Keyword Indexes for the Behavioral Sciences

by Kenneth Janda

Keyword-in-context, KWIC, indexing is the most widely-used computer-derived information retrieval method in use today. Its major drawback is the lack of descriptiveness of some of the journal titles that are KWIC-indexed, but it is nevertheless useful, flexible, and operational. Professor Janda, of the Department of Political Science, Northwestern University, describes a recent project in which he KWIC-indexed all volumes of The American Political Science Review, *describes several variations and improvements on KWIC indexing, and summarizes some of its planned and potential applications.*

TO many students in the behavioral sciences, information retrieval schemes involving the use of computers probably appear interesting and ambitious but still of dubious utility for present-day research problems. While it is true that much of the really exciting work being done by information retrieval specialists is still experimental, some of their techniques have long been operational and have already been applied to practical problems of scientific research. Computer-generated keyword indexes to current research literature, for example, have found standard usage within the physical sciences, at least as a partial solution to the problem of keeping abreast of publications in one's field. The success of this general method of using computers to prepare keyword indexes to bibliographical material within the physical sciences has been amply proven; the practical applications of keyword indexing within the behavioral sciences are no less promising.

Keyword indexing in any discipline operates on the assumption that certain terms or "keywords" constitute useful handles for pulling out or retrieving information of interest to the researcher. An ordinary book index serves as a good example of this assumption.

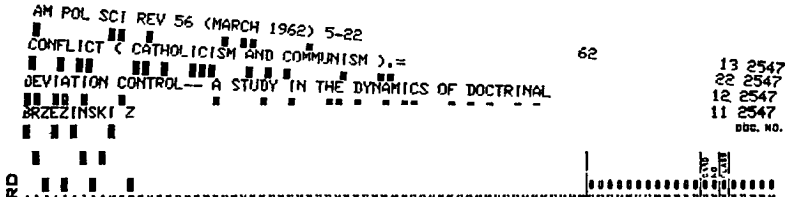
The preparation of any index involves many mechanical tasks: scanning pages for important terms, recording page numbers, cross-listing, and alphabetizing. Familiarity with the subject is undoubtedly an asset in preparing an index, but a good job can ordinarily be expected from any thorough, accurate, and efficient research assistant who is instructed in advance about the terms to be included in the index. Computers, which are notoriously thorough, accurate, and efficient, possess the requisite virtues for routinized keyword indexing, and need only be instructed how to do the job.

Computer programs providing these instructions have been available now for some time.[1] By means of the proper program, a computer can be instructed to read natural language material, search the material for the appearance of pre-defined keywords, alphabetize the keywords found, and print out the alphabetized keywords along with some of the "context" in which they occur. Two basic versions of these keyword indexing computer programs have become standard in the field of information retrieval. One version prints the keywords embedded in the original context; this method is commonly called "KWIC" indexing, for "*Key-Word-In-Context.*" The other prints the keywords alongside the original context and has become known as "KWOC" indexing, for "*Key-Word-Out-of-Context.*" Both KWIC and KWOC indexing can now be regarded as operational information retrieval techniques within both the physical and behavioral sciences.

The Nature of a KWIC Index

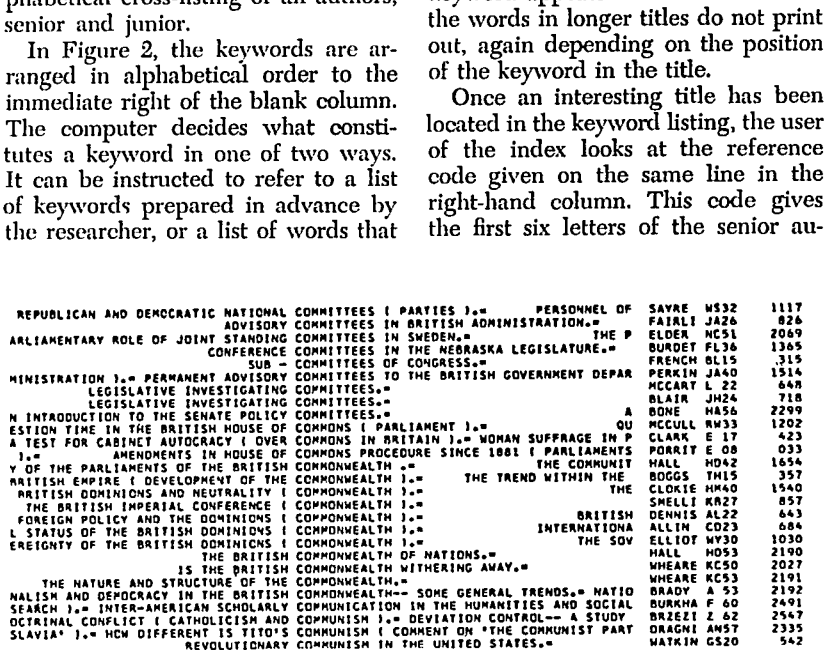
The current capabilities of KWIC indexing can be illustrated by reference to a computer-generated cumulative index to the *American Political Science Review,* which will be published in the fall of 1964 by Northwestern University Press on behalf of the American Political Science Association.[2] A total of 2,614 articles had appeared in the *Review* from its first issue in 1906 through its 57th volume in 1963. These articles were punched on cards as in Figure 1, with different "classes" of cards reserved for different types of information: author, title, and facts of publication.[3] Thousands of

Figure 1: IBM cards for the preparation of a KWIC index. At the bottom is the author card, second and third are title cards, the top is the source card.



these cards were fed into an IBM 709 computer, and the machine produced three types of output: an alphabetical listing of all keywords found in the titles of the articles along with part or all of the titles in which the keyword occurred (Figure 2), an alphabetical listing of complete citations by first-named author (Figure 3), and an alphabetical cross-listing of all authors, senior and junior.

In Figure 2, the keywords are arranged in alphabetical order to the immediate right of the blank column. The computer decides what constitutes a keyword in one of two ways. It can be instructed to refer to a list of keywords prepared in advance by the researcher, or a list of words that



REPUBLICAN AND DEMOCRATIC NATIONAL COMMITTEES (PARTIES).• PERSONNEL OF	SAVRE WS32	1117
ADVISORY COMMITTEES IN BRITISH ADMINISTRATION.•	FAIRLI JA26	826
ARLIAMENTARY ROLE OF JOINT STANDING COMMITTEES IN SWEDEN.• THE P	ELDER NC51	2069
CONFERENCE COMMITTEES IN THE NEBRASKA LEGISLATURE.•	BURDET FL36	1365
SUB - COMMITTEES OF CONGRESS.•	FRENCH BL15	315
MINISTRATION).• PERMANENT ADVISORY COMMITTEES TO THE BRITISH GOVERNMENT DEPAR	PERKIN JA40	1514
LEGISLATIVE INVESTIGATING COMMITTEES.•	MCCART L 22	648
LEGISLATIVE INVESTIGATING COMMITTEES.•	BLAIR JH24	718
N INTRODUCTION TO THE SENATE POLICY COMMITTEES.• A	BONE HA56	2299
ESTION TIME IN THE BRITISH HOUSE OF COMMONS (PARLIAMENT).• QU	MCCULL RW33	1202
A TEST FOR CABINET AUTOCRACY (OVER COMMONS IN BRITAIN).• WOMAN SUFFRAGE IN P	CLARK E 17	423
).• AMENDMENTS IN HOUSE OF COMMONS PROCEDURE SINCE 1881 (PARLIAMENTS	PORRIT E 08	033
Y OF THE PARLIAMENTS OF THE BRITISH COMMONWEALTH .• THE COMMUNIT	HALL HD42	1654
RITISH EMPIRE (DEVELOPMENT OF THE COMMONWEALTH).• THE TREND WITHIN THE	BOGGS TH15	357
BRITISH DOMINIONS AND NEUTRALITY (COMMONWEALTH).• THE	CLOKIE HM40	1540
THE BRITISH IMPERIAL CONFERENCE (COMMONWEALTH).•	SMELLI KR27	857
FOREIGN POLICY AND THE DOMINIONS (COMMONWEALTH).• BRITISH	DENNIS AL22	643
L STATUS OF THE BRITISH DOMINIONS (COMMONWEALTH).• INTERNATIONA	ALLIN CD23	684
EREIGNTY OF THE BRITISH DOMINIONS (COMMONWEALTH).• THE SOV	ELLIOT WY30	1030
THE BRITISH COMMONWEALTH OF NATIONS.•	HALL HD53	2190
IS THE BRITISH COMMONWEALTH WITHERING AWAY.•	WHEARE KC50	2027
THE NATURE AND STRUCTURE OF THE COMMONWEALTH.•	WHEARE KC53	2191
NALISM AND DEMOCRACY IN THE BRITISH COMMONWEALTH-- SOME GENERAL TRENDS.• NATIO	BRADY A 53	2192
SEARCH).• INTER-AMERICAN SCHOLARLY COMMUNICATION IN THE HUMANITIES AND SOCIAL	BURKHA F 60	2491
OCTRINAL CONFLICT (CATHOLICISM AND COMMUNISM).• DEVIATION CONTROL-- A STUDY	BRZEZI Z 62	2547
SLAVIA').• HOW DIFFERENT IS TITO'S COMMUNISM (COMMENT ON 'THE COMMUNIST PART	DRAGNI AN57	2335
REVOLUTIONARY COMMUNISM IN THE UNITED STATES.•	WATKIN GS20	542

Figure 2: A portion of a KWIC index of The American Political Science Review.

are *not* to be considered keywords. In the first case, the computer looks at every word contained in the title of the publication and compares it with its own stored list of keywords. The words in the titles found on the list are then selected for indexing. The process operates in a comparable way when a list of *non-keywords* is used: the computer includes the word in the index only when the word does *not* appear in the list. The latter procedure was used in the preparation of the cumulative index of the *Review*. Sample *non*-keywords (words not indexed) are "an," "of," "the," and "other."

To use the index in Figure 2, scan the vertical column of keywords for one that is of interest to you. Then read the context of the title printed on the same line as the keyword. There is room provided on the line for only 78 character-spaces, though this proved adequate for printing in full 86% of the 2,614 titles from the *Review*. A title with no more than 78 characters and spaces prints out in full, although a portion of the title may be "wrapped around" and printed before or after the keyword, depending on where the keyword appears in the title. Some of the words in longer titles do not print out, again depending on the position of the keyword in the title.

Once an interesting title has been located in the keyword listing, the user of the index looks at the reference code given on the same line in the right-hand column. This code gives the first six letters of the senior author's last name, his initials, the year of publication of the article, and the identification number of the article. The code enables the user to locate the complete citation in the author-alphabetized bibliography shown in Figure 3.

This type of index is also known as a "permuted" keyword index, for an article will appear as many times as the number of keywords it contains. The first title listed in Figure 2, for example, will be found in five other places in the index: under "REPUBLICAN," "DEMOCRATIC," "NATIONAL," "PARTIES," and "PERSONNEL." A total of 10,089 keyword lines were produced for the 2,614 articles from the *Review*. Therefore each title appears in the index on an average of 3.9 times.

The Nature of a KWOC Index

Conventional KWIC indexes are "double-entry" indexes; after the user has found a title of interest, he must check the reference code and then enter the author-alphabetized bibliography to get complete information on journal, volume, month, year, and pages. Although it would be possible to develop a KWIC index which would give this information in abbreviated form in the reference code, the purpose of a "single-entry" index is perhaps better served by the KWOC version or keyword indexing, which prints out the complete citation for each appearance of a keyword in the title.

A reproduction of a page of KWOC computer output is given in Figure 5. The punchcard format for the input to the KWOC program is identical to the KWIC format (Figure 1). The input to the KWOC example in Figure 4, however, consisted of 928 titles on "Africa" and "The Middle East" which were reported in the "Foreign and Comparative Government" bibliography published in the back pages of the *American Political Science Review* during the last four years.

Both the KWIC and KWOC programs identify the keywords by searching the title cards for the existence of previously defined keywords or non-keywords. But when the KWOC program finds a keyword, it is reprinted out of context to the left of the complete citation. KWIC, by printing only one line per index entry, is definitely more economical in its use of space and is possibly easier to use in retrieving titles of interest. KWOC, on the other hand, does provide a single-entry look-up and does allow for printing the complete title regardless of length.

Advantages of KWIC Indexing

The advantages of automatic indexing of bibliographical material by the keywords contained in their titles are those generally associated with the use of computers in most data processing operations. First of all, the indexes are easily and inexpensively prepared. The input to the KWIC index of the *Review* was prepared by a girl keypunching directly from the title pages

of the bound library volumes. It took less than 200 hours for her to punch and correct all the titles in 57 volumes of the *Review*. It took the 709 computer less than 12 minutes to process the 2,614 titles—searching a 418 non-keyword list for each word in every title, preparing 10,089 KWIC index lines, and producing a cross-reference listing for 2,801 senior and junior authors. Another 29 minutes were required to sort the output into alphabetical order, and about 30 minutes were needed to print the output on the IBM 1401.

In addition to the advantages of speed and economy, computer-generated keyword indexes are easily updated with new material and readily reproduced. Once a comprehensive bibliography is punched on cards, it becomes a simple matter to prepare specialized bibliographies by instructing the computer to index the literature only on a smaller number of previously identified keywords, such as "legislator," "legislature," "parliaments," "representation," etc.

For some people, a keyword index also has the advantage of convenience in use. Some users report that an alphabetized list of keywords provides a more efficient means of locating articles of interest than the conventional subject-heading index, where articles are alphabetized by authors within subjects. It must be noted, however, that others have found keyword indexes less convenient than conventional subject-heading indexes.[4] Perhaps it takes time to adjust to using the computer output.

The Descriptiveness of Titles

Obviously the *major* disadvantage in computer methods of keyword indexing lies in the "descriptiveness" of the titles fed into the computer. Keyword indexing was originally developed for application to "technical" literature, and indeed the longer and more descriptive titles of journal publications in the physical and biological sciences seem better suited to this technique than titles in the behavioral sciences. Lane has investigated the variation within different fields concerning the suitability of indexing titles solely by their keywords.[5] He pointed out that

```
1279  SALTER JT35  SALTER JT
                     GOVERNOR PINCHOT AND THE LATE MAGISTRATE STUBBS ( PATRONAGE
                     AND BOSSES IN PARTIES ).=
                     AM POL SCI REV 29 (APRIL 1935) 249-256
1511  SALTER JT40  SALTER JT
                     PERSONAL ATTENTION IN POLITICS
                     ( REPRESENTATIVE - CONSTITUENCY RELATIONS ).=
                     AM POL SCI REV 34 (FEBRUARY 1940) 54-66
1036  SANDEL W 31  SANDELIUS W
                     NATIONAL SOVEREIGNTY VERSUS THE RULE OF LAW.=
                     AM POL SCI REV 25 (FEBRUARY 1931) 1-20
2080  SANDEL WE51  SANDELIUS WE
                     REASON AND POLITICAL POWER.=
                     AM POL SCI REV 45 (SEPTEMBER 1951) 703-715
 489  SARKAR BK18  SARKAR BK
                     DEMOCRATIC IDEALS AND REPUBLICAN INSTITUTIONS IN INDIA.=
                     AM POL SCI REV 12 (NOVEMBER 1918) 581-606
 525  SARKAR BK19  SARKAR BK
                     HINDU THEORY OF INTERNATIONAL RELATIONS.=
                     AM POL SCI REV 13 (AUGUST 1919) 400-414
2581  SARTOR G 62  SARTORI G
                     CONSTITUTIONALISM-- A PRELIMINARY DISCUSSION.=
                     AM POL SCI REV 56 (DECEMBER 1962) 853-864
```

Figure 3: A KWIC index of complete citations by first-named author.

the reader interested in "duck shooting" probably would have missed an article entitled "Good Day in Bad Marsh" if he had consulted a keyword index but he would have found the article under the above subject-heading in the *Readers' Guide to Periodical Literature*.

The evidence indicates, however, that the professional journals in the behavioral sciences have a higher proportion of "descriptive" titles than periodicals reported in the *Readers' Guide*.[6] Furthermore, the problem of inadequate titles can be largely solved through editorial supervision during the preparation of the computer input. Scanning the texts of articles whose titles seem unclear or literary in nature will usually disclose some terms or phrases which might be enclosed in parentheses and added to the titles. Keywords added in this manner will be indexed as if they had been in the title. This procedure was followed in preparing the index to the *Review*, and the first line in Figure 2 shows how the word "PARTIES" was entered as a keyword addition. Keywords were added to 604 of the 2,614 *Review* titles, indicating that original titles were considered suitably "indexable" in about 77% of the cases.

Thus considerable improvement can be made in the quality of an index if some editorial supervision can be exercised in recording titles before processing them on the computer. This is relatively easy to do when the input is punched directly from journals as in the preparation of a cumulative index, but it is impossible to do when the input is punched from a bibliography. Although it is faster and usually more convenient to punch from bibliographies, the cumulative index approach will undoubtedly provide a greater pay-off in the long run in terms of more exhaustive coverage, more accurate reporting, more detailed publication information, and improved descriptiveness of titles. The cumulative indexes prepared for individual journals within a given discipline could then be merged to form a comprehensive master cumulative index.[7]

Future Applications of Keyword Indexing

As originally developed and applied within the physical sciences, keyword indexing has been limited to the retrieval of bibliographic items, but the

Figure 4: A sample of KWOC computer output.

```
2138  CONGO        AUTHOR NOT GIVEN
                     EQUATORIAL AFRICA - GABON, CENTRAL AFRICAN REPUBLIC, CONGO
                     REPUBLIC, CHAD.=
                     AFRICA SPECIAL REPORT MARCH, 1960
2243  CONGO        BERRIAN AH
                     THE FORMATION OF PROFESSIONAL AND TECHNICAL CADRES FOR THE
                     CONGO.=
                     JOURNAL OF HUMAN RELATIONS (AUTUMN 1963)
2389  CONGO        LEGUM C
                     THE BELGIAN CONGO (1)-- REVOLT OF THE ELITE.=
                     AFRICA SOUTH (OCTOBER-DECEMBER 1959)
2440  CONGO        AUTHOR NOT GIVEN
                     ATLANTIC REPORT-- THE CONGO.=
                     ATLANTIC MONTHLY, SEPTEMBER, 1963
2450  CONGO        HOSKYNS C
                     SOURCES FOR A STUDY OF THE CONGO SINCE INDEPENDENCE.=
                     JOURNAL OF MODERN AFRICAN STUDIES, SEPTEMBER 1963
2461  CONGO        RUDIN HR
                     AFTERMATH IN THE CONGO.=
                     CURRENT HISTORY, DECEMBER 1963
```

method has also shown definite, albeit limited, uses in genuine problems of *information* retrieval. The American Bar Foundation, for example, has applied KWIC indexing techniques to legislation recently enacted in the 50 states.[8] This index makes it possible for lawyers, legislators, and other interested persons to keep abreast of the enactment of state legislation by searching keyword listings obtained from brief descriptions of the statutes.

Another application of keyword indexing to information retrieval problems in political science is under way at Northwestern University. Systematic collections of roll call votes assembled for state legislatures, Congress, or the United Nations often grow so extensive that it becomes a task simply to find all votes recorded that deal with the substantive issues being researched, such as "labor," "foreign policy," and so forth. Keyword indexing provides a means for locating relevant votes out of hundreds of issues for which roll call votes were recorded. KWOC indexing of vote descriptions has been applied at Northwestern to 259 roll call votes taken in the plenary sessions and committee meetings of the 15th U.N. General Assembly. This technique is now being extended to processing the roll call vote descriptions reported in *Congressional Quarterly* and descriptions of votes taken in the Illinois state legislature.

These additional applications of keyword indexing are mentioned to illustrate the flexibility of the method; undoubtedly there are other possible applications to behavioral science research. Although keyword indexing is certainly not the ultimate technique in literature or information retrieval, it is *useful, practical, flexible,* and—perhaps most significantly—*operational.* The technique deserves to be investigated by those who sense a close fit between its capabilities and their needs.

References

[1] The earliest application of keyword indexing by computers to be published seems to have been the IBM bibliography, "Literature on Information Retrieval and Machine Translation," September 1958. H. P. Luhn discussed the indexing technique in "Keyword-in-Context Index for Technical Literature (KWIC Index)," IBM Advanced Systems Development Division Report RC-127, August 1959. *Chemical Titles,* which appeared in 1960, seems to have been the first published application of KWIC indexing outside of the computing industry. Various keyword indexing programs for computers are now in operation across the country. A thorough discussion of one such program for the popular IBM 1401 computing system is contained in "Keyword-in-Context (KWIC) Indexing Program for the IBM 1401 Data Processing System," Reference Code 1401-CR-02X, White Plains, N.Y.: International Business Machines, 1963. A recent revision of earlier KWIC programs for computers in the IBM 700 series has been prepared by Professor James S. Angaard of Northwestern University's Department of Electrical Engineering. This program, available through SHARE, is supplied as a series of relocatable subroutines and is designed for use in conjunction with IBM's Basic Monitor IBSYS. The indexes reported on in this paper were prepared with the use of Professor Angaard's EIKWIC revision.

[2] The cumulative index to the *Review* was processed on an IBM 709 computer with the EIKWIC program in use at Northwestern University (see Footnote 1). Dr. Evron M. Kirkpatrick, Executive Director of the American Political Science Association, arranged for the Association's financial support of the indexing project. Miss Louise Cowen, Director of the Northwestern University Computing Center, extended many kindnesses during both the sponsored and unsponsored phases of the project.

[3] There is no practical limit to the number of authors, length of titles, or amount of publication information that can be recorded on punchcards for input to the computer. Additional cards within a given "author," "title," or "source" classification can be used as needed. The standard format for recording bibliographic information on punchcards for keyword indexing is given in the IBM General Information Manual E20-8091, "Keyword-In-Context (KWIC) Indexing."

[4] See C. W. Hanson's review of a KWIC publication in the *Journal of Documentation,* 19 (September 1963), 137-38.

[5] B. B. Lane, "Key Words in—and out of—Context," *Special Libraries,* (January 1964), 45-46.

[6] Donald H. Kraft, "A Comparison of Keyword-In-Context (KWIC) Indexing of Titles with a Subject Heading Classification System," *American Documentation,* 15 (January 1964), 48-52. Kraft's analysis of 3,248 entries in an issue *of The Index to Legal Periodicals* and an issue of *The Index to Legal Theses and Research Projects* disclosed that 89.5% had titles judged suitable for keyword indexing.

[7] A computer-generated cumulative index has been prepared by Murray A. Straus and Susanne C. Graham for *Marriage and Family Living,* volumes 1 to 24 (1939-1962). *The American Journal of Sociology* will also publish a KWIC index in early 1965 which will include all the articles published in the *Journal* during its 70 volume history. A comprehensive KWIC index to publications in the field of social welfare from 1924 to 1962 has been prepared by Joe R. Hoffer, Executive Secretary of the National Conference on Social Welfare. The social welfare index has both cumulative and bibliographical characteristics, as does *The Index to Legal Theses and Research Projects,* first prepared with the use of a computer in 1962.

[8] *Current State Legislation Index,* American Bar Foundation, Indianapolis: Bobbs-Merrill Company, beginning publication in 1962.