Electronic data processing for the International Vocabulary of Terms used in Information Processing

By J. S. Gatehouse*

This paper describes the uses made of data processing in the construction of a vocabulary or glossary of terms.

The vocabulary of terms used in information processing that has been prepared under the joint auspices of the International Federation for Information Processing and the International Computation Centre and which is about to be published has been described in the paper by Gould & Tootill (1965)†. The British contribution to this work was made through the glossary committee of The British Computer Society; the committee also took a major part of the responsibility for editing. As the work progressed it was realized that the lack of an up-to-date index, which could readily be modified as changes were incorporated, was holding up the work. It was not easy to follow the cross references and ensure compatibility, and even harder to check that the definitions did not form loops. An example of the latter is that it is too easy to define PROGRAM in terms of INSTRUCTIONS and INSTRUCTION in terms of PROGRAM. Since this was a data-processing committee it was clear that data-processing techniques should be used to overcome these difficulties. The program and computer used had to be such that the programming would not be difficult, the amount of labour required should be reasonable and the work had to be done on a computer readily available to at least one member of the committee. These restrictions arise from the fact that the committee was staffed by voluntary labour, as is the usual case with B.C.S. committees. The programs were successfully written for a scientific computer, Mercury, in a scientific language with little difficulty, and ran successfully. *It was particularly interesting to find that techniques that would normally be considered to belong to the commercial field were readily programmed in a scientific language and successful object programs obtained in two or three weeks of spare-time work.*

The index

This was the easier program to prepare; like the other program it was written purely in Mercury Autocode, with the exception of instructions for "read one character" and "print one character" which do not exist in that language‡. The data was punched onto five-track paper tape directly from the draft of the vocabulary, the teleprinter operator punching a line each of the identifier for the term followed by that term. The terms were stored in the computer, five characters at a time as twenty-five bit numbers, by taking the value of the holes punched in the paper tape for each character and multiplying these by 32^n where $n = 4, 3, 2, 1, 0$. Each term is thus stored in the computer as a sequence of numbers, together with the identifier for that term and an end-of-block character. If these numbers are regarded as left-justified it is found that the numerical order is also the alphabetical order (using the values of Ferranti tape code) with the single exception that the hyphen has a lower value than the space, and this needs to be corrected. For example, DOUBLE-LENGTH would precede DOUBLE PRECISION contrary to the usual practice of alphabetical ordering. Variable block length was employed so as to pack the information more closely on to the magnetic-drum backing store of the Mercury computer. There are a total of 1,623 terms in the final version of the vocabulary, and these terms may have up to 60 characters each.

The sorting procedure used was that of Quicksort (Hoare, 1962). This procedure has been found to work very well on Mercury since no additional storage is required, and it has been found to operate so fast that there has never been any need to improve on the very first routine that was written, and which was intended to be experimental. The method was modified slightly in this context; a separate list was made on the magnetic drum containing only the first machine word for each term together with the address at which this word was stored in the main list. This list was then sorted by Quicksort and used to output§ the index in alphabetical order. If two or more adjacent machine words in the sorted secondary list were identical when the output reached this point, these machine words were exchanged for the second machine word of the term drawn from the main list and these sorted. The procedure was repeated over all the machine words containing the terms until a different word was found.

This procedure was simple enough for amended versions of the index to be provided for committee meetings. For meetings of the British committee, six copies were made, perhaps a week before the meeting, for reference at that meeting; for the International committee the print-out was taken on a spirit master and enough copies made so that each member of the International committee could have one. This procedure was necessary since the index at the end of a committee meeting was rarely the

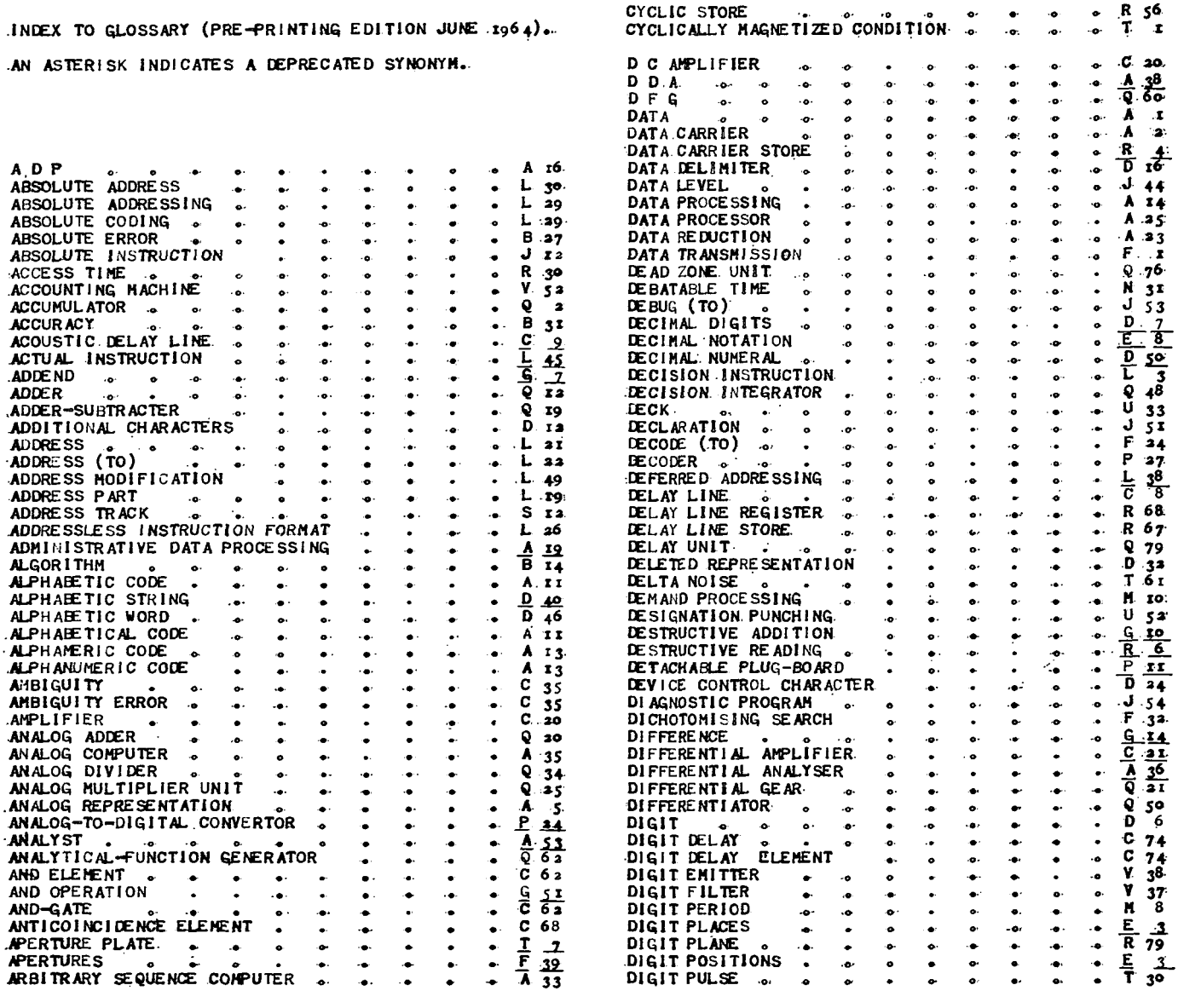
† *See p. 264 of this issue.* Ed.
‡Available in Extended Mercury Autocode (1964).
§This verb is now defined in the Vocabulary. *Ed.*
**The General Electric Company Limited, Erith, Kent.*

same as at the beginning, and quite frequently had changed considerably.

The final run of this program was made for the publisher and was modified so that the input contained identifiers showing whether the term was a main term, a deprecated synonym or a term defined *en passant*. Fig. 1 shows two pages of this final list where the identifiers for *en passant* terms are underlined, and deprecated synonyms are indicated by an asterisk.

The loop-finding procedure

As Gould & Tootill have described, when a reference to a term that has been defined elsewhere in the vocabulary occurs in the definition of another term this reference will be printed in italics. When the use of a computer for this project was first proposed it was suggested that we could feed the whole vocabulary into the computer and check each word against a list of the defined terms. This would show where the italicized words should be, and we could then use the information to determine whether loops of definitions existed. This procedure, however, was not adopted since the amount of information to be processed would be colossal—somewhere around 50,000 words each to be checked against 1,600 key words—and there was the added difficulty that the cross-referenced words were not always of the same form as the defined term: for example, plurals and different tenses of a verb. Another difficulty is that a word may be present in a



INDEX TO GLOSSARY (PRE-PRINTING EDITION JUNE 1964).

AN ASTERISK INDICATES A DEPRECATED SYNONYM.

A D P A 16
ABSOLUTE ADDRESS L 30
ABSOLUTE ADDRESSING L 29
ABSOLUTE CODING L 29
ABSOLUTE ERROR B 27
ABSOLUTE INSTRUCTION J 12
ACCESS TIME R 30
ACCOUNTING MACHINE V 52
ACCUMULATOR Q 2
ACCURACY B 31
ACOUSTIC DELAY LINE C 9
ACTUAL INSTRUCTION L 45
ADDEND G 7
ADDER Q 12
ADDER-SUBTRACTER Q 19
ADDITIONAL CHARACTERS D 12
ADDRESS L 21
ADDRESS (TO) L 22
ADDRESS MODIFICATION L 49
ADDRESS PART L 29
ADDRESS TRACK S 12
ADDRESSLESS INSTRUCTION FORMAT L 26
ADMINISTRATIVE DATA PROCESSING A 19
ALGORITHM B 14
ALPHABETIC CODE A 11
ALPHABETIC STRING D 40
ALPHABETIC WORD D 46
ALPHABETICAL CODE A 11
ALPHAMERIC CODE A 13
ALPHANUMERIC CODE A 13
AMBIGUITY C 35
AMBIGUITY ERROR C 35
AMPLIFIER C 20
ANALOG ADDER Q 20
ANALOG COMPUTER A 35
ANALOG DIVIDER Q 34
ANALOG MULTIPLIER UNIT Q 25
ANALOG REPRESENTATION A 5
ANALOG-TO-DIGITAL CONVERTOR P 24
ANALYST A 53
ANALYTICAL-FUNCTION GENERATOR Q 62
AND ELEMENT C 62
AND OPERATION G 51
AND-GATE C 62
ANTICOINCIDENCE ELEMENT C 68
APERTURE PLATE T 7
APERTURES F 39
ARBITRARY SEQUENCE COMPUTER A 33

INDEX PAGE 1

CYCLIC STORE R 56
CYCLICALLY MAGNETIZED CONDITION T 1

D C AMPLIFIER C 20
D D A A 38
D F G Q 60
DATA A 1
DATA CARRIER A 2
DATA CARRIER STORE R 4
DATA DELIMITER D 16
DATA LEVEL J 44
DATA PROCESSING A 14
DATA PROCESSOR A 25
DATA REDUCTION A 23
DATA TRANSMISSION F 1
DEAD ZONE UNIT Q 76
DEBATABLE TIME N 31
DEBUG (TO) J 53
DECIMAL DIGITS D 7
DECIMAL NOTATION E 8
DECIMAL NUMERAL D 50
DECISION INSTRUCTION L 3
DECISION INTEGRATOR Q 48
DECK U 33
DECLARATION J 51
DECODE (TO) F 24
DECODER P 27
DEFERRED ADDRESSING L 38
DELAY LINE C 8
DELAY LINE REGISTER R 68
DELAY LINE STORE R 67
DELAY UNIT Q 79
DELETED REPRESENTATION D 32
DELTA NOISE T 61
DEMAND PROCESSING M 10
DESIGNATION PUNCHING U 52
DESTRUCTIVE ADDITION G 10
DESTRUCTIVE READING R 6
DETACHABLE PLUG-BOARD P 11
DEVICE CONTROL CHARACTER D 24
DIAGNOSTIC PROGRAM J 54
DICHOTOMISING SEARCH F 32
DIFFERENCE G 14
DIFFERENTIAL AMPLIFIER C 21
DIFFERENTIAL ANALYSER A 36
DIFFERENTIAL GEAR Q 21
DIFFERENTIATOR Q 50
DIGIT D 6
DIGIT DELAY C 74
DIGIT DELAY ELEMENT C 74
DIGIT EMITTER V 38
DIGIT FILTER V 37
DIGIT PERIOD M 8
DIGIT PLACES E 3
DIGIT PLANE R 79
DIGIT POSITIONS E 3
DIGIT PULSE T 30

INDEX PAGE 7

Fig. 1.—Two pages from the index to the vocabulary

definition with its normal English language meaning and not the data-processing meaning. However, I now think such a procedure is practicable. Taking two words as being the same if they agree for, say, the first 75% of the characters of the longer word this would produce some apparent cross-references that did not truly exist, and these would have to be removed by hand. One would still be left with the major difficulty that the whole vocabulary would have to be transcribed on to punched paper tape.

The procedure that was in fact adopted was to locate the cross-references by hand and then to use the computer to see if any loops of definitions did exist. This was not as onerous a task (to the members of the committee) as it may sound, since by this time the committee were fairly familiar with the terms that were defined and it was easy enough to check when one was uncertain once the index had been produced. The data was prepared for each term in the form of the identifier for that term, followed by the identifiers for the cross-references, and terminated by an end of block character. Where a term was defined *en passant* the only cross-reference was taken as the identifier for the main term in which the *en passant* definition occurred. When this data was stored on the magnetic drums a marker was added for each term showing the number of cross-references that occurred in the definition for that term, that is, the number of identifiers that occurred between the identifier for the term and the end-of-block character.

The routine for searching for loops of definitions was the same as that required to find loops of activities in a critical-path analysis. From this point of view the structure of the vocabulary can be regarded as a network and, using the analogy with critical-path techniques, each term can be regarded as an activity and those terms to which cross-references applied within the definition regarded as immediate predecessor activities. It is then only necessary to assemble the terms in a "topological" order until a point is reached where this is no longer possible. When this procedure fails a loop must exist. The first step, then, is to search through the list for those terms for which the marker has been set to zero: that is, those terms for which there are no cross-references within the definition. These terms which can be regarded as the starting definitions of the network can be listed directly. As each is considered the total list is scanned, and where the identifier occurs as a cross-reference the marker for that term is reduced by 1. Proceeding in this way, removing cross-references from the list when that is possible, and listing the terms for which all cross-references have been removed, produces a topological order for the terms (this order is not, of course, a unique order). If a loop exists there comes a stage when none of the terms have their markers set to zero. The procedure adopted was to look through the list to find a marker set to 1, and remove the remaining cross-reference. This is arbitrary; there is no evidence that it will be the best way to break the loop nor even that the loop will necessarily be broken. To find the best position to break the loop requires a search of the whole network and this is not a practicable procedure, as has been found with critical-path analysis work. A section of the output from the final run was

ORDERED LIST OF TERMS IN GLOSSARY(CORRECTED TO JAN 1964)

THE ORDER IS THAT OF THE NEW STRUCTURE EXCEPT THAT NO
TERM IS LISTED UNTIL ALL TERMS TO WHICH THE DEFINITION REFERS
HAVE BEEN LISTED. REFERENCES IN NOTES, AND IN EXAMPLES
THAT HAVE THE FORM OF A NOTE, ARE IGNORED, EXCEPT
FOR TERMS THAT ARE DEFINED WITHIN THE NOTE OR EXAMPLE.

LOOPS OF DEFINITIONS ARE BROKEN BY REMOVING
AN ARBITRARY CROSS-REFERENCE.

```
A  1    DATA
A  5    ANALOG REPRESENTATION
A  6    INCREMENTAL REPRESENTATION
A 40    NETWORK ANALOG
A 57    LINEAR OPTIMISATION
A 58    NON-LINEAR OPTIMISATION
A 59    MONTE CARLO METHOD
A 60    AUTOMATION
A 61    AUTOMATICS
A 62    CYBERNETICS
B  1    NUMBER
B  5    BINARY
B  6    TERNARY
B  9    MANTISSA
B 11    RANGE
B 12    OUT OF RANGE
B 13    SPAN
B 15    HEURISTIC
B 16    RECURSIVE DEFINITION OF A FUNCTION
B 18    NORMALISE (TO)
B 22    SCALE (TO)
B 23    SCALE FACTOR
B 24    ERROR
B 25    BALANCED ERROR
B 27    ABSOLUTE ERROR
B 28    RELATIVE ERROR
B 31    STATIC ERROR
B 32    DYNAMIC ERROR
B 35    RESOLUTION ERROR
B 38    ACCURACY
B 39    PRECISION
B 14    ALGORITHM
B 44    FORMAL LOGIC
B 45    SYMBOLIC LOGIC
B 46    MATHEMATICAL LOGIC
C  2    SOLID-STATE COMPONENTS
C  2    ELECTRONIC
C  3    MINIATURISATION
C  4    MICROMINIATURISATION
C  5    INTEGRATED CIRCUIT
C  6    BISTABLE
C  7    ECCLES-JORDAN
C  8    MONOSTABLE
C 21    HIGH-GAIN AMPLIFIER
```

Fig. 2.—The first page of results from the loop-finding program

V71 PAPER THROW
V72 FORM FEED

LOOP—REMOVE REFERENCE TO
J11 COMPUTER INSTRUCTION
FROM DEFINITION OF
A30 CONSECUTIVE SEQUENCE COMPUTER

A30 CONSECUTIVE SEQUENCE COMPUTER
K53 JUMP
K49 TRIGGER (TO)
K50 SELF-TRIGGERED PROGRAM
L6 UNCONDITIONAL JUMP INSTRUCTION
L7 UNCONDITIONAL JUMP
L11 REPETITION INSTRUCTION
M4 COMPUTER OPERATION
J11 COMPUTER INSTRUCTION
A31 ARBITRARY SEQUENCE COMPUTER

Investigation showed that the loop consisted of A30, J11, M4, K53, A30. When we checked back on the definitions it was found that the loop only existed because JUMP was given as an example of a computer operation, and so no action was taken.

This procedure was rather time-consuming on the computer, mainly because of the need each time a term was listed as being available to scan the remaining list to remove all terms to which it was a cross-reference. No attempt was made to speed up the procedure or to optimize the program since it was only going to be used a few times.

The first page of the output obtained from the final run is shown as Fig. 2. It can be seen that the order bears little relation to that of the final publication, illustrating the impossibility of listing the terms in an order so that no back referencing is required, and at the same time keeping related terms together.

References

GOULD, I., and TOOTILL, G. C. "The terminology work of IFIP and ICC," *The Computer Journal*, Vol. 7, p. 264.
HOARE, C. A. R. (1962). "Quicksort," *The Computer Journal*, Vol. 5, p. 10.

Book Review

Automation in Bankwesen, by HANS PETER BAUER, 1962; 151 pages. (Tübingen: *J. C. B. Mohr (Paul Siebeck)*, DM 14.50).

This book is obviously intended to be a guide for people at the managerial and higher levels in banking who are considering the possibilities of electronic computers in the European banking world. It assumes no prior knowledge of, or reading in, its subject-matter and, although much of its content is historical and refers to outmoded machines, the lessons drawn from this history are still valid and should save much catastrophic experimentation in European banking organizations.

The author has planned his book in two logical sections. The first is a general survey of the field in America. The author has made a general study of the progress of automation in the Bank of America, with reference to the bank's structure, policy and growth. He covers its introduction of "ERMA" and the IBM 702, taking into account the reasons for the installation of these systems, the staff reorganization necessary and the experience gained. In this part of his book he also deals with such vital questions as machine reliability and accuracy.

The remainder of the book is devoted to the possibilities of bank organization in Europe. Here the author is on his home ground in dealing with banking structures considerably more intricate than those in America.

He details those machines currently available and their likely application (this section of the book cannot help but be of limited value, as the machine market is constantly changing). He is particularly interesting on the economics of making an installation—both as regards the installation itself and the benefits accruing from it. He also makes many cogent remarks on the staffing aspect, particularly the "problem" of redundancy. Finally the author includes a thought-provoking section on future developments in bank automation, sections on the training and prospects of specialists in EDP, and the effect of automation on the total economy.

In total, this book is a worthwhile (indeed almost essential) addition to the library of any one interested in European banking.

Appendix—Translation of the Contents of the book.

INTRODUCTION: The problem.

1st Chapter: America

Survey

1. Bank of America
   Structure—Business policy—business economy and staffing problems—growth.
2. The automatic cheque accounting machine "ERMA"
   History of its origin—installation into the structure of the bank—data transmission—magnetic lettering—electronic accounting.
3. The electronic multi-purpose machine IBM 702
   Reasons for its purchase—integrated data processing: mortgage and consumer credits—consolidation of branch clearing systems—other fields of application—experiences and new plans.
4. Organization
   Planning and executive departments—distributions of powers and duties.
5. Technical questions
   Does the machine make mistakes? Maintenance.
6. The extent of automation.

2nd Chapter: Possibilities in Europe

1. Organizational questions
   The available types of machine—size and business structure of European banks—centralization—planning and organization.
2. Aspects from the point of view of business economy (prac. economics)—customer service—modern business management—economy—reduction of office and admin. costs—statistics and customer service—installation and operating costs—expenditure and profit—replacement.
3. Staff
   New categories—redundant staff.
4. Summary

(*Review continued on page* 289)