GREENE, M. (1960). "Abbreviating Words Systematically," *Comm. Assoc. Comp. Mach.*, Vol. 3, p. 323.
BOURNE, C. P., and FORD, D. F. (1961). "A Study of Methods for Systematically Abbreviating English Words and Names," *J. Assoc. Comp. Mach.*, Vol. 8, p. 538.
BRACE, D. A. (1963). "Direct Coding of English Language names," *The Computer Journal*, Vol. 6, No. 2, p. 113.
DAVIDSON, L. (1962). "Retrieval of Misspelled Names in an Airline Passenger Record System," *Comm. Assoc. Comp. Mach.*, Vol. 5, p. 169.
KOROLEV, L. N. (1958). "Coding and Code Compression," *J. Assoc. Comp. Mach.*, Vol. 5, p. 328.
OETTINGER, A. G. (1957). "Account Identification for Automatic Data Processing," *J. Assoc. Comp. Mach.*, Vol. 4, p. 245.

Book review: Data processing

Data Processing Yearbook 1963-64; 306 pages. *American Data Processing Inc.*, 22nd Floor Book Tower, Detroit 26, Michigan, $15.00.

This is the 13th volume of the publication that started as *The Punched-Card Annual*, the first U.S. nationally-circulated non-manufacturer publication dealing with punched cards and computer systems. It is an interesting publication for the U.K. business user, who wishes to maintain contact with the way business computer users in U.S.A. are thinking, without going there to see: we get so little about business applications in the publications of ACM, or for that matter in our BCS publications.

The *Yearbook* is published by Mr. Frank H. Gille, assisted by ten other individuals, named in the frontispiece, who form an Editorial panel independent of any manufacturer: there are only ten pages of advertisements, none of them for computer systems, mainly for business forms and computer accessories. No claim is made for completeness, which would be very difficult.

The first article is a ten-page survey of company approaches and results by McKinsey and Company Inc., of Park Avenue, New York, enquiring what American industry has gained or learned from its heavy investment in computers. Twenty-seven companies with 4–8 years' experience in computer systems were surveyed. Several companies claimed to have made major gains, but for most companies the effort is still costly and has so far produced only minor benefits. Company systems are always very individualistic, but the authors have devised a classification scheme for the inquiry, from which summary charts are presented. Of the 27 companies in the survey, ten reported administrative cost reductions and 11 had operating cost savings. Intangible benefits were also reported, 10 companies claiming increased speed and 15 that they were getting new information. In 21 companies, the computer manager was not more than two levels below the company's chief executive, and the survey concludes that where top management plays its essential role, important consequences can follow.

This survey is followed by brief biographical notes on the 28 participating authors.

The remainder of the volume is divided into four sections: *Tools* (70 pp), *Techniques* (58 pp), *Tactics* (56 pp), and a reference section (87 pp) on courses, associations, users organizations and data-processing centres. Where the reference section lists organizations outside North America, there are a number of omissions, due no doubt to the difficulties of data collection; for example, Northampton College (London) is omitted from a list of English Colleges running courses. The University Grants Committee in the United Kingdom might note that in 1963 there were 518 digital computers installed, (varying in size from 47 LGP30's to 16 IBM7090 systems), plus 24 on order, in North American

Universities. There were apparently only 24 analogue computers in the reporting Universities. Some of the large and small computers included in the 518, include machines, such as the Ford Motor Company's Engineering Division three IBM systems, to which access is granted for the Community College at Dearborn (150 students).

The *Tools* section comprises surveys of hardware additions announced during the year, input/output equipment (in which magnetic tape is included—considerable room for reductions in cost!) and an authoritative review of memory systems (omitting nickel-delay lines!) by J. Presper Eckert. These chapters are followed by articles on trends in design, real-time data accumulation, progress in data communications, character recognition, and teaching machines and programmed learning.

In the reviewer's opinion, it is a great pity that the only article on software, by Howard Masters and Geraldine Bowen of Univac, is on "FORTRAN—Formula for Business", which might have been written several years ago. Did we not have Dr Grace Hopper herself with us in London four years ago, hinting at further developments, which by now should have resulted in effective use of COBOL, FACT, etc., in U.S.A? (Give us a few months and we may yet overtake them?) If this *Yearbook* is really to keep management informed, more than four pages should be given to software in future editions, without which the engineering skill of the builders and the systems experience or mathematical ability of the problem-experienced user cannot quickly be brought together.

The section on *Techniques* includes business-manager oriented papers on new microfilm systems, CPM and PERT (well illustrated with figures), operations research, decision processes research, and single information-flow philosophy. Arthur F. Anderson writes on "Protecting Magnetic Tape," and recommends a "safe within vault" system for important tapes.

Dr. Herbert W. Robinson (CEIR) engages attention to the influence of computers and computer technology as "amplifiers of intellectual powers of reasoning and computation." His organization probably employs more mathematical statisticians, and has access to more large computer systems, than any other non-government, non-manufacturer organisation in the world. It would appear that financial controllers are now looking hard at their data-processing costs: hardware manufacturers produce faster and faster machines, quicker than Americans can learn how to apply them. The really big challenges, for statistical analysis, in his opinion, go unanswered. He ventures to predict that if a company's costs for computer facilities, software and personnel are £36,000 per month after startup, the work could be done by professionally staffed data-processing centres at half the cost. But what about large volumes of data-preparation and knowledge of the company's problems, advantages from company programmers *knowing* from experience what they are trying to do; and what of confidentiality aspects? If this argument is true, why haven't large companies also sacked their accountants and employed professional firms to prepare their monthly accounts? Why has not a standardized service-bureau procedure already been adopted for insurance companies in U.S.A? This argument if true must apply to bread and butter work, as well as to more elegant O.R. applications.

This is a very stimulating paper, from many viewpoints. His concluding paragraphs sum up the argument that computer people haven't yet started on their real task, that of *Strategos*– the leader, uniting all the arts and sciences under one system to solve the "total" problems of man. "To this end a world symposium should be held." These are serious and stimulating ambitions. Many of us in U.K. will be happy, if, before our retirement, we can bring together the various arts and sciences (not to mention personalities) to solve one company's problems! What we would appreciate from U.S.A. business experts is a few more papers on what has been *done* and *how*, not long-range predictions of what *might* be done.

The section on *Tactics* goes some way to answer this need. There are papers on applications to advertising (e.g. linear programming of media mix), cheque handling within the Federal Reserve system, Local Government and Tax Administration (Survey of 185 reports), legal applications (e.g. patent search, indexing of statutes, etc), manufacturing and publishing (computer assisted composing). All these papers are well illustrated with tables and charts, or references to original papers elsewhere.

The reviewer found this *Yearbook* stimulating reading, for a weekend free from failure in the air-conditioning system, while awaiting delivery of a second-generation computer. The book is nearly as stimulating as a visit to U.S.A.

H. W. GEARING.