U. S. DEPARTMENT OF INTERIOR

U.S. GEOLOGICAL SURVEY

[Reports - Open File series]

A computer program for creating keyword indexes to textual data files

by

D. W. Moody

Open-File Report

Washington, D. C.

June 1972

CONTENTS

ILLUSTRATIONS

ABSTRACT

A keyword-in-context (KWIC) or out-of-context (KWOC) index is a convenient means of organizing information. This keyword index program can be used to create either KWIC or KWOC indexes of bibliographic references or other types of information punched on cards, typed on optical scanner sheets, or retrieved from various Department of Interior's data bases using the Generalized Information Processing System (GIPSY). The index consists of a "bibliographic" section and a keyword section based on the permutation of document titles, project titles, environmental impact statement titles, maps, etc. or lists of descriptors. The program can also create a back-of-the-book index to documents from a list of descriptors. By providing the user with a wide range of input and output options, the program provides the researcher, manager, or librarian with a means of maintaining a list and index to documents in a small library, reprint collection, or office file.

INTRODUCTION

Machine produced indexes have been used for more than a decade to inexpensively organize information. The most common types of indexes prepared by the computer are the keyword-in-context (KWIC) index and the keyword-out-of-context (KWOC) index. Both indexes are produced in much the same way. A string of characters, such as the title of a bibliographic citation, is fragmented into substrings (e.g., words) using break characters such as blanks, commas, and periods. The words are then used as index terms. Because many words occur so frequently as to have little information value, provision is usually made to eliminate them by providing the computer with a list of stop words, i.e., words which are not to be used as index terms. The remaining terms are referred to as keywords.

The KWIC index format lists the keywords alphabetically in the center of the index page. The remaining words in the title are printed to the left and right of the keyword in their original order. If there is insufficient space on the line for the words to the right of the keyword in the title, the title is wrapped around and printed beginning at the far left of the same line. Thus, the keyword appears in the context in which it is used in the title. Each line is identified by a number which references the full citation of the title in the bibliographic section of the index.

The KWOC index format presents each keyword alphabetically in the left margin of the index followed by a list of all the titles which contain that keyword. In this case the keyword is out of context with the title.

This report describes the use of a keyword index program to produce KWIC and KWOC indexes of document titles punched on cards, typed on optical scanner sheets, or retrieved from various Department of the Interior's data bases using the Generalized Information Processing System (GIPSY). While bibliographic records are used as illustrations, the program can also be used to produce indexes of projects, maps, computer programs, geologic names, index terms, or any other information contained in a GIPSY data base or punched on cards in the keyword index program format. Instructions are given for preparing and maintaining a personal bibliographic file.

PROGRAM ORGANIZATION

This keyword index program is written in PL/1 for the IBM 360/65 computer and is a modification of the University of California's Quick Index Compiler (IBM SHARE Program Library 360D 03.3.002). The program is organized into 3 parts. The first part of the program reads a set of control cards, a list of stop words, and the bibliographic data. The program then breaks the title or other information into words, removes stop words, and prepares two output files: the file named OUTPUT contains the bibliographic section of the index; the file named INPUT contains the keyword index. Note that indexing will occur on words followed by a pound sign (#), blank, period, comma, double apostrophe, left and right parentheses, hyphen, semi-colon, colon, question mark, and quote mark.

The second part of the program uses the cataloged procedure DISKSORT to alphabetize the keywords in file INPUT and add the sorted index to file OUTPUT. The third part of the program prints the bibliographic and keyword sections of the index contained in file OUTPUT.

PROGRAM OPTIONS

Control Cards

The following control words are punched on separate cards beginning in column 1. Numbers are separated from the control words by a blank. The order of the cards is unimportant with the exception that 'STOP WORDS' must precede the control card 'TITLES'. The 'TITLES' card is the only card required. All others are optional.

PRINT BIBLIOGRAPHY - This card causes all citations to be printed. Each citation is separated by a double space. If this option is used, the input data should be sorted by reference number prior to using the program.

LENGTH n - The length parameter sets the index line length to n characters. The maximum line length is 132 characters. If no length is specified, n defaults to 80.

PAGESIZE n - This card specifies the number of lines to be printed on each page of the index. If the page size is not specified, n defaults to 60.

COPIES n - This card specifies the number of copies of the index to be printed. If the number of copies is not specified, n defaults to one.

STORAGE n - The storage card specifies the maximum length of a character string that can be processed by the program. The default value of n, the number of characters in the string, is 180. If using punch cards (or optical scanning sheets) as input, 180 characters will be sufficient to handle document titles which fill 3 cards (60 characters per card). Any excess characters will be ignored unless the STORAGE parameter is increased in multiples of 60. If using input from a GIPSY retrieval, the default value will accommodate the first 180 characters of the document title record. STORAGE may be set to any value; however, it must equal the length of the "title" string put out by the COPY command in the GIPSY retrieval (see p. 13). In this case the record length and the block size parameters

on the DATA DD card (job control language card) will also have to be adjusted (see p. 14).

COLUMN - The column command is used with the INDEX TERMS to format a double column list of terms. It should only be used with this option.

MARGIN - The margin command disables the automatic margin adjustments made for line lengths less than 132 characters. In general this card will not be used.

SCANNER INPUT - The program will expect an input data set on a tape or disk pack whose file name is defined as DATA by the Job Control Language. The input data must be in the optical scanner format.

GIPSY INPUT - The program will expect an input data set on a tape or disk pack whose file name is defined as DATA by the Job Control Langauge. The input data must be in the GIPSY copy format.

(Default Input)- If the type of input is not specified, the program assumes that the data is punched on cards in the format required

by the program. Data cards immediately follow the TITLES cards.[1/]

KWIC INDEX - The program will produce a keyword-in-context index.

KWOC INDEX - The program will produce a keyword-out-of-context index using terms from the citation titles.

KWOC TERMS - The program will produce an index in KWOC format using the index terms. Only the first 10 characters of each term word is used.[2/]

INDEX TERMS c - The program will produce a back-of-the-book type of index from the list of index terms associated with a citation. If "c" is a blank character (col. 13), then every term will be indexed. If term sets are separated by a comma or other punctuation, then place this character in column 13 and the term set will be treated as a single unit. A maximum of 25 characters from the multiple-term descriptor is printed.

1/ If bibliography is stored on disk in card-imate format, use DISK INPUT on control card. The program will expect an input data set on a disk pack whose file name is defined as DATA by the Job Control Language.

2/ To use this option a list of descriptors (index terms) must be associated with each bibliographic citation.

STOP WORDS - Up to 500 words may be punched on separate cards following this one. Each word begins in column 1. Only the first 10 characters (including blanks) of each word are used for comparison. Columns 19-80 should be blank.

TITLES - The 'TITLES' card is the only required control card. Any information punched in columns 11-60 will be used as a heading for each page of the bibliography section.

INDEX AUTHORS - The author(s) names will be indexed with the keywords in the index.

Card Input

Card input is assumed by the program, if no other input option is specified. Sets of one or more punched cards must immediately follow the 'TITLES' card. The format of the punched cards is as follows:

Cols.

1-60 Text. The text on each card may be continued on up to nine additional cards as indicated by the card's sequence number. Because the text is printed exactly as it appears on the card, do not break text in the middle of a word at the end of a line. The punctuation used in the text depends upon the type of information contained in the card (see types of cards listed below).

61 Type of card.

'1' - Author card. Place author's last name first followed by a comma, first name or initials, and a semi-colon. Repeat for each author to be listed. For example: 'JOE, JOHN; SMITH, A. A.;'

'2' - Title card. Place colon (or comma) at end of title.

'3' - Citation. Use standard USGS bibliographic format.

'4' - Descriptor or index term card. Separate

terms by commas. For example: 'WATER, CHEMISTRY, ATLANTIC OCEAN, WATER RESOURCES.' End the term list with a period.

'5' - Identifiers terms (or other information) related to card 4. If card type 5 appears, it will be combined with the information on card type 4. Otherwise, do not include card type 5.

Other types of cards may be added as needed (card type 6 - 9).

62-63 Sequence number. Number successive cards of the same type from 01-10.

64-73 Reference code of document. This is a 10 character alphanumeric code which appears next to each reference in the bibliography and to the right of each line in the index.

Optical Scanner Input

Bibliographic data can also be prepared on optical scanning sheets using a 10-pitch IBM Selectric typewriter equipped with an OCS typing element. The following format is used:

Cols.

1 Type of card:

'1' - Author card

'2' - Title card

'3' - Citation card

'4' - Descriptor term card

'5' - Identifier term card

Other types of cards may be added as needed (card type 6-9).

2-3 Sequence number. Number successive cards of the same type from 01-10.

4-13 Reference code of document.

14-73 Text. The text may be continued on up to nine continuation cards.

Since the optical scanner reads up to 75 characters on a line, there is space to make two corrections of one letter each on any given line of input text. Otherwise, the entire line must be deleted. After the citations are typed they are read by the optical scanner and placed as card images on magnetic tape. The tape can then be processed by the keyword program.

GIPSY Input

Records can also be used as input to the keyword program. It is assumed that the user of this version of the program has read the GIPSY application manual (Addison and others, 1969) and is familiar with the GIPSY cataloged procedures[1/] on the Geological Survey's IBM 360/65.

1/ Copies of the GIPSY cataloged procedures may be obtained from Chief, Office of Systems Research and Development, Computer Center Division, Geological Survey, Washington, D. C.

The GIPSY program is an integral part of the Geological Survey's Bibliographic System. Using GIPSY the user can search the title, index terms, or abstracts of bibliographic references for specific combinations of words, word stems, or phrases. The selected records can then be sorted by author or accession number and printed. A key-word index of these records provides a convenient means of scanning the retrieved documents for items of interest.

The user prepares the key-word program input data set by (1) making a GIPSY search of one of the GIPSY data files, (2) sorting the retrieved records by reference number, and (3) placing the records into a temporary disk storage area on the computer system, e.g., TEMP01. The temporary data set is then ready for processing by the keyword program. The use of the TEMP01 disk pack is described in the U.S.G.S. Computer User's Manual (p. 5.3), and the necessary job control language instructions and GIPSY commands are shown in figures 1 and 2.

The principal limitations to this use of the keyword program are the length of the data elements copied, e.g., AUTHOR, TITLE, CIT, and DESCR from the bibliographic file, and the number of records selected. Only the first 180 characters of the author, title, reference, and index terms fields are copied. If any of these fields are more than 180 characters, the remaining characters will be dropped. The total length of each record is 193 characters.

This includes the "card type," sequence number, and the 10-character reference code. Sixty tracks of temporary disk space is sufficient to store 740 references. If larger character strings are copied, the STORAGE parameter will need to be set accordingly. The length of record parameter on the DATA DD card will also need to be changed to reflect the increase in the length of the record copied on to the temporary disk. The new length will be equal to STORAGE plus 13 characters, e.g., LRECL=STORAGE+13. Similarly, the BLKSIZE parameter will need to be adjusted. The overall effect of these changes will be to reduce the number of records which can be stored on a disk track.

The required job control language cards to execute the keyword program are shown in figure 3.

```
//.... JOB (...)
/*SETUP     CCDnnn/DISK
//JOBLIB DD DSN=SYS1.GIPLIB,DSN=SHR
//STEP1 EXEC QUESTRAN,DNAME=dictionary,DVOL=CCDnnn,
// RNAME=recordfile,RVOL=CCDnnn
//QUESTRAN.SYSWRKO DD UNIT=SYSDA,DISP=(NEW,KEEP,DELETE),
// VOL=SER=TEMP01,DCB=(RECFM=FB,LRECL=193,BLKSIZE=7141),
// SPACE=(TRK,(10,10),RLSE),DSN=Annnnnn.aaaB200.name
//QUESTRAN.SYSRDR DD *

      (Gipsy control cards go here.)

/*
```

Figure 1. -- Job Control Language cards needed to make a GIPSY search and create a data set for input to the keyword program.

```
Col. 1
|

FORM
 WRSIC
SELECT
 A. DESCR <....>
 B. $ <....>
 .
 .
 .
 LOGIC A AND B AND ...
SORT
 NUMBER 10
COPY
 AUTHOR 180
 '101'
 NUMBER 10
 NEW RECORD
 TITLE 180
 '201'
 NUMBER 10
 NEW RECORD
 CIT 180
 '301'
 NUMBER 10
 NEW RECORD
 DESCR 180
 '401'
 NUMBER 10

 Selected Water Resources
 Abstracts
```

Figure 2. -- GIPSY control cards needed to search and copy selected records from Selected Water Resources Abstracts.

```
//.... JOB (...)
//STEP1 EXEC PL1LFCLG,PARM.PL1L='SIZE=0077824,ST,A,NT,X',
// TIME.PL1L=(0,30),TIME.LKED=(0,5),TIME.GO=(2,10),
// REGION.GO=193K
//PL1L.SYSIN DD *

     (Source deck KWIC program- part 1)

/*
//GO.OUTPUT DD UNIT=SYSDK,DISP=(NEW,PASS,DELETE),
// SPACE=(TRK,(60,20)),DCB=(RECFM=FB,LRECL=133,BLKSIZE=7182),
// DSN=&&OUTPT
//GO.INPUT DD UNIT=SYSDK,DISP=(NEW,PASS,DELETE),
// SPACE=(TRK,(30,10)),DCB=(RECFM=FB,LRECL=133,BLKSIZE=7182),
// DSN=&&INPT
//GO.DATA DD UNIT=SYSDA,DISP=SHR,
// VOL=SER=TEMP01,DSN=Annnnnn.aaaB200.name
//GO.SYSIN DD *

TITLES     (Must be last control card

           (If CARD input option is used place cards
            here and "dummy" the "GO.DATA" DD card.)

/*

//STEP2 EXEC PROC=DISCSORT,REGION=100K,TIME=(0,30)
//SORTIN DD DISP=(OLD,DELETE,DELETE),DSN=&&INPT
//SORTOUT DD DISP=(MOD,PASS,DELETE),DSN=&&OUTPT
//SYSIN DD *

 SORT FIELDS=(62,26,CH,A),SIZE=E30000     (Use for KWIC index.)

 SORT FIELDS=(6,10,CH,A),SIZE=E30000      (Use for KWOC index.)

 SORT FIELDS=(21,25,CH,A),SIZE=E30000     (Use for INDEX TERMS.)

/*

//STEP3 EXEC PL1LFCLG,TIME.PL1L=(0,10),TIME.Lked=(,5),
// TIME.GO=(,30),REGION.GO=100K
//PL1L.SYSIN DD *

     (Source deck KWIC program- part 2)

/*
//GO.PRINTER DD SYSOUT=A,DCB=(RECFM=UA,BLKSIZE=133)
//GO.OUTPUT DD DISP=(OLD,DELETE,DELETE),DSN=&&OUTPT
//GO.SYSIN DD DUMMY,DCB=(RECFM=F,BLKSIZE=80)
/*
```

Figure 3. -- Job Control Language cards needed to execute the keyword program.

APPLICATIONS

The card version of the keyword program can be used to develop an index to a personal reprint collection or a small office library. Each item is given a reference code as it is received which identifies the item's owner (or location) and perhaps the date it was received. Thus, if more than one collection were to be merged for indexing purposes, the reference code will identify the report's location. The following reference code (columns 64-73 on card input) might be used to identify and locate documents

Cols.	
64-66	Document owner's initials (3 characters)
67	Hyphen
68-69	Year reprint received or year of publication
70	Hyphen
71-73	Accession number of reprint within year

For example, JDD-70-001 represents the first reprint received by John D. Doe in 1970. When submitting cards for processing, the citations must be in accession number order. By placing each private collection in alphabetical order by the owner's initials, more than one collection can be merged into one index. Thus, as research personnel move from office to office they can take their index with them and merge with existing collections at their next assignment. Figures 4 to 7 are examples of indexes from a private reprint collection.

UNDERGROUND WASTE DISPOSAL - CONCEPTS AND MISCONCEPTIONS.
ENVIRONMENTAL SCIENCE AND TECHNOLOGY, V. 4, NO. 8, P. 642-647, 1970.
GROUNDWATER, POLLUTION, WASTE DISPOSAL, DEEPWELL DISPOSAL.

DWM-70-002 DUBBINS, D. A.; RAGLAND, P. C.; JOHNSON, J. D.;
WATER-CLAY INTERACTIONS IN NORTH CAROLINA'S PAMLICO ESTUARY:
ENVIRONMENTAL SCIENCE AND TECHNOLOGY, V. 4, NO. 9, P. 743-748, 1970.
ESTUARY, SEDIMENT, WATER QUALITY, SALINITY, SEDIMENTATION, ION EXCHANGE, SUSPENDED SEDIMENT, NORTH CAROLINA.

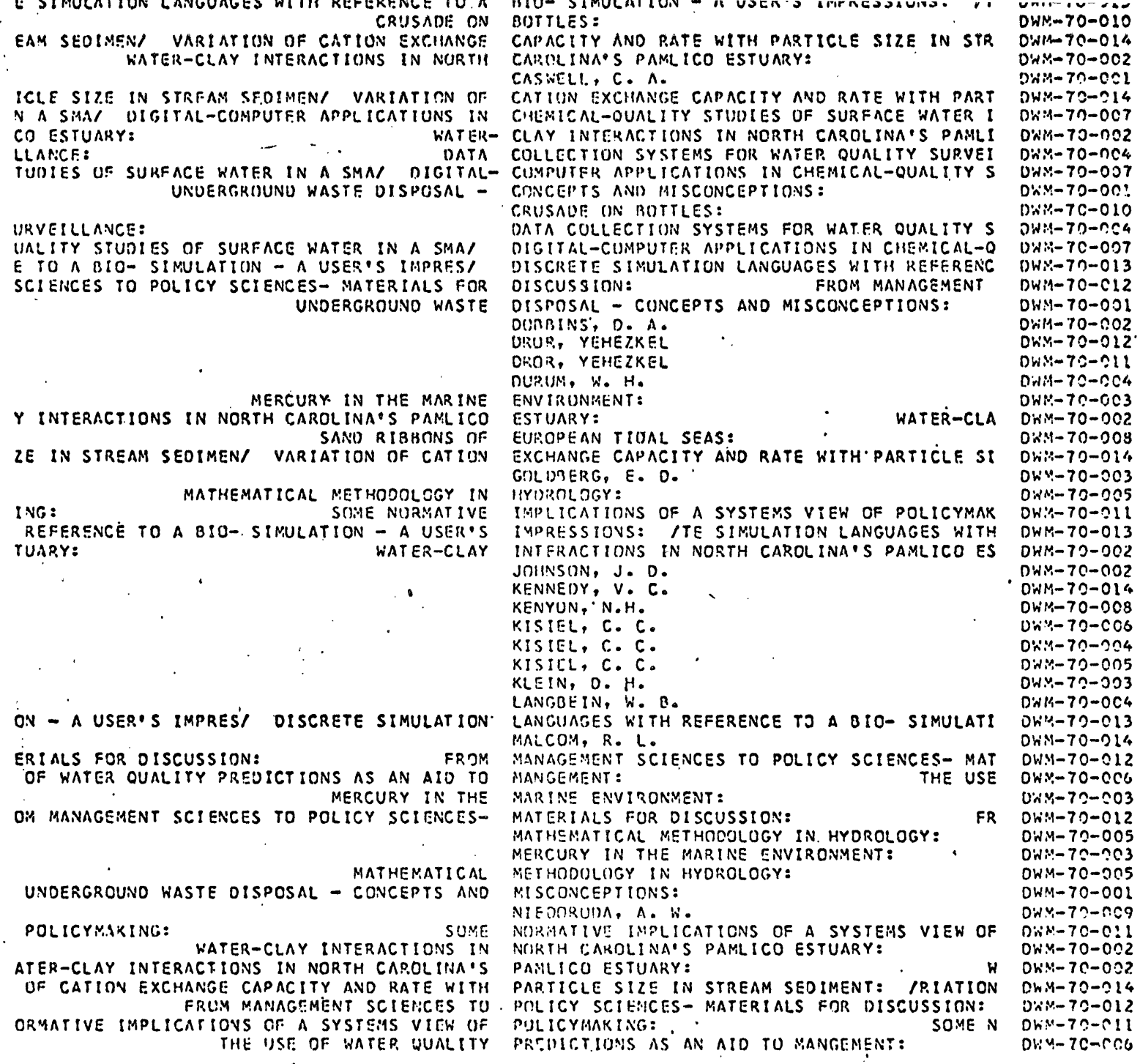
DWM-70-003 KLEIN, D. H.; GOLDBERG, E. D.;
MERCURY IN THE MARINE ENVIRONMENT:
ENVIRONMENTAL SCIENCE AND TECHNOLOGY, V. 4, NO. 9, P. 765-768, 1970.
MERCURY, MARINE ORGANISMS, SEDIMENTS, CALIFORNIA, POLLUTION.

DWM-70-004 KISIEL, C. C.; DURUM, W. H.; LANGBEIN, W. B.;
DATA COLLECTION SYSTEMS FOR WATER QUALITY SURVEILLANCE:
U.S. GEOL. SURVEY, UNPUBLISHED MANUSCRIPT, 21 P., 1968.
WATER QUALITY, DATA COLLECTION, DATA PROGRAMS, SAMPLING THEORY, NETWORK DESIGN.

DWM-70-005 KISIEL, C. C.;
MATHEMATICAL METHODOLOGY IN HYDROLOGY:
INTERNAT. SEMINAR FOR HYDROLOGY, PROC., JULY 1969, URBANA, ILLINOIS, P. 362-399.
EDUCATION; MODELS; MATHEMATICAL MODELS, HYDROLOGIC MODELS.

DWM-70-006 KISIEL, C. C.;
THE USE OF WATER QUALITY PREDICTIONS AS AN AID TO

Figure 4. Example of the bibliographic section of the index.



E SIMULATION LANGUAGES WITH REFERENCE TO A
CRUSADE ON
EAM SEDIMEN/ VARIATION OF CATION EXCHANGE
WATER-CLAY INTERACTIONS IN NORTH

ICLE SIZE IN STREAM SEDIMEN/ VARIATION OF
N A SMA/ DIGITAL-COMPUTER APPLICATIONS IN
CO ESTUARY: WATER-
LLANCE: DATA
TUDIES OF SURFACE WATER IN A SMA/ DIGITAL-
UNDERGROUND WASTE DISPOSAL -

URVEILLANCE:
UALITY STUDIES OF SURFACE WATER IN A SMA/
E TO A BIO- SIMULATION - A USER'S IMPRES/
SCIENCES TO POLICY SCIENCES- MATERIALS FOR
UNDERGROUND WASTE

MERCURY IN THE MARINE
Y INTERACTIONS IN NORTH CAROLINA'S PAMLICO
SAND RIBBONS OF
ZE IN STREAM SEDIMEN/ VARIATION OF CATION

MATHEMATICAL METHODOLOGY IN
ING: SOME NORMATIVE
REFERENCE TO A BIO- SIMULATION - A USER'S
TUARY: WATER-CLAY

ON - A USER'S IMPRES/ DISCRETE SIMULATION

ERIALS FOR DISCUSSION: FROM
OF WATER QUALITY PREDICTIONS AS AN AID TO
MERCURY IN THE
OM MANAGEMENT SCIENCES TO POLICY SCIENCES-

MATHEMATICAL
UNDERGROUND WASTE DISPOSAL - CONCEPTS AND

POLICYMAKING: SOME
WATER-CLAY INTERACTIONS IN
ATER-CLAY INTERACTIONS IN NORTH CAROLINA'S
OF CATION EXCHANGE CAPACITY AND RATE WITH
FROM MANAGEMENT SCIENCES TO
ORMATIVE IMPLICATIONS OF A SYSTEMS VIEW OF
THE USE OF WATER QUALITY

BIO- SIMULATION - A USER'S IMPRESSIONS: /T DWM-70-013
BOTTLES: DWM-70-010
CAPACITY AND RATE WITH PARTICLE SIZE IN STR DWM-70-014
CAROLINA'S PAMLICO ESTUARY: DWM-70-002
CASWELL, C. A. DWM-70-001
CATION EXCHANGE CAPACITY AND RATE WITH PART DWM-70-014
CHEMICAL-QUALITY STUDIES OF SURFACE WATER I DWM-70-007
CLAY INTERACTIONS IN NORTH CAROLINA'S PAMLI DWM-70-002
COLLECTION SYSTEMS FOR WATER QUALITY SURVEI DWM-70-004
COMPUTER APPLICATIONS IN CHEMICAL-QUALITY S DWM-70-007
CONCEPTS AND MISCONCEPTIONS: DWM-70-001
CRUSADE ON BOTTLES: DWM-70-010
DATA COLLECTION SYSTEMS FOR WATER QUALITY S DWM-70-004
DIGITAL-COMPUTER APPLICATIONS IN CHEMICAL-Q DWM-70-007
DISCRETE SIMULATION LANGUAGES WITH REFERENC DWM-70-013
DISCUSSION: FROM MANAGEMENT DWM-70-012
DISPOSAL - CONCEPTS AND MISCONCEPTIONS: DWM-70-001
DOBBINS, D. A. DWM-70-002
DROR, YEHEZKEL DWM-70-012
DROR, YEHEZKEL DWM-70-011
DURUM, W. H. DWM-70-004
ENVIRONMENT: DWM-70-003
ESTUARY: WATER-CLA DWM-70-002
EUROPEAN TIDAL SEAS: DWM-70-008
EXCHANGE CAPACITY AND RATE WITH PARTICLE SI DWM-70-014
GOLDBERG, E. D. DWM-70-003
HYDROLOGY: DWM-70-005
IMPLICATIONS OF A SYSTEMS VIEW OF POLICYMAK DWM-70-011
IMPRESSIONS: /TE SIMULATION LANGUAGES WITH DWM-70-013
INTERACTIONS IN NORTH CAROLINA'S PAMLICO ES DWM-70-002
JOHNSON, J. D. DWM-70-002
KENNEDY, V. C. DWM-70-014
KENYON, N.H. DWM-70-008
KISIEL, C. C. DWM-70-006
KISIEL, C. C. DWM-70-004
KISIEL, C. C. DWM-70-005
KLEIN, D. H. DWM-70-003
LANGBEIN, W. B. DWM-70-004
LANGUAGES WITH REFERENCE TO A BIO- SIMULATI DWM-70-013
MALCOM, R. L. DWM-70-014
MANAGEMENT SCIENCES TO POLICY SCIENCES- MAT DWM-70-012
MANGEMENT: THE USE DWM-70-006
MARINE ENVIRONMENT: DWM-70-003
MATERIALS FOR DISCUSSION: FR DWM-70-012
MATHEMATICAL METHODOLOGY IN HYDROLOGY: DWM-70-005
MERCURY IN THE MARINE ENVIRONMENT: DWM-70-003
METHODOLOGY IN HYDROLOGY: DWM-70-005
MISCONCEPTIONS: DWM-70-001
NIEDORUDA, A. W. DWM-70-009
NORMATIVE IMPLICATIONS OF A SYSTEMS VIEW OF DWM-70-011
NORTH CAROLINA'S PAMLICO ESTUARY: DWM-70-002
PAMLICO ESTUARY: W DWM-70-002
PARTICLE SIZE IN STREAM SEDIMENT: /RIATION DWM-70-014
POLICY SCIENCES- MATERIALS FOR DISCUSSION: DWM-70-012
POLICYMAKING: SOME N DWM-70-011
PREDICTIONS AS AN AID TO MANGEMENT: DWM-70-006

Figure 5. Example of the KWIC index.

CAPACITY	VARIATION OF CATION EXCHANGE CAPACITY AND RATE WITH PARTICLE SIZE IN S	DWM-70-014
CAROLINA	WATER-CLAY INTERACTIONS IN NORTH CAROLINA'S PAMLICO ESTUARY:	DWM-70-002
CATION	VARIATION OF CATION EXCHANGE CAPACITY AND RATE WITH PARTICLE SIZE IN S	DWM-70-014
CHEMICAL	DIGITAL-COMPUTER APPLICATIONS IN CHEMICAL-QUALITY STUDIES OF SURFACE W	DWM-70-007
CLAY	WATER-CLAY INTERACTIONS IN NORTH CAROLINA'S PAMLICO ESTUARY:	DWM-70-002
COLLECTION	DATA COLLECTION SYSTEMS FOR WATER QUALITY SURVEILLANCE:	DWM-70-004
COMPUTER	DIGITAL-COMPUTER APPLICATIONS IN CHEMICAL-QUALITY STUDIES OF SURFACE W	DWM-70-007
CONCEPTS	UNDERGROUND WASTE DISPOSAL - CONCEPTS AND MISCONCEPTIONS:	DWM-70-001
CRUSADE	CRUSADE ON BOTTLES:	DWM-70-010
DATA	DATA COLLECTION SYSTEMS FOR WATER QUALITY SURVEILLANCE:	DWM-70-004
DIGITAL	DIGITAL-COMPUTER APPLICATIONS IN CHEMICAL-QUALITY STUDIES OF SURFACE W	DWM-70-007
DISCRETE	DISCRETE SIMULATION LANGUAGES WITH REFERENCE TO A BIO- SIMULATION - A	DWM-70-013
DISCUSSION	FROM MANAGEMENT SCIENCES TO POLICY SCIENCES- MATERIALS FOR DISCUSSION:	DWM-70-012
DISPOSAL	UNDERGROUND WASTE DISPOSAL - CONCEPTS AND MISCONCEPTIONS:	DWM-70-001
ENVIRONMEN	MERCURY IN THE MARINE ENVIRONMENT:	DWM-70-003
ESTUARY	WATER-CLAY INTERACTIONS IN NORTH CAROLINA'S PAMLICO ESTUARY:	DWM-70-002
EUROPEAN	SAND RIBBONS OF EUROPEAN TIDAL SEAS:	DWM-70-008
EXCHANGE	VARIATION OF CATION EXCHANGE CAPACITY AND RATE WITH PARTICLE SIZE IN S	DWM-70-014
HYDROLOGY	MATHEMATICAL METHODOLOGY IN HYDROLOGY:	DWM-70-005
IMPLICATIO	SOME NORMATIVE IMPLICATIONS OF A SYSTEMS VIEW OF POLICYMAKING:	DWM-70-011
IMPRESSION	DISCRETE SIMULATION LANGUAGES WITH REFERENCE TO A BIO- SIMULATION - A	DWM-70-013
INTERACTIO	WATER-CLAY INTERACTIONS IN NORTH CAROLINA'S PAMLICO ESTUARY:	DWM-70-002
LANGUAGES	DISCRETE SIMULATION LANGUAGES WITH REFERENCE TO A BIO- SIMULATION - A	DWM-70-013
MANAGEMENT	FROM MANAGEMENT SCIENCES TO POLICY SCIENCES- MATERIALS FOR DISCUSSION:	DWM-70-012
MANGEMENT	THE USE OF WATER QUALITY PREDICTIONS AS AN AID TO MANGEMENT:	DWM-70-006
MARINE	MERCURY IN THE MARINE ENVIRONMENT:	DWM-70-003
MATERIALS	FROM MANAGEMENT SCIENCES TO POLICY SCIENCES- MATERIALS FOR DISCUSSION:	DWM-70-012
MATHEMATIC	MATHEMATICAL METHODOLOGY IN HYDROLOGY:	DWM-70-005
MERCURY	MERCURY IN THE MARINE ENVIRONMENT:	DWM-70-003
METHODOLOG	MATHEMATICAL METHODOLOGY IN HYDROLOGY:	DWM-70-005
MISCONCEPT	UNDERGROUND WASTE DISPOSAL - CONCEPTS AND MISCONCEPTIONS:	DWM-70-001
NORMATIVE	SOME NORMATIVE IMPLICATIONS OF A SYSTEMS VIEW OF POLICYMAKING:	DWM-70-011
NORTH	WATER-CLAY INTERACTIONS IN NORTH CAROLINA'S PAMLICO ESTUARY:	DWM-70-002
PAMLICO	WATER-CLAY INTERACTIONS IN NORTH CAROLINA'S PAMLICO ESTUARY:	DWM-70-002
PARTICLE	VARIATION OF CATION EXCHANGE CAPACITY AND RATE WITH PARTICLE SIZE IN S	DWM-70-014
POLICY	FROM MANAGEMENT SCIENCES TO POLICY SCIENCES- MATERIALS FOR DISCUSSION:	DWM-70-012
POLICYMAKI	SOME NORMATIVE IMPLICATIONS OF A SYSTEMS VIEW OF POLICYMAKING:	DWM-70-011
PREDICTION	THE USE OF WATER QUALITY PREDICTIONS AS AN AID TO MANGEMENT:	DWM-70-006
PRELIMINAR	PRELIMINARY STUDY OF TRANSVERSE BARS:	DWM-70-009
QUALITY	DATA COLLECTION SYSTEMS FOR WATER QUALITY SURVEILLANCE:	DWM-70-004
QUALITY	THE USE OF WATER QUALITY PREDICTIONS AS AN AID TO MANGEMENT:	DWM-70-006
QUALITY	DIGITAL-COMPUTER APPLICATIONS IN CHEMICAL-QUALITY STUDIES OF SURFACE W	DWM-70-007
RATE	VARIATION OF CATION EXCHANGE CAPACITY AND RATE WITH PARTICLE SIZE IN S	DWM-70-014
REFERENCE	DISCRETE SIMULATION LANGUAGES WITH REFERENCE TO A BIO- SIMULATION - A	DWM-70-013
RIBBONS	SAND RIBBONS OF EUROPEAN TIDAL SEAS:	DWM-70-008
SAND	SAND RIBBONS OF EUROPEAN TIDAL SEAS:	DWM-70-008
SCIENCES	FROM MANAGEMENT SCIENCES TO POLICY SCIENCES- MATERIALS FOR DISCUSSION:	DWM-70-012
SCIENCES	FROM MANAGEMENT SCIENCES TO POLICY SCIENCES- MATERIALS FOR DISCUSSION:	DWM-70-012
SEAS.	SAND RIBBONS OF EUROPEAN TIDAL SEAS:	DWM-70-008
SEDIMENT	VARIATION OF CATION EXCHANGE CAPACITY AND RATE WITH PARTICLE SIZE IN S	DWM-70-014

Figure 6. Example of the KWOC index.

CALIFORNIA	DWM-70-003
CATION EXCHANGE	DWM-70-014
COASTAL GEOMORPHOLOGY	DWM-70-009
COASTAL SEDIMENTATION	DWM-70-009
COMPUTERS	DWM-70-007
COMPUTERS	DWM-70-013
CONTINENTAL SHELF	DWM-70-008
DATA COLLECTION	DWM-70-004
DATA PROGRAMS	DWM-70-004
DELPWELL DISPUSAL	DWM-70-001
EDUCATION	DWM-70-005
ENGLISH CHANNEL	DWM-70-008
ENVIRONMENT	DWM-70-010
ENVIRONMENTAL MOVEMENT	DWM-70-010
ESTUARY	DWM-70-002
FINGER BARS	DWM-70-009
FLORIDA	DWM-70-009
FORECASTING	DWM-70-006
GROUNDWATER	DWM-70-001
HYDROLOGIC MODELS	DWM-70-005
ION EXCHANGE	DWM-70-002
MANAGEMENT	DWM-70-006
MANAGEMENT SCIENCE	DWM-70-012
MARINE ORGANISMS	DWM-70-003
MARINE SEDIMENTATION	DWM-70-008
MARYLAND	DWM-70-010
MATHEMATICAL MODELS	DWM-70-013
MATHEMATICAL MODELS	DWM-70-006
MATHEMATICAL MODELS	DWM-70-005
MATHEMATICAL MODELS	DWM-70-007
MERCURY	DWM-70-003
MODELS	DWM-70-005
NETWORK DESIGN	DWM-70-004
NORMATIVE GENERAL SYSTEMS	DWM-70-011
NORTH CAROLINA	DWM-70-002
POLICY ANALYSIS	DWM-70-011
POLICY ANALYSIS	DWM-70-012
POLICY SCIENCE	DWM-70-012
POLICY SCIENCE	DWM-70-011
POLICYMAKING	DWM-70-011
POLLUTION	DWM-70-003
POLLUTION	DWM-70-001
RECYCLING WASTE	DWM-70-010
REGRESSION ANALYSIS	DWM-70-007
RETURNABLE BOTTLES	DWM-70-010
SALINITY	DWM-70-002
SAMPLING THEORY	DWM-70-004
SAND DEPOSITS	DWM-70-008
SAND RIDGES	DWM-70-009
SAND WAVES	DWM-70-008
SEDIMENT	DWM-70-002

Figure 7. Example of the Index term index.

Although the bibliographic example just given will probably be the most frequent application of the keyword index program, other potential applications should not be overlooked. Recall that up to 9 types of "cards" (records) can be identified with one reference code. These cards in the bibliographic example were called:

Card type	
1	Author card
2	Title card
3	Citation card
4	Descriptor card
5	Identifier card
6-7	Additional card types

The information on all card types is printed in the bibliographic section of the index. Information on card type 1 (INDEX AUTHORS) and on card types 4 and 5 (INDEX TERMS) may be extracted in a back-of-the-book type of index. Information on card type 2 may be permuted in a keyword-in-context index (KWIC INDEX) or a keyword-out-of-context index (KWOC INDEX); in both cases information on card type 1 may be merged into the index (INDEX AUTHORS). Similarly, information on card types 4 and 5 may also be permuted in KWIC (KWIC TERMS) or KWOC (KWOC TERMS) indexes; again information on card type 1 may be merged into the index (INDEX AUTHORS). These commands are summarized in table 1.

Table 1. -- Summary of keyword program control cards.
X - required card; 0 - optional card.

Keyword program control cards								Sort program control cards				Description of output
PRINT BIBLIOGRAPHY	INDEX AUTHORS	KWIC INDEX	KWOC INDEX	INDEX TERMS c	KWIC TERMS	KWOC TERMS	COLUMN	Index terms sort	KWIC sort	KWOC sort	Skip sort step	
X	-	-	-	-	-	-	-	-	-	-	X	Bibliographic section of report only - all card types printed
0	X	-	-	-	-	-	X	X	-	-	-	Double column index of information on card type 1
0	0	X	-	-	-	-	-	-	X	-	-	KWIC index of card type 2
0	0	-	X	-	-	-	-	-	-	X	-	KWOC index of card type 2
0	0	-	-	X	-	-	X	X	-	-	-	Double column index of information on card types 4 and 5
0	0	-	-	-	X	-	-	-	X	-	-	KWIC index of card types 4 and 5
0	0	-	-	-	-	X	-	-	-	X	-	KWOC index - information on card types 4 and 5 appears in left-hand margin; information on card type 2 appears on line.

TITLES control card required for all runs and must appear last.
STOP WORDS control card and Stop words list, if used must immediately precede TITLES card.

Having reviewed the effects of the commands, now consider a research project file stored on GIPSY. Using the COPY command, an index might be created to the file making the following card type assignments:

Card type	
1	Project leader
2	Project title
3	Project location (city and state)
4	Descriptors
6	Project summary

The possible types of indexes were discussed above.

Other potential applications that come to mind are map catalogs, lists of geologic and geographic names, mineral and fossil inventories, certain types of personnel rosters, and indexes to publications (see the cumulative index to the U. S. Geological Survey's Sysnotes).

STORAGE REQUIREMENTS AND RUN TIME

The core storage requirements of the keyword program depend upon the line length of the index to be produced. The REGION parameter on the KWIC program-part 1 execute card should be specified as follows:

LENGTH n	REGION (or REGION.GO) parameter
LENGTH 80 (default)	REGION=114K
LENGTH 120	REGION=178K
LENGTH 132	REGION=198K

Run times depend upon the number of documents to be indexed and the length of the card type 2, card types 4, and cards type 5 records as specified by the STORAGE control card (STORAGE 180 is the default value). If STORAGE 180, a KWIC index to 120 bibliographic references requires the following run times:

	Program step	Run time (seconds)[1/]
Step 1	KWIC program-part 1	37
Step 2	Sort program - Balsort	3
Step 3	KWIC program - part 2	7

Run times for larger collection of documents or larger storage values must be adjusted upwards.

1/ Run times do not include times required to compile and link edit program source decks.

REFERENCES CITED

Addison, C. H., Coney, M. D., Jones, M. A., Shields, R. W., and Sweeney, J. W., 1969, General information processing system application description: Univ. Oklahoma Information Science Series, mono. no. 4, 126 p.

U. S. Geological Survey, 1970, U.S.G.S. Computer user's manual: Washington, D. C., U. S. Dept. of the Interior, Geological Survey, Computer Center Division.