KENNETH JANDA

Political research with MIRACODE : A 16 mm microfilm information retrieval system*

« Information retrieval » is a term which is inevitably linked with digital computer technology. While the tremendous speed and manipulative power of computers provide good reason for their dominance in information retrieval methodology, computing tehchnology at present does not offer the best solution to all information retrieval problems. So pervasive is the computer's influence in information retrieval, however, that alternative techniques are often overlooked by scientists and scholars. Notwithstanding the great capabilities and many appropriate applications of computers in information processing, researchers owe it to themselves to keep a watchful eye and open mind for alternative solutions to their information problems. One such alternative solution that deserves to be considered for retrieving information from large files of original material is Eastman Kodak's MIRACODE system for storage and retrieval with 16 mm microfilm.

MIRACODE is an acronym for « Microfilm Information Retrieval Access CODE » and is the name applied to the system and equipment developed by the Eastman Kodak Company. The MIRACODE system was introduced in 1963 and demonstrated at several conventions of information processing specialists, including the 1964 Annual Meeting of the American Documentation Institute. But perhaps because of the computer's widespread influence in information processing, the existence and capabilities

* Paper presented at the Third Technical Conference of the Council of Social Science Data Archives, Ann Arbor, Mich., May 10-12, 1966.

of the MIRACODE system have apparently escaped the notice of most information retrieval methodologists [1].

The MIRACODE system deserves attention especially for its advantages over certain drawbacks in contemporary computer technology. Its use of microfilm storage enables the system to handle large amounts of textual material without abstracting or endless keypunching. It allows direct man-machine interaction with browsing capabilities that are impossible with computer systems operating in a batch-processing mode. Furthermore, its relatively low purchase price (less than $25,000 for the basic system) permits its consideration as a system for tasks which could not justify the expense of renting and operating a suitable computer. Finally, it has some of the same powerful searching capabilities of a computer, employing Boolean logic on machine-readable optical codes.

The basic components of the MIRACODE system are a special microfilm camera and microfilm reader. The system can store and retrieve individual pages of original documents according to one or more three-digit code numbers assigned to the input material. At the microfilming stage, the MIRACODE camera transforms the code numbers into machine-readable binary codes, which are recorded on film next to the appropriate page image. The film, which is loaded into magazines for convenient handling, is searched for logical combinations of code numbers at the rate of ten feet per second at the MIRACODE retrieval station. Upon retrieval, the page image is projected on a ten by twelve inch viewing screen. Black-on-white photographic prints of projected pages can be produced by pressing a button on the microfilm reader. Depending on the amount of coding per image, several hundred pages of material can be stored on one 100 foot film magazine and searched for specified combinations of code numbers in ten seconds.

Instead of describing the technical features of the MIRACODE equipment, this article will try to illustrate its operation with reference to its application in a comparative study of political parties, now in progress.

THE COMPARATIVE PARTIES PROJECT

The comparative parties project at Northwestern University is a comprehensive, empirically-based, comparative study of political parties throughout the world. Defining a political party as any political organization whose electoral candidates won at least 5% of the membership of the lower house in a national legislature in two successive elections from 1950 through 1962, the project includes some 250 parties in 90 countries.

1. An examination of all issues of *American Documentation* since 1963 discloses no articles discussing the MIRACODE system of information retrieval. Moreover, our discussions and correspondence with information retrieval specialists have disclosed little or no knowledge of the system.

Data for analysis in the project will not be collected through costly field research in each of these countries but from the rich literature on political parties that has developed since 1950. The project proposes to mine this information through the use of various modern information retrieval techniques, to analyze the data brought to the surface, and to make the retrieved information available for research by other scholars.

At least five major information handling problems confront this proposal to conduct systematic and comprehensive research on the world's political parties by mining the existing literature. These are :

1) Developing an effective method for retrieving information from the parties literature;
2) Locating literature which contains relevant information on parties included in the study;
3) Building an inventory of propositions and theories about political parties and party systems;
4) Operationalizing variables in the propositions with reference to information from the literature; and
5) Analyzing data for hundreds of parties coded according to variables included in the study.

To some extent, similar problems are present in virtually every research project. The scope of the parties project, however, magnifies the tasks far beyond an effort conceivable with traditional research methods. The demands of this project require the utilization of modern information retrieval and information processing technology. A variety of specific techniques have been proposed as solutions to the information handling problems listed above [2]. Most of the techniques involve computer applications for keyword indexing [3], search and retrieval by logical relations among keywords [4], and data processing [5]. The use of computer technology was rejected, however, as a solution for the central problem in the project — that of developing an effective method of retrieving information from the parties literature.

2. K. JANDA, « Retrieving information for a comparative study of political parties », in : W.F. CROTTY, ed., *Approaches to the study of party organization*, Boston, Allyn and Bacon [forthcoming in 1967].

3. K. JANDA, « Keyword indexing for the behavioral sciences », *American behavioral Scientist* 7, June 1964 : 55-58.

4. K. JANDA and W. TETZLAFF, « TRIAL : A computer technique for retrieving information from abstracts of literature », *Behavioral Science* 11, November 1966 : 484-496.

5. K. JANDA, *Data processing : applications to political research*, Evanston, Ill., Northwestern University Press, 1965. Especially chapter VI.

RETRIEVING INFORMATION FROM PARTIES LITERATURE

In the early stages of the comparative parties project, considerable attention was given to the development and application of computer techniques to retrieve discussions from the literature about party functions, membership, goals, leaders, activities, and other relevant information. Computer programs were written to search natural language text and retrieve such discussions on command. The fundamental drawback in using computers for this purpose in the parties project was the tremendous amount of keypunching required to put the vast literature on political parties into machine-readable form. Keypunching costs could go down, of course, if one chose to punch only abstracts of literature rather than entire texts, but this would decrease the amount of information entering the system and increase costs of preparing input for keypunching. At least until optical scanners of printed texts become both reliable and economical, computer techniques of information retrieval seem poorly suited for handling the thousands of books and articles that will eventually form the input to the parties project.

The MIRACODE system offers a far more efficient method for harnessing this vast literature. Material is prepared for the system by coding the topics discussed on each page with reference to a set of coding categories, similar to the practice followed in the Human Relations Area Files [6]. These code numbers are then recorded on slips of paper, which are placed in the page margins before microfilming. A typical page coded and prepared for photographing is illustrated in Figure 1, which shows a discussion of Japanese political parties « tagged » with numbers from the coding scheme (described below). Coding the information and microfilming the coded page saves considerable time and expense over keypunching the original text or abstracting the information and then keypunching. With microfilm, the entire text is recorded instantaneously, with perfect accuracy, and stored in a fraction of the space required by punchcards or even magnetic tape (See Figure 1).

The advantages of microfilm for recording and storing large files of material have long been recognized, but no method has heretofore been provided for effective retrieval of information once recorded on film. To be sure, an index can be prepared to show the content and location of information on film, and the appropriate reel can be selected and run through a microfilm reader to zero in on the desired frame by visually checking sequence numbers on the film. But this method of retrieval is too crude and cumbersome for many research purposes, which require an

6. G.P. MURDOCK and others, *Outline of cultural materials*, New Haven, Conn., Human Relations Area Files, 1961.

FIGURE 1. — *Sample page tagged with code numbers and ready for microfilming.*

Yanaga, Chitoshi, Japanese People and Politics, 069
New York: Wiley, 1956

278 Japanese People and Politics

of every 4. This was 2.6 times the next largest group, Waseda University, which was represented by 47 members or 1 out of 10. Even in the Socialist parties the Tokyo University group was the largest, with Waseda University coming second.[44]

There is a striking social disparity between the members of the Diet and the rank and file members of the party outside the parliament. This is true of all the parties but is more clearly demonstrated in the conservative parties as can be seen by the educational background of the members who come from the upper and upper middle classes. Within the parliamentary parties themselves, however, there is remarkable educational-level homogeneity.

As compared with the British Labor Party members of Parliament in 1950, of whom about 4 out of 11 or better than one-third had some kind of university education, the overwhelming majority of Socialist Party members of the Diet, to the extent of 80 to 84 percent, had some kind of college or university education. This gives quite an intellectual flavor to the leadership in their activites.

Occupation

Occupational breakdown presents a difficult problem since accuracy in classification categories becomes almost impossible. However, an analysis can provide a useful basis for understanding the bias of the Diet. Table V represents the occupational distribution of the members of the House of Representatives who were elected in the General Election of April 19, 1953.

Several generalizations can be made from the figures given above. "Big business" has the biggest representation, taking up well over one-third of the entire House of Representatives membership on their side. This compares with the conservative parties in which 3 in every

[44] The preponderance of Tokyo University graduates was maintained in the Diet as the result of the House of Representatives election of April 19, 1953, though there was a slight decrease in the total number. The educational background of the newly elected members was as follows:

Tokyo University	113
Waseda University	50
Nippon University	34
Kyoto University	26
Chuo University	22
Other universities and colleges	141
Secondary education only	70
No mention	10
Total	466

ability to retrieve material swiftly and automatically according to logical connections among subjects being discussed, e. g., retrieving only discussions of subjects A *and* B that do *not* mention C.

Previously, such logical searching capability was available only through digital computer technology. The MIRACODE system, however, incorporates electronic circuitry to detect logical combinations among machine-readable code numbers associated with the input material. These code numbers are rendered machine-readable during microfilming by manipulating sliding switches at the MIRACODE camera to translate the numbers into a binary code of transparent and opaque rectangles recorded on the film next to the page image. As an optional feature, the codes can be recorded on film from punchcards read by an IBM keypunch machine modified for this purpose. The page image and the codes are recorded on film in accordance with the illustration in Figure 2.

The binary codes on the film are sensed by an optical scanning device, which reads the codes flashing by the scanning head at the normal film transport speed of ten feet per second. The retrieval station has the capability of testing for logical relationships among as many as fifteen different three-digit codes as the film passes the optical scanner. A code is involved in a search by pressing down the appropriate keys on a bank of buttons at the MIRACODE keyboard. The keyboard is modular in design, allowing from one to a maximum of fifteen banks of keys to operate a retrieval station. One keyboard configuration involves six banks of keys, which permit testing for logical relationships among six three-digit codes. At present, the available logic for MIRACODE searches consists of « and », « not », « or », « greater than », « less than », and « equal to ».

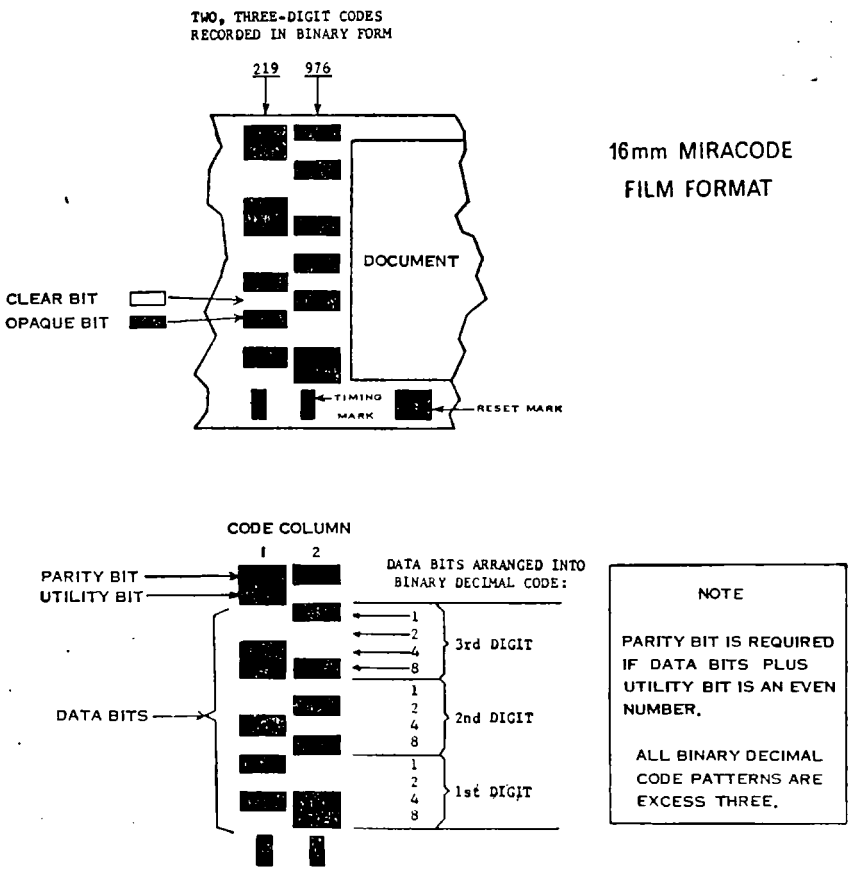
A search command is communicated to the reader by pressing the SEARCH button, which starts the film transport. When the machine senses the appropriate relationship among the numbers entered on the keyboard, the film immediately comes to a halt and backs up several frames to display the image retrieved by the search command. The operator has the opportunity to examine the page being projected for its relevance to his request. Should a hard copy be desired, a black-on-white photographic print can be made from the projected image in twenty-five seconds.

If the retrieved image does not satisfy the user, the search can be continued by pressing the search button again. The film will advance and stop to project the next image on the film that satisfies the search command. Rewinding occurs automatically when the end of the film is reached. If desired, the retrieval station can be set to operate automatically, printing each page on the film that satisfies a given search command.

While seated at the MIRACODE reader-retrieval station, the user can interact with his data files by changing his search command to alter the character and amount of information retrieved. To increase the number of « hits », the user can relax the search command by turning off a small toggle switch associated with each bank of keys, thus removing code numbers from the search. To decrease the number of « hits » and make his search more

selective, the user can enter additional code numbers on remaining keys, depending on the number of keyboard banks available. The user can determine in advance of retrieval how many « hits » he will get for any given command through the operation of an optional device, a « response monitor», which reads the film and tallies the number of satisfied conditions without actually stopping to display the retrieved images. This tally is instantaneously displayed on an electronic « scoreboard » as the film is read.

FIGURE 2.



BINARY DECIMAL CODE FOR 1st COLUMN: 219

3rd DIGIT 8 + 4 — 3 = 9
2nd DIGIT 4 — 3 = 1
1st DIGIT 4 + 1 — 3 = 2
} = 219

BINARY DECIMAL CODE FOR 2nd COLUMN: 976

3rd DIGIT 8 + 1 — 3 = 6
2nd DIGIT + 2 — 3 = 7
1st DIGIT 8 + 4 — 3 = 9
} = 976

CONSTRUCTING CODING CATEGORIES

Obviously, it is crucial that the input information be coded properly for effective retrieval with the MIRACODE system. Constructing useful coding categories and training people in their proper application constitute difficult but not insurmountable problems. A sizable literature has developed around coding procedures and the evaluation of coding reliabilities [7], and these topics will not be discussed here. Suffice it to say that serviceable code categories and coding instructions have been developed for a wide variety of information, limited only by the imagination of the researcher. The use of codes in conjunction with the MIRACODE equipment will be illustrated with reference to coding categories developed for the parties project.

Two different sets of numbers are used in coding the political parties literature. One set, consisting of three digit numbers from 000 through 999, is used exclusively as *identification codes* for specific parties. The other set, consisting of two-digit codes from 00- to 99-, is used to index *substantive information* about parties. The two sets of codes can be differentiated in the MIRACODE system by means of a « utility bit » recorded on the film with every column of code. The utility bit position can assume a value of 0 or 1, depending on how the utility bit switch is set at the time the codes are recorded on film. Party identification codes are recorded with the switch set at 1; information codes are recorded with the utility bit at 0. The MIRACODE retrieval station can decipher the utility bit code during the searching process so that a given number can be interpreted properly as an identification or information code.

Identification Codes : The party identification codes are organized on the basis of ten broad cultural-geographical categories. The first digit of the three-digit code stands for each main division as follows :

Code Cultural-Geographical Division

0 — Anglo-American political culture
1 — West Central and Southern Europe
2 — Scandinavia
3 — South America
4 — Central America and the Caribbean
5 — Asia and the Far East
6 — Eastern Europe
7 — Middle East and North Africa
8 — West Africa
9 — Central and East Africa

7. For the most recent evaluation of these activities, see O. R. HOLSTI, « Content Analysis », in G. LINDZEY and E. ARONSON, eds., *The handbook of social psychology*, 2nd ed., Cambridge, Mass., Addison-Wesley [forthcoming].

The second digit of the three-digit code stands for a particular country within each division. This scheme permits recording up to ten countries within each division, thus accommodating a maximum of 100 countries. Although there are 115 countries in the United Nations alone, the coding scheme is adequate for the ninety countries in the parties project. The third digit stands for a particular party within each country, providing for a maximum of ten parties within each country and 1000 parties overall. These ranges are quite adequate for the parties project, which includes only about 250 parties and not more than seven in any single country. Sample identification codes for Japanese political parties in the project are as follows :

541 Progressive (Kaishinto)
542 Left-Wing Socialist (Saha Shakaito)
543 Right-Wing Socialist (Uha Shakaito)
544 Liberal Democratic (Jiyu Minshuto)
545 Socialist (Shakaito, Social Democratic before 1955)

Party identification codes are used to tag locations in texts where information about specific parties is presented. The *substantive* nature of the information is recorded by means of information codes.

Information Codes : The MIRACODE system has the capability of dealing with three-digit codes, and the party identification codes are in fact three-digit numbers. The initial set of information codes constructed for the project were also three-digit numbers. Our experience in applying three-digit codes to selected articles on political parties, however, revealed that these codes were too detailed. Coding the material with 1000 coding categories required far more time than anticipated. Moreover, agreement among coders often extended to the first two digits of the code, but not to the third.

Upon re-examination of the nature of the codes and the objectives of the project, the decision was made to discontinue making the fine distinctions that the third digit required and to code only at the two-digit level of classification. This scheme provides 100 coding categories for information on political parties and at the same time leaves room for expansion of the code (by activating the third digit) to accommodate up to 1000 categories, should finer distinctions prove necessary. Because of technical considerations in the MIRACODE system, the two digit codes are recorded with « = » as a dummy third digit. Should any two digit code be expanded to a third digit, the use of the equal sign for that position in the old codes will permit retrieval of either old or new material with the same keyboard settings.

The information codes have been organized in an attempt to answer several basic questions about political parties. Each of these questions encompasses up to ten coding categories. The first digit of the information codes stands for a given question.

Code Questions about Political Parties

0 — What is a political party? — Definition, function, theory
1 — How do political parties begin? — The origin of parties
2 — What does a party do? — Party activities
3 — Who belongs to the party? — Party composition
4 — How is the party organized? — Party structure
5 — What does the party seek to accomplish? — Party goals
6 — Under what conditions does the party operate? — Political environment
7 — Under what conditions does the party operate? — Social, economic and geographical environment
8 — Are there any other parties in the country? — Party system
9 — How have parties been studied? — Methodology

Each of the code divisions have been subdivided into a maximum of ten concept categories. The complete set of codes as they stand in the present stage of the parties project is given in Table 1.

These codes have been applied on a pretest basis to 1445 pages on Japanese political parties and 1015 pages on party politics in Yugoslavia. The pretest phase of the project is now involved in coding literature on political parties in Argentina. As expected in a pretest experience, the codes continually undergo revision and refinement. Reliabilities among coders are systematically evaluated by having different personnel recode a ten percent random sample of literature previously coded. Intercoder reliabilities are assessed with this formula : Reliability $= 2M/(N_1 + N_2)$, where M equals the number of matching codes on a page, N_1 equals the total number of codes entered on that page by the first coder, N_2 equals the total number of codes used by the second coder. Our pretest experience with partially developed coding categories has revealed coding reliabilities of .50 and .47 for substantive information on Japan and Yugoslavia respectively. We are now working to improve these reliabilities before the pretest phase ends.

It should be noted that these initial coding reliabilities, which appear small by comparison to reliabilities of .8 and .9 often achieved in coding open-ended survey responses, are generated for the most exacting of coding tasks. In the first place, the number of substantive coding categories, at 100, is much larger than the ten codes or so used for survey responses. Secondly, each page of material emits many stimuli to the coder, who may or may not respond to a stimulus by entering a code number. For sample survey data, however, each stimulus normally demands a code and *only* one code, which removes the decisions as to whether or not codes should be applied and how many should be used. For this reason, it is submitted that the familiar percentage of agreement formula for calculating coding

reliabilities is appropriate for the parties material, notwithstanding the legitimate criticisms raised against its use in different contexts [8].

Our experience shows that coding can be done at the rate of about twenty pages per hour. Assuming an average of 3000 pages to be codep for parties literature in each country, 150 hours are required per country, which amounts to $ 375 at $ 2.50 per hour for graduate student coders. The cost of film, magazine, and processing is minimal — about $6.00 for 100 feet. A 100 foot reel of film will accommodate 2400 one-half inch « frames » generated by the MIRACODE camera. This does not mean that 2400 pages of material can be put on one reel, because the codes also take up film space. One column of code on the film is required for each three-digit code number, and a column of code requires approximately 1/3 of a frame. Therefore, material coded to a depth of three codes per page will require as much film for the codes as for the material being photographed.

We have estimated that the hourly wages involved in all aspects of processing information for retrieval with the MIRACODE equipment will be approximately $ 855 per country. This cost might be judged in comparison with the expense of a national sample survey, which can easily run $ 30 an interview for a sample size of 1000. Clearly, the types of information generated by the two methodologies — survey research and content analysis — are quite different, which may vitiate the comparison. The point is, however, that systematic exploitation of library-type material with the MIRACODE system is both technically and economically feasible. The system certainly deserves consideration for other social science research applications.

TABLE 1. *Outline of substantive information codes for the Parties Project.*

0— WHAT IS A POLITICAL PARTY—DEFINITION, FUNCTIONS, THEORY
00– DEFINITION OF A POLITICAL PARTY
01– TYPOLOGY OF PARTIES
02– PURPOSE OF STUDYING PARTIES
03– THEORY ABOUT PARTIES
04– FUNCTIONS OF PARTIES
05–
06–
07–
08–
09–
1— HOW DOES A POLITICAL PARTY BEGIN
10– WHEN WAS IT FORMED
11– WHO FORMED IT AND WHAT WAS ITS BASE OF ELECTORAL SUPPORT
12– WHY WAS IT FORMED
13– HOW WAS IT FORMED
14– POLITICAL HISTORY OF PARTY
15– ORGANIZATIONAL HISTORY OF PARTY
16–
17–
18–
19–

8. W.A. SCOTT, « Reliability of content analysis : the case of nominal scale coding », *Public Opinion Quarterly* 3 (19), 1955 : 321-325.