DOCUMENT RESUME

ED 078 884 LI 004 419

AUTHOR Conaway, Charles W.
TITLE A User's Guide to Rice's KWAC (Key Word Alongside of Context) Indexing Program. Version 3.0.
INSTITUTION State Univ. of New York, Buffalo. School of Information and Library Studies.
PUB DATE Jul 73
NOTE 16p.; (0 References); SILS Technical Memorandum No. 1

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Automatic Indexing; Bibliographic Citations; Computer Programs; Indexes (Locaters); *Indexing; *Information Processing; Information Science
IDENTIFIERS Information Science Education; *Key Word Alongside of Context; KWAC

ABSTRACT

The KWAC Index Generation Program was implemented at the State University of New York at Buffalo. It consists of only 252 statements for the COBOL Compiler Edition V310222 on the CDC6400 computer under the KRONOS operating system in the batch mode. The KWAC program is essentially a KWIC index generator designed for a special purpose for use in a particular course, though it has sufficient flexibility to be used in other, similar contexts. The program takes free form natural language input, and generates an index in alphabetical order of each significant word appearing in the input alongside which appears the bibliographical description of the document in which the word was located. Output options permit free form title page, introduction, and epilog. Page headers and footers are permitted, and index page numbers are assigned and printed automatically. Determination of significant words is made in two ways: 1) a word is determined to be significant by default, if it does not appear in a user-input stop list and provided it begins with an alphabetic character and it is more than one character long; and 2) an otherwise significant word may be eliminated from the indexing by a simple procedure at the time of input. (Author/SJ)

A User's Guide to Rice's KWAC (Key Word Alongside of Context) Indexing Program. Version 3.0.

Prepared by Charles Wm. Conaway

July, 1973

SILS Technical Memorandum No. 1.

School of Information and Library Studies
State University of New York at Buffalo
Buffalo, New York 14214

TABLE OF CONTENTS

		PAGE
I.	INTRODUCTION	1
II.	PROGRAM DESCRIPTION	2
III.	INPUT REQUIREMENTS	3
IV.	DECK MAKE-UP TO GENERATE A KWAC INDEX	6
	IV. A. JOB CONTROL CARDS	6
	IV. B. KWAC CONTROL CARDS AND DATA INPUT CARDS	6
	APPENDIX A. JOB CONTROL CARDS REQUIRED AT SUNY AT BUFFALO	10
	APPENDIX B. SAMPLE OUTPUT PAGES	11
	APPENDIX C. SAMPLE BIBLIOGRAPHIC DATA INPUT CARD LISTING	13
	APPENDIX D. KNOWN BUGS IN KWAC	14

I. INTRODUCTION

The KWAC (Key Word Alongside of Context) Index Generation Program was written and debugged by Lester A. Rice* at the School of Information and Library Studies, SUNY at Buffalo, during the winter 1972-73. The purpose of the program was primarily pedagogical. It was developed specifically for LI 561 - Information Storage, Retrieval, and Selective Dissemination Systems where it was used successfully during the Spring Term 1973. The intention was to produce a group of indexes to the same data base for the purpose of comparative evaluation of their performances in an experimental environment. The corpus indexed at that time was 36 substantive articles appearing in the Journal of the American Society for Information Science, v. 22 (1971).

KWAC is a KWIC-type index written for the specific purpose indicated above, and consequently has some rather severe input and output limitations. It is acknowledged that these limits reduce its general usefulness and while it is not particularly difficult to expand them for more general use, this has not been done for two reasons: 1) the ready availability of existing program packages with considerably more flexible capability; and, 2) the simplicity and proven effectiveness of the KWAC index for the purposes intended.

* Dr. Rice is currently employed in the Reference Department, University Libraries, University of Pennsylvania, Philadelphia, Pennsylvania 19104.

II. PROGRAM DESCRIPTION

The KWAC Index Generation Program Version 3.0 was implemented at the Computing Center of SUNY at Buffalo on May 10, 1973. It consists of only 252 statements for the COBOL Compiler Edition V310222 on the CDC6400 computer under the KRONOS operating system used at Buffalo. There should be few, if any, problems implementing the program at other installations which have a COBOL compiler.* The program is used only in the batch mode.

The KWAC program is essentially a KWIC index generator designed for a special purpose for use in a particular course, though it has sufficient flexibility to be used in other, similar contexts. Basically, the program takes free form natural language input, and generates an index in alphabetical order of each significant word appearing in the input alongside which appears the bibliographical description of the document in which the word was located. Output options permit free form title page, introduction, and epilog. Page headers and footers are permitted, and index page numbers are assigned and printed automatically. Determination of significant words is made in two ways: 1) a word is determined to be significant by default, if it does not appear in a user-input stop list and provided it begins with an alphabetic character and it is more than one character long; and, 2) an otherwise significant word may be eliminated from the indexing by a simple procedure at the time of input.

Input procedures are straightforward and students with minimal

---

* Anyone desiring a program listing and/or deck should make arrangements with the author of this User's Guide at SILS/SUNY at Buffalo.

key punching experience have been able to do them with very little difficulty. While no cost studies have been undertaken; for the one use already made, CPU time required for the whole index generation was less than 11 seconds. In this case there were 36 pages of index output consisting of a total of 318 index entries, generated from 36 documents. The cost of preparation of the first copy of the index (exclusive of input costs) at our installation was $1.87.

III. INPUT REQUIREMENTS

III. 1- Each document in the corpus to be indexed must have exactly three 80-column cards for input. One or two of these cards may be blank if 240 characters of input are not required, but each document <u>must</u> be represented by three cards.

III. 2- Each set of three cards may be thought of as a single 240 column "supercard". Free form data may be punched beginning at any point, but it is recommended that the first column be used as uneven left margins may occur in the printed output otherwise. The data on the three cards will be printed exactly in the same order (i.e. spelling, line endings, etc.) as the "context" beside each index entry generated from the document description. Thus it is essential that they be arranged in the correct order. Care should be taken that words not be hyphenated at the end of a card (i.e. columns 80 and 160). Further, if a word ends in column 80 or 160, a blank should be left in the first column of the card following, otherwise the two words will be joined and no index entry will be generated for the second word.

III. 3- The IBM 029 Keypunch and the CDC 6400 Printer do not always use

the same code to represent the same graphic symbol (e.g., the IBM 029 "&" code is translated to "^" by the 6400 Printer). While this is not often a problem with bibliographic data, it is recommended that "&" appearing in a document title be translated and input as the character string AND.

III. 4- The program will <u>not</u> generate entries for the following character strings:

a- Those which are only one character long

b- Those which begin with any non-alphabetic character except the colon, " : "

The printing of character strings longer than 30 characters will be truncated at that point.

III. 5- For improved readability and to link character strings that are by convention orthographically separate, but which are semantically linked (e.g. LOS ANGELES), it is recommended that a hyphen be placed between the strings instead of a blank (e.g. LOS-ANGELES). This can happen in columns 60 and 120, Paragraph 2 above notwithstanding.

III. 6- The standard COBOL sort order is used to arrange the index terms which are generated. However as only significant words are indexed, there will be none generated for character strings beginning with blanks, numbers, or special symbols with the exception of the colon " : ". In the COBOL sort the colon has low order; thus, for practical purposes the entire sort order for the first character in each string is A, B, C...Z,:. The low sort order of the colon may be exploited for several purposes. For example, it is possible to preceed each author's name with a colon as it is input. This results in all of the non-author names being generated and

sorted at beginning of the printed index, followed by all of the author's names (each preceeded by a colon) in a separate alphabetical sequence. The sort goes as far into the character string as necessary to insure complete and perfect ordering, thus the authors' names are printed in alphabetical order, just as are the title words. Similarly, if desired, some other bibliographic element (e.g., the journal title) could be preceeded by two colons (i.e. " :: ") to cause a third separate alphabetical sequence, and so on as many times as necessary. It has been observed that the colon preceeding an author's name is not very obtrusive and that it does not seriously interfere with the scanning of a list of such names.

III. 7- Any character string may be made non-significant simply by preceeding it with any non-alphabetical character except the colon (see Paragraph 6). However, it is recommended that the logical not "¬" always be used for this purpose to exploit a special feature of the program. Any character punched in the input cards will be printed exactly as it was punched (see Paragraph 2), with one exception: the logical not "¬"; and then only when a further condition is met. In any card on which a logical not "¬" appears, and where it is not desirable to have it printed, it can be suppressed (i.e., replaced by a blank) simply by putting <u>another</u> logical not "¬" <u>in the last column of that card</u>. When this is done, the printing of both of them is suppressed.

III. 8- Many other modifications of an additions to the raw bibliographic data can be made before input in order to produce a better index. Indexes of the KWIC family are "quick and dirty" and in this

statement is revealed both their great virtue and their great fault. They may be made quickly and inexpensively; but, unfortunately they are far from perfect retrieval tools and are ordinarily used only where these two characteristics are demanded. If better retrieval tools are needed, it is recommended that another indexing method be used. Consequently only the index improvements mentioned above should be made.

IV. DECK MAKEUP TO GENERATE A KWAC INDEX

An assembled deck to generate and print a KWAC index consists of the following parts:

1- Job control cards,

2- KWAC control cards, and

3- Data input cards.

IV. A. JOB CONTROL CARDS

Job control cards are computer installation specific. Do whatever is necessary to have the job accepted and to invoke the COBOL compiler. Use of the program thus far has been satisfactory with a field length of 55K; but this is dependent upon the amount of space required for the COBOL sort which is in turn dependent upon the number of index entries generated. (See Appendix A for the appropriate SUNY at Buffalo job control cards).

IV. B. KWAC CONTROL CARDS AND DATA INPUT CARDS

The data input cards and the control cards for the KWAC program are not separated from each other. The following list of control

and data cards is arranged in the proper order for the making up of a deck for generating a KWAC index.

Control cards contain one or more characters and must begin in column 1 of each card. They may extend as far as necessary to the right with the exceptions noted below. Data input cards may have a free form input with the exceptions noted within their descriptions below. On control cards, numbers must fill columns 1-3, with leading zeros supplied if necessary.

IV. B. 1- PREFACE LINE SPACING CONTROL CARD indicates the number of blank lines to appear between each line of printing in the preface.

IV. B. 2- EPILOG LINE SPACING CONTROL CARD indicates the number of blank lines to appear between each line of printing in the epilog.

IV. B. 3- BEGINNING OF PREFACE MARKER (*) is required. An asterisk must appear in column 1, even if there is no preface.

IV. B. 4- PREFACE DATA INPUT CARDS must appear in pairs. As the preface will be printed exactly as it is input, certain precautions must be observed in the punching of this data. A card has 80 columns but the width of the printed line of output is limited to 136 characters. Accordingly, for the purposes of centering the data of the preface in the printed output, it is necessary to think of the midpoint of the printed line as falling between column 68 and 69 of the first of two cards. For example, if it were desired to have the word PREFACE centered at the top of the first page of the preface, the following procedures should be used. The word PREFACE contains seven characters. Subtract 7 from 136, yielding 129. Divide by 2 yielding 64.5; thus, the word PREFACE should begin in either column 64 or 65 for approximate centering. Other lines may be centered and left-or right- justified by

similar methods.

IV. B. 5- ENDING OF PREFACE MARKER (*) is required, even if there is no preface.

IV. B. 6- HEADER/FOOTER OPTION CONTROL CARD. A pound sign (#) is required if headers and footers are to be printed on each page of the index. If this option is not chosen, a blank card must be used in its place.

IV. B. 7- HEADER/FOOTER INPUT DATA CARDS must be used if there is a # in column 1 of the preceeding card. If used, there must be exactly two cards with data entered as follows. The word PAGE and the page number are automatically generated and printed at the center of both the top and the bottom of each index page. As each of the two cards is printed exactly as the data is entered into the cards, care should be taken at input to insure appropriate spacing of the printed output.

IV. B. 8- NUMBER OF STOPWORDS CONTROL CARD must be used to specify the number of stopwords to be input (i.e., the number of words to be declared non-significant throughout the entire index generation). The maximum number that may be declared is 153. If more are needed, use the method described in Paragraph III. 7. above. A stopped word can be unstopped by adding two periods (..) immediately after it (i.e., creating a non-stopword character string). This is required because the program automatically searches for a punctuation mark at the end of each word and strips it off.

IV. B. 9- STOPWORD LIST INPUT DATA requires as many cards as the number specified on the preceeding card. Only one stopword may be entered on each card, and the first letter must be in column 1. The maximum length of any stopword is 20 characters.

IV. B. 10- BIBLIOGRAPHIC DATA INPUT CARDS may be in free form format, but see Paragraph III. 2. above for suggestions. Each logical record must consist of 3 physical cards, with blank cards used if necessary.

IV. B. 11- END OF DATA MARKER (*) in column 1 must follow the last bibliographic data input card.

IV. B. 12- EPILOG INPUT DATA CARDS have exactly the same format requirements as the preface input data cards discussed in Paragraph IV. B. 4. above.

APPENDIX A

JOB CONTROL CARDS REQUIRED AT SUNY AT BUFFALO
(KRONOS version 2.0.9 on a CDC 6400)

Column 1

```
BATCH,T=40,F=55000,P=30,R=E. CONAWAY--JOB NAME.
LISCHAS,CONAWAY,PASSWORD.
COBOL(LR)
REDUCE,NO.
LGO.
7-8-9*
(KWAC Source Deck)
7-8-9*
(Preface Line Spacing Control Card)
(Epilog Line Spacing Control Card)
(Beginning of Preface Marker)
(Preface Data Input Cards)
(Ending of Preface Marker Card)
(Header/Footer Option Control Card)
(Header/Footer Input Data Cards)
(Number of Stopwords Control Card)
(Stopword List Input Data Cards)
(Bibliographic Data Input Cards)
(End of Data Marker)
(Epilog Input Data Cards)
7-8-9*
6-7-8-9*
```

---

* 6-7-8-9 and 7-8-9 cards have all digits multipunched in column 1.

APPENDIX B

Sample Output Pages

HUMAN
::MCALLISTER,CARYL ::BELL,JOHN-M
HUMAN FACTORS IN THE DESIGN OF AN INTERACTIVE LIBRARY SYSTEM
:::(JASIS,V22,1971,P096-104)

IMPLICATIONS
::SHOFFNER, RALPH-M
SOME IMPLICATIONS OF AUTOMATIC RECOGNITION OF BIBLIOGRAPHIC ELEMENTS
:::(JASIS,V22,1971,P275-282)

IMPLICATIONS
::WILLIAMS,J-H
FUNCTIONS OF A MAN-MACHINE INTERACTIVE INFORMATION RETRIEVAL SYSTEM SOME IMPLICA
TIONS OF AUTOMATIC RECOGNITION OF BIBLIOGRAPHIC ELEMENTS ::NJASIS,V22,1971,P311

IMPROVING
::PAISLEY,WILLIAM
IMPROVING A FIELD-BASED ERIC-LIKE INFORMATION SYSTEM
:::(JASIS,V22,1971,P399-408)

INDEX
::ROSENBERG,VICTOR
A STUDY OF STATISTICAL MEASURES FOR PREDICTING TERMS USED TO INDEX DOCUMENTS
:::(JASIS,V22,1971,P041-050)

INDEXES
::AULD,LARRY
KWOC INDEXES A VOCABULARY COMPARISONS OF SUMMARIES OF LC AND
CLASSIFICATION SCHEDUALS :::(JASIS,V22,1971,P322-325)

INDEXES
::BLANKEN,ROBERT-R THE PREPARATION OF INTERNATIONAL AUTHOR INDEXES, WITH
PARTICULAR REFERENCE TO THE PROBLEMS OF TRANSLITERATION, PREFIXES, AND COMPOUND
FAMILY NAMES :::(JASIS,V22,1971,P051-063)

INDEXING
::BL[illegible],FRED
TWO MACHINE INDEXING PROJECTS AT THE CATHOLIC UNIVERSITY OF AMERICA
:::(JASIS,V22,1971,P185-186)

INDEXING
::ROSENBERG,VICTOR
COMPARATIVE EVALUATION OF TWO INDEXING METHODS USING JUDGES
:::(JASIS,V22,1971,P251-259)

BRIGGS ELEANORE — ::JOHNSON,CLAIRE ::BRIGGS,ELEANORE
HOLOGRAPHY AS APPLIED TO INFORMATION STORAGE AND RETRIEVAL SYSTEMS
:::(JASIS,V22,1971,P187-192)

COOPER MICHAEL-D — ::LEIMKUHLER,FERDINAND-F ::COOPER,MICHAEL-D
ANALYTICAL MODELS FOR LIBRARY PLANNING
:::(JASIS,V22,1971,P390-398)

CRAVENS DAVID-W — ::CRAVENS,DAVID-W
PREDICTING PERFORMANCE OF INFORMATION SPECIALISTS
:::(JASIS,V22,1971,P805-811)

CRAWFORD SUSAN — ::CRAWFORD,SUSAN
INFORMAL COMMUNICATION AMONG SCIENTISTS IN SLEEP RESEARCH
:::(JASIS,V22,1971,P301-310)

HARMON GLYNN — ::HARMON,GLYNN
OPINION PAPER ON THE EVOLUTION OF INFORMATION SCIENCE
:::(JASIS,V22,1971,P235-241)

HELMUTH NANCY-A — ::HELMUTH,NANCY-A
THE USE OF EXTRACTS IN INFORMATION SERVICES
:::(JASIS,V22,1971,P382-389)

HILLINGER CLAUDE — ::KRAUZE,TADEUSZ-K ::HILLINGER,CLAUDE
CITATIONS,REFERENCES AND THE GROWTH OF SCIENTIFIC LITERATUREA MODEL OF DYNAMIC INTERACTION :::(JASIS,V22,1971,P333-336)

HOLM B-E — ::HOLM,B-E
FID COMMITTEE INFORMATION FOR INDUSTRY (FID/II)
:::(JASIS,V22,1971,P409-410)

JACKSON E-B — ::JACKSON,E-B
FID AT BUENOS AIRES, SEPTEMBER 14 TO 24, 1970
:::(JASIS,V22,1971,P064)

APPENDIX C

Sample Bibliographic Data Input Card Listing

::MCALLISTER,CARYL ::BELL,JOHN-M
HUMAN FACTORS IN THE DESIGN OF AN INTERACTIVE LIBRARY SYSTEM
:::(JASIS,V22,1971,P096-104)

::SHOFFNER, RALPH-M
SOME IMPLICATIONS OF AUTOMATIC RECOGNITION OF BIBLIOGRAPHIC ELEMENTS
:::(JASIS,V22,1971,P275-282)

::WILLIAMS,J-H
FUNCTIONS OF A MAN-MACHINE INTERACTIVE INFORMATION RETRIEVAL SYSTEM SOME IMPLIC
TIONS OF AUTOMATIC RECOGNITION OF BIBLIOGRAPHIC ELEMENTS :::NJASIS,V22,1971,P31

::PAISLEY,WILLIAM
IMPROVING A FIELD-BASED ERIC-LIKE INFORMATION SYSTEM
:::(JASIS,V22,1971,P399-408)

::ROSENBERG,VICTOR
A STUDY OF STATISTICAL MEASURES FOR PREDICTING TERMS USED TO INDEX DOCUMENTS
:::(JASIS,V22,1971,P041-050)

::AULD,LARRY
KWOC INDEXES ^ VOCABULARY COMPARISONS OF SUMMARIES OF LC ^DC
CLASSIFICATION SCHEDUALS :::(JASIS,V22,1971,P322-325)

::BLANKEN,ROBERT-R THE PREPARATION OF INTERNATIONAL AUTHOR INDEXES, WITH
PARTICULAR REFERENCE TO THE PROBLEMS OF TRANSLITERATION, PREFIXES, AND COMPOUND
FAMILY NAMES :::(JASIS,V22,1971,P051-063)

::BLUM,FRED
TWO MACHINE INDEXING PROJECTS AT THE CATHOLIC UNIVERSITY OF AMERICA
:::(JASIS,V22,1971,P105-106)

::ROSENBERG,VICTOR
COMPARATIVE EVALUATION OF TWO INDEXING METHODS USING JUDGES
:::(JASIS,V22,1971,P251-259)

APPENDIX D

KNOWN BUGS IN KWAC (JULY 24, 1973)

1- The footer, but not the header, loses character in column 1 in the printed index. To avoid the problem, begin the header/footer input data in column 2.

2- A character string ending in column 80 of the last card of a set of 3 bibliographic data input cards will not be indexed. Avoid the bug by always leaving the last column of the last card blank.