Bothalia 11, 3: 325–328 (1974)

An automatic bibliography indexing programme

J. W. MORRIS*

ABSTRACT

A relatively simple FORTRAN IV programme, designed for a small computer, for author and key-word indexes to bibliographic records is described, and examples of output are given. It is compared with some other systems. Suggested improvements to the programme are given.

INTRODUCTION

In order to make available for reference within the Institute the 300 reprints and 200 literature references collected during an overseas tour, some method of indexing was required. A simple alphabetical list by author, even within broad subject divisions, was considered insufficient for efficient use of the collection. The collection covers in-depth the fields of ecosystem modelling, multivariate analysis, information retrieval and related subjects to a less comprehensive degree.

It was, for several reasons, decided to try a computerized system of indexing. Such a system would satisfy the needs of author and subject indexes. It would also serve as an example for a more comprehensive computer programme for the whole Institute, or at least it would form the basis for discussion of such a bibliographic index. Another reason is that the system has been seen to work on a number of similar bibliographies overseas. Other considerations were that the data set was fairly small, not open-ended, and the subject catergories were limited in number although covering a fairly wide field.

The indexing programme is described and then compared with some other systems before conclusions are made as to the utility of the programme.

COMPUTER PROGRAMME AND INPUT DATA LAY-OUT

The programme, called BIBLO of some 500 cards was written in standard FORTRAN IV and implemented on an IBM/1130 computer with 32K words core memory. At the time no other computer was available and the programme was written specifically for a small computer. As the sort routine took too long on the small computer, the programme was modified to run on an IBM/360. Even on the large computer with optimization for speed and a faster sort routine, the programme was time-consuming. Programme listings and card decks are available, on request, from the author.

Five card types, numbered 1 to 5, are used for data entry and between five and seven cards are punched for each reference. Examples of input data are given in Fig. 1.

On each card, the accession number is punched in the first four columns and the card type number in the fifth. References and reprints may be accessioned in any order and given any four-digit number. In this case, reprints in the collection were numbered consecutively from 1 000 upwards and references from 2 000. The accession number was written on the reprint and the reprints filed in numerical order to aid retrieval. Column six of each card is left blank (or a zero punched) except in card types 3 and 4 when a 'one' indicates a continuation card of that type.

The first card (type 1) contains the author's name and initials starting in the seventh column. When there is more than one author, each name is separated by a comma. Periods between and after initials are omitted. The second card (type 2) contains the date of publication (year) punched in the seventh through tenth columns. The title of the article or book is given on the third card (type 3), starting in column seven. One continuation card may be used if the title is too long to fit on one card. The citation is given on the fourth card type. A continuation card may be used. The fifth card type contains title-enriching terms.

* Botanical Research Institute, Department of Agricultural Technical Services, Private Bag X101, Pretoria.

```
COLUMN   1111111111222222222233333333334444444444555555555566666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

123410ROSS JH,MORRIS JW
1234201971
123430PRINCIPAL COMPONENTS ANALYSIS OF ACACIA BURKEI AND ACACIA NIGRESCENS IN NA
123431TAL
1234408OTHALIA 10 (3), 437-450.
123450NUMERICAL-TAXONOMY PCA
147010MATHER PM
1470201970
147030PRINCIPAL COMPONENTS AND FACTOR ANALYSIS
147040COMPUTER APPLICATIONS IN THE NATURAL AND SOCIAL SCIENCES NO. 10 UNIVERSITY
147041 OF NOTTINGHAM.
147050PCA FACTOR-ANALYSIS
209410ROHLF FJ,SOKAL RR
2094201965
209430COEFFICIENTS OF CORRELATION AND DISTANCE IN NUMERICAL TAXONOMY
209440UNIV. KANSAS SCI. BULL. 45  1-27.
209450
```

FIG. 1.—Example of three references punched on 80-column cards for input to BIBLO.

Each term is separated by a blank space and two-word terms, like NUMERICAL TAXONOMY, are joined by a hyphen (see Fig. 1).

COMPUTER OUTPUT

An alphabetical list of authors and the accession numbers of their references is the first output from the programme (Fig. 2). The words of the title and title-enriching terms are then listed in alphabetical order.

PRINCIPAL AUTHORS

1147	ALLISON H
1124	ANON
1119	ANON
1123	ANON
1062	ANON
1125	ANON
1255	ANON
1188	ANON
1130	ANON
1282	AUCLAIR AN
1212	AUSTIN MP
1260	AUSTIN MP
1061	BALLEYDIER R
1233	BEAMAN JH
1241	BEAMAN JH
1008	BESCHEL RE
1245	BESCHEL RE
1259	BISBY FA
1044	BRISSE H
1117	BROUGH R
1265	BROWN RA
1170	BUNCE RGH
1108	BUNCE RGH
1107	BUNCE RGH
1268	BURTON HD
1269	BURTON HD
1175	CARTER CI
1152	CASWELL H
1053	CEDERGREN RJ
1037	CESKA A
1134	CHASE RH
1209	CHEETHAM AH
1128	CLINE HF
1301	CONNOR RJ
1202	COOK CW
1196	COOK CW
1198	COOK CW
1213	COOK CW

FIG. 2.—Part of an alphabetical list of principal authors and reference numbers.

Words are listed together with the senior author of the reference and the accession number (Fig. 3). Title words with low information content, such as WITH, FOR, and BUT, specified prior to running the programme, are omitted from the list. Up to 200 words, chosen by the user, may be excluded in this way. The word index is a KWOC (Key-Word Out of Context) one as the title must be referred to elsewhere to determine the context of the word. Finally, listings of the references in order of accession number and of principal author are given (Fig. 4).

TITLE AND KEY-WORDS

2029	ACACIA	ROSS JH
2136	ACTUAL	GODRON M
2185	ADSORPTION	GOLDSTEIN RA
2012	ADVANCED	RAO CR
2054	AERIAL	HOWARD JA
2192	AFRICA	ACOCKS JPH
2193	AFRICA	MORRIS JW
2024	AFRICA	EDWARDS D
2154	AID	STEWART DH
2177	AIDED	CEDERGREN RJ
2165	ALEAS	ESCOUFIER Y
2162	ALGARIOS	SCHEINVAR L
2005	ALGEBRA	SEARLE SR
2082	ALGEBRAIC	KRZANOWSKI WJ
2183	ALLOMETRY	JOLICOEUR P
2180	ALLOMETRY	JOLICOEUR P
2015	AMERICA	MORSE LE
2181	AMERICANA	JOLICOEUR P
2132	ANAL-FACT-CORRE	ROMANE F
2128	ANAL-FACT-CORRE	GORDIER B
2133	ANAL-FACT-CORRE	ROMANE F
2125	ANAL-FACT-CORRE	BENZECRI JP
2135	ANAL-FACT-CORRE	THOMASSONE R
2129	ANAL-FACT-CORRE	DO THINHUNG M
2128	ANALYSE	GORDIER B
2060	ANALYSIS.	CATTELL RB
2061	ANALYSIS.	CATTELL RB
2085	ANALYSIS.	MATHER PM
2099	ANALYTICAL	SOKAL RR
2103	ANGIOSPERM	WATSON L
2102	ANGIOSPERM	WATSON L
2093	ANGIOSPERM	PRANCE GT
2185	ANIMALS	GOLDSTEIN RA
2191	ANTIARCH	HEMMINGS SK
2162	ANTOMATICOS	SCHEINVAR L
2063	APPLICATION	CRADDOCK JM
2182	APPLICATION	JOLICOEUR P
2137	APPLICATION	GODRON M
2141	APPLICATION	ROMANE F
2040	APPLICATION	AUSTIN MP
2115	APPLICATIONS	GODRON M
2007	APPLICATIONS	HARBAUGH JW

FIG. 3.—Part of KWOC index of title- and key-words.

SOME OTHER SYSTEMS

Details of some comparable bibliographic systems are summarized in Table 1. They are described in the following section.

TABLE 1.—Summary of some automatic indexing programmes.

Author (date)	Mnemonic	Programme language	Computer used
Burton, Hilary D. *et al.* (1969)	FAMULUS	FORTRAN IV	CDC 6400 CDC 6600 IBM/360-40 Univac 1108
Bridges, K.W. (1970)	INDEX	PL/1	IBM/360
Cedergren, R. J. (1971)	CHERCHE	FORTRAN IV	CDC 6400
Creighton, R. A. *et al.* (1971 & 1972)	SELGEM	COBOL	Honeywell
Morris J. W. (this paper)	BIBLO	FORTRAN IV	IBM/1130*

* later modified for IBM/360.

```
 1263  1971  PHIPPS JB
             DENDROGRAM TOPOLOGY
              SYST. ZOOL. 20  306-308
KEY WORDS *    CLASSIFICATION-COMPARISON

 1220  1972  PHIPPS JB
             STUDIES IN THE ARUNDINELLEAE - GRAMINEAE -. XI. TAXIMETRICS OF CHANGING CLASSIFICATIONS
              CANADIAN J BOT 50,787-802
KEY WORDS *    PCA NUM-TAXONOMY

 1219  1972  PHIPPS JB
             STUDIES IN THE ARUNDINELLEAE - GRAMINEAE -. XIII. TAXIMETRICS OF THE LOUDETIOID,
             TRISTACHYOID, AND DANTHONOIPSCID GROUPS
              CANADIAN J BOT 50,935-948
KEY WORDS *    HIERARCHY PCA NUM-TAXONOMY

 1013  1969  PIELOU EC
             ASSOCIATION TESTS VERSUS HOMOGENEITY TESTS - THEIR USE IN SUBDIVIDING QUADRATS
             INTO GROUPS
              VEGETATIO 18,4-18

 1046  1972  POISSONET P
             RELATIONS DE VOISINAGE ENTRE VEGETAUX D UNE FORMATION HERBACEE DENSE - DISPOSITIF
             EXPERIMENTAL ET PARAMETRES DE LA PRODUCTION
              OECOL PLANT 7,23-43
KEY WORDS *    COMPETITION

 1064  1965  PRINGLE JS
             HYBRIDIZATION IN GENTIANA - GENTIANACEAE - A RESUME OF JT CURTIS STUDIES
              WISC ACA SCI,ARTS AND LETT 54,283-293

 1045  1968  RAPP M, ROMANE F
             CONTRIBUTION A L'ETUDE DU BILAN DE L'EAU DANS LES ECOSYSTEMES MEDITERRANEENS
              OECOL PLANT 3,271-284
KEY WORDS *    RAINFALL THROUGHFALL
```

FIG. 4.—Example of bibliography listed alphabetically by principal author. First number on each author line is reprint accession number and second is date of publication.

A very simple computer-aided bibliography programme was developed by Cedergren (1971). Alphabetical lists of principal authors and up to four key-words per reference are produced along with the citations and one line of comment. The comment may be the title or additional key-words. Words in the title (comment line) and secondary authors are not indexed.

Bridges (1970) discusses the application of computer processing to maintain a personal bibliography and produce a sophisticated set of printed indexes. He considers that personal bibliographies, especially when well indexed, are important tools for scientific research, teaching, administration and writing. Some benefits of computer processing over index-card bibliographies which he lists are: ease of making multiple copies, ease of transport and the small amount of assistance required by users from the compiler as the indexing criteria are based on a consistent set of procedures. In addition to a straight bibliography listing and author index (principal and secondary authors), a KWIC (Key-Word In Context) index of title and key-words and an index to sources is given. The word index is in context as a few words before and after the indexed term are also printed. The same type of KWIC index is given by Biological Abstracts publications and other commercial abstracting services.

Complex, automated documentation systems have been developed by Burton *et al.* (1969) of the U.S. Department of Agriculture and Creighton *et al.* (1971 & 1972) of the Smithsonian Institution in Washington. The user is free to design input, content and format. A wide range of outputs, from searches for specific authors, words or references, to indexes of all kinds, is available with each system. As a large computer is a pre-requisite, these systems, as well as those of the commercial organizations, are not discussed further.

CONCLUSION

BIBLO is less sophisticated than FAMULUS and SELGEM (Table 1) but provides more information in the way of key-word indexes and lists of secondary authors than CHERCHE. It is the only one capable of being run on a small computer (Table 1). The programme can be improved in a number of ways. At present two lists of authors are produced (principal authors and all authors) whereas a list of all authors could be given with principal authors marked in some way. A means of identifying new accessions to the bibliography (c.f. CHERCHE) could be built in. A third improvement would be the linking of two-word terms with a special character (such as an ampersand) in the place of the hyphen used at present. The special character would be suppressed in the printing of the indexes and would improve readability. More sophisticated indexing is also possible.

In that it made available for reference the literature collected on my overseas tour in the form of alphabetical indexes of authors and key-words, it is considered a successful computer application. The possibility of applying the programme to other bibliographies is being actively considered. As powerful computer facilities are available, the question of whether to modify BIBLO or adopt SELGEM, FAMULUS or some other system, such as that of T. J. Crovello (in Morris, 1973), should be carefully considered.

ACKNOWLEDGEMENT

The author is grateful to Dr D. Edwards for his interest in the project and for his comments on this paper.

OPSOMMING

'n Betreklik eenvoudige FORTRAN IV program vir outeur- en sleutelwoordindekse tot bibliografiese gegewens, word vir 'n klein rekenoutomaat beskryf. Voorbeelde van resultate word gegee. Dit word met ander sisteme vergelyk. Moontlike verbeterings aan dié program word aan die hand gedoen.

REFERENCES

BRIDGES, K. W. 1970. Automatic indexing of personal bibliographies. *Bioscience* 20: 94–97.

BURTON, HILARY D., RUSSEL, R. M. & YERKE, T. B. 1969. FAMULUS: a computer-based system for augmenting personal documentation efforts. *U.S.D.A. Forest Service Research Note PSW*-193: 1–5.

CEDERGREN, R. J. 1971. Computer-aided bibliographies for personal or group use. *J. Chem. Documentation* 11: 224–226.

CREIGHTON, R. A. & CROCKETT, J. J. 1971. SELGEM: A system for collection management. *Smithsonian Institution Information Systems Innovations* 2 (3): 1–24.

CREIGHTON, R. A., PACKARD, P. & LINN, H. 1972. SELGEM Retrieval: A general description. *Smithsonian Institution Procedures in Computer Science* 1 (1): 1–38.

MORRIS, J. W. 1973. *Overseas advances in quantitative ecology and computerised data banking*. Unpublished Departmental Report, Botanical Research Institute, Pretoria.