DOCUMENT RESUME

ED 061 993 LI 003 664

TITLE	Proceedings, Annual Meeting, Western Canada Chapter, American Society for Information Science (Third, Banff School of Fine Arts, October 3,4,5, 1971).
INSTITUTION	American Society for Information Science. Western Canada Chapter.; Calgary Univ. (Alberta).
PUB DATE	Oct 71
NOTE	142p.; (84 References)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Annual Reports; Automation; *Computer Science; *Information Processing; Information Retrieval; *Information Science; *Information Scientists; Information Services; Information Systems; Meetings
IDENTIFIERS	American Society for Information Science; *Canada

ABSTRACT

The proceedings contain papers given by the members of the chapter who come from both the University and Business environments. Some operational indexing, bibliographic, SDI and Retrospective Search Systems which include CAN/SDI, Compendex, TEXT-PAC, SIS II & III, KWOC and FAMULUS are discussed. Also included are papers on two projects conducted by the Computing Science department of the University of Alberta; the one project is an on-line thesaurus and the second an Information Retrieval Laboratory. Other papers are about the computerized circulation system at the University of Calgary's library, the Marc project at the University of Saskatchewan and the problems of design and coding questionnaires. (Author)

PROCEEDINGS

Third Annual Meeting

Western Canada Chapter

AMERICAN SOCIETY FOR INFORMATION SCIENCE

October 3, 4, 5, 1971

Banff School of Fine Arts

Information Systems
The University of Calgary
Calgary, Alberta

October, 1971

PROGRAM

Opening
A Few Words from Your President

by

F.T. DOLAN

OPENING
A FEW WORDS FROM YOUR PRESIDENT

Dear ASIS Members and Visitors:

It is my pleasure, as President of the American Society for Information Science - Western Canada Chapter, to welcome you heartily to our Third Annual Meeting.

Our cordial welcome is due to our ASIS distinguished lecturer, Dr. Leimkuhler, and to Dr. Norrie, Head of the Division of Information Services, The University of Calgary, who will be giving the key-note address a few minutes later.

Our annual conferences have become major milestones, not only in our Chapter's evolution, but in our own professional development as individuals.

At our first meeting in Edmonton in 1969, we laid the foundations of our Chapter. At the Vancouver meeting in 1970, we already had quite a few remarkable achievements behind us. It is rewarding to see, at this point, schools of Information Science being formed, new information systems working well, but most of all, interest growing among users as indicated by their active participation in the information process: which we will never cease to point out - is a two-way process.

Established only two years ago, our Chapter is still relatively young, however it is indeed gratifying to know that our organization has proven not to have been merely the whim of a few people. Our Chapter is a consolidated organizational structure, which not only has a head, but more important, it has a living body. This was substantiated not only by the increasing number of papers submitted, but also by their dynamic content.

I will not go into the particulars which may be seen in the agenda or other materials, but I would like to say a few words concerning the general concept of this meeting.

It is the rather unusual set-up of papers and the pertaining discussion which may have drawn your attention. We believe that the discussions confined into smaller rooms and to only one paper at a time for a devoted audience will create an informal atmosphere which is more likely to attract those who are really interested in that particular subject matter and save some time to those who are not and would be wasting their time. This should also promote animated discussion.

Our Third Annual Meeting is taking place in the dignified and enjoyable environments of Banff which, we hope, will create a serene

atmosphere for our work. In this respect, our thanks are due to the School of Fine Arts in Banff which is hosting our conference, as well as to the Department of Continuing Education at The University of Calgary, for looking after the organizational details.

Most of all, I would like to thank all of you who have come here, for your continuing interest and support, without which this event never would have materialized. May I also encourage those of you who are not yet members of our Chapter, to join us in our effort to do something worthwhile for Information Science in Canada.

F. T. Dolan

Information Science -- Its Social Responsibility

by

Dr. D.H. NORRIE

INFORMATION SCIENCE - ITS SOCIAL RESPONSIBILITY

Dr. D. H. Norrie
The University of Calgary
Calgary, Alberta

ABSTRACT

Information is meaning encoded in form. the meaning will be incomplete, biased, or distorted, if the coded message or the form is altered in any of these ways during the collection, transmission, processing, storage, or retrieval stages. A society can only exist if there is an adequate information flow between its component parts. A democratic society can only truly develop if this information flow is protected from manipulation and bias. Those in the profession of information science have a special responsibility in protecting the information flow from distortion and alteration.

Some Aspects of Canada's Highly Qualified Manpower Resources with Special Reference to the Can/SDI Project

by

G.R. MAUERHOFF

SOME ASPECTS OF CANADA'S HIGHLY QUALIFIED MANPOWER RESOURCES WITH SPECIAL REFERENCE TO THE CAN/SDI PROJECT

Georg R. Mauerhoff
National Science Library of Canada
Ottawa, Ontario

ABSTRACT

In 1972, Canada will have an estimated 174,500 highly qualified researchers and scientists employed in the various areas of engineering, physical and life sciences. Since this manpower is dependent on information, a study was undertaken to ascertain where this population was located, how they were employed, and how many could be in fact categorized as potential users of bibliographic information systems such as CAN/SDI. It was concluded that 16% or 28,000 of all the STI manpower can utilize SDI services, with 2,400 presently doing so.

INTRODUCTION

This investigation was prompted by the need for an estimate on Canada's scientists and technologists as potential users of information systems. Details of the CAN/SDI Project, the largest current awareness information system in Canada have been made available by Brown (1969), Mauerhoff (1970), Wolters (1971) and Gaffney (1971). The University of Calgary's Compendex system has been described by Dolan (1970), but it was not attempted to describe in Canada or, for that matter in the United States, the actual manpower that could utilize scientific and technical bibliographic information.

Lipetz (1970), who analyzed and critically appraised the myriad activity in the field of information needs and uses for the latest Annual Review of Information Science and Technology neglected numbers, was concerned with the processing of information, the methology used in studying users and their needs, and the theoretical aspects of information utilization.

This paper attempts to view the size of Canada's STI (Scientific and Technological Information) community and such characteristics as geographical distribution, sectors of employment, and fields or profession. The CAN/SDI Project will be referred to throughout this paper.

HIGHLY QUALIFIED MANPOWER

In this report, highly qualified manpower (HQM) comprises two groups. First, it is regarded as that portion of the total labour force possessing a university degree or its equivalent. This group consisted of 280,000 graduates in 1961, and made up 4.4% of the entire labour force.

The second group comprises the "professional and technical occupational group" as given in the 1961 Census of Canada, and included 10% of the Canadian labour force. Numbering 629,000 people in 1961, this class embraces the fields of architecture, engineering, physical sciences, life and health sciences, social sciences, law and education.

Since approximately one-third of the professional and technical occupational group was also university graduates, as the 1961 census figures indicate, the HQM resources for 1961 were actually 280,000 graduates and 419,000 professional and technical persons, or a total of 699,000 people. Table 1 shows these figures, and also attempts to project the labour force, professional and technical group, university degree holders and HQM for the years 1971 and 1976. The projected values were calculated by the author based on the growth curves of the past thirty years.

Canada's total resources of HQM are, however, too broad a population to be investigated in this study because of the National Science Library's main responsibility to researchers and scientists in the fields of engineering, the physical and life sciences. This reduced manpower population consists of the various areas of engineering, such as aeronautical, electrical, mechanical, etc.; the physical sciences, such as chemistry, physics, mathematics, etc.; and the fields of life sciences, such as agriculture, biology, forestry, psychology, etc. Consisting of about 113,000 in 1966, this subset of the HQM has been estimated by the author to number 174,500 in 1972, and over 300,000 in 1978. Architecture, social sciences, health, law and education have been excluded. According to A.G. Atkinson, K.J. Barnes, and Ellen Richardson (1970) of the Department of Manpower and Immigration, "the true population of scientific and engineering manpower in Canada is unknown". In order to arrive at some kind of a picture of Canada's resources of scientists and engineers, the author has decided to accept the results of the 1967 sample survey of scientists and engineers by Atkinson. Ratios and distributions for the 1967 survey population of 61,300 are assumed to stay constant for 1972, when the population is estimated to number 174,500 scientists and engineers. The information in 1967 was obtained by means of a questionnaire mailed to members of a large number of governmental, educational organizations and professional associations in Canada.

Table 2 illustrates the distribution of the 1967 survey population, and also indicates how the 1972 population is expected to be employed. For example, in the 1967 survey, 55% of the 61,300 scientists were engineers, 55% of the 174,500 scientists and engineers (S&E) in 1972 are also engineers.

FIELD OF EMPLOYMENT

Of the scientists and engineers to be employed in Canada in 1972, some 95,000 will be in the field of engineering; 26,000, or 15% are to

TABLE 1

CANADA'S POPULATION, LABOR FORCE, PROFESSIONAL AND TECHNICAL OCCUPATIONAL GROUP, UNIVERSITY GRADUATES AND HIGHLY QUALIFIED MANPOWER

	TOTAL POPULATION	LABOR FORCE	PROFESSIONAL & TECH. CLASS	UNIVERSITY GRADUATES	HIGHLY QUALIFIED MANPOWER
	10,377,000	3,908,000	235,000 (6%)		
	11,507,000	4,498,000	288,000 (6.7%)		
	14,009,000	5,277,000	433,000 (8.2%)		
	18,238,000	6,290,000 *	629,000 (10%)	280,000	699,000*
	20,015,000	7,400,000	851,000 (11.5%)*	379,000*	946,000*
	21,965,000*	8,308,000	1,096,000 (13.2%)*	488,000*	1,218,000*
	24,105,000*	9,320,000*	1,491,000 (16%)*	664,000*	1,658,000*

culated by author

TABLE 2

Scientists and Engineers Employed in Canada

in 1972

By Field of Principal Employment

	1967 POP'N	1972 POP'N	%	CAN/SDI USERS	%
Engineering - Total	33,401	95,079	55%	378	16%
Aeronautical	357	1,016		24	
Ceramic	120	342			
Chemical	1,546	4,401		24	
Civil	5,426	15,446		6	
Electrical - Total	6,314	17,973		141	
Electronics	3,345	9,522			
Power	2,969	8,451			
Geological	400	1,138		81	
Industrial	4,664	13,276			
Marine	216	615			
Materials	593	1,688			
Mechanical	2,656	7,561		87	
Metallurgical	972	2,767		15	
Mining	1,239	3,527			
Nuclear	162	461			
Petroleum	1,569	4,466			
Surveying	340	968			
Textile	162	461			
Transportation	504	1,435			
Engineering n.e.s.	6,161	17,538			
Physical Sciences- Total	9,265	26,374	15%	795	34
Chemistry	4,428	12,605		519	
Atm. Hydro, Litho	2,415	6,875		99	
Mathematics	1,048	2,983		6	
Physics	1,374	3,911		171	
Physical Sciences n.e.s.					
Life Sciences- Total	8,219	23,396	13%	519	22
Agriculture	2,653	7,552		27	
Biology	1,901	5,411		462	
Forestry	2,016	5,739		18	
Psychology	441	1,255		12	
Veterinary	1,146	3,262			
Life Sciences n.e.s.	62	177			
Other	10,415	29,647	17%	672	28
Total	61,300	≈174,500	100%	2,364	100

be engaged in the various areas of the physical sciences; and 23,000, or 13% in the life sciences. Another group of almost 30,000, or 17% makes up the "other" category, i.e. those for whom field of principal employment was not identified.

Users of the CAN/SDI Project of the National Science Library are also shown in Table 2, but rather than engineering being predominant, the physical sciences are with a membership of 34%, or 795 users. Life sciences, engineering, and "other" account for the remainder with 519 (22%), 378 (16%) and 672 (28%) respectively.

Within the field of engineering, employment is most dense in electrical, civil, industrial, and mechanical, and less dense in ceramics, nuclear engineering, textile and marine engineering. CAN/SDI utilization seems to favor electrical and mechanical, as well as geological engineering. Ceramics, industrial, marine, materials, mining, nuclear, petroleum, surveying, textile and transportation engineering are not yet represented. This, however, could be a fault of our general subject coding scheme which does not properly identify all users by area of research.

In the physical sciences, employment and CAN/SDI usage correlates mostly in chemistry, while in the life sciences, biology outdistances all other employment areas in terms of subscribership, even though it is third to agriculture in number of scientists and engineers employed.

SECTOR OF EMPLOYMENT

If the 1967 trend continues, it is to be expected that the industrial sector of the economy will employ 64% of the S&E manpower, with government agencies employing 20%, and educational institutions the remainder. Industrial CAN/SDI users, on the other hand, constitute less than 10% of all users; education 32% and the government an overwhelming 59%.

By field of employment, as Table 3 indicates, the comparisons also vary. Of all engineers to be employed, approximately 80% will be in industry, but only 3% in educational institutions. In CAN/SDI, almost 48% of the engineers are now found in the government, 33% in educational institutions, and 18% in industry (Table 3.1). This concentration of CAN/SDI users is almost the same in the physical sciences, even though in terms of employment, the ratios are nearly identical for all three sectors of employment. Of all life scientists to be employed, four out of ten will be in industry and government, and one in ten in education. Subscribership, however, seems to favor the government sector, with 6 out of 10 users originating from government agencies compared to three from education and one from industry.

TABLE 3

Scientists and Engineers Employed in Canada, 1972
By field of employment, by sector of employment, if 1967 trend continues.

FIELD OF EMPLOYMENT	SECTOR OF EMPLOYMENT			
	INDUSTRY	EDUCATION	GOVERNMENT	OTHER
Engineering	79.7	3.4	15.6	1.3
Physical Sciences	23.9	22.5	20.6	33.0
Life Sciences	41.3	13.2	44.5	1.0
All Fields	64.0 %	14.5 %	20.0 %	1.5%

TABLE 3.1

Scientists and Engineers Utilizing CAN/SDI, 1972
By field of employment, by sector of employment, if 1971 subscription trend continues.

FIELD OF EMPLOYMENT	SECTOR OF EMPLOYMENT			
	INDUSTRY	EDUCATION	GOVERNMENT	OTHER
Engineering	18.3	33.3	48.4	
Physical Sciences	12.1	31.7	56.2	
Life Sciences	6.4	31.8	61.8	
All Fields	9.1 %	31.8 %	59.1 %	

MANPOWER RESOURCES AND CAN/SDI

REGION OF EMPLOYMENT

Tables 4 and 4.1 show how scientists and engineers from the various regions are to be employed by industry, education and government, and how they use CAN/SDI. Quebec has the relatively heaviest concentration of industrial employment, but the lowest in educational institutions and government agencies. The Atlantic provinces are relatively the lowest in terms of industrial manpower, but highest in governmental employment. Compared with CAN/SDI usage, Quebec has the highest industrial participation, the highest ratio of educational usage, and no governmental utilization. Ontario is highest with governmental participation, but lowest in terms of education. British Columbia has the lowest ratio of industrial utilization of CAN/SDI in all geographical regions and sectors of employment.

The geographical distribution of highly qualified S&E manpower indicates that Ontario accounts for almost 45%, and Quebec under 24%. The Prairies (Alberta, Saskatchewan, Manitoba) make up 16%, British Columbia 10% and the Atlantic provinces 6%. This compares quite well with SDI usage, because from Ontario there are 57% of the 2,364 users of CAN/SDI; and from Quebec 11%, making 68% altogether for these two provinces. Most interesting is that of all scientists and engineers in Canada, and of all S&E's on CAN/SDI, an almost identical 16% are from the Prairies. For all other regions, employment and subscriptions are at variance, with the exception of Ontario, where the ratios of the three employment fields are almost identical. These are found in Tables 5 and 5.1.

WORK FUNCTION

Table 6 indicates how the 1967 survey population occupied the largest portion of its time during a normal work week, and how our population of 174,500 S&E is expected to be engaged in 1972.

Of all the engineers, 29.6% will be involved with administration and management, while another 14.3% will be supervisors; this amounts to almost half of the 95,000 engineers. Physical scientists and life scientists have under 25% of their population in the administrative capacity.

Only 10% of the engineers are engaged in R&D, and only 2.1% in teaching. Physical scientists, on the other hand are well caught up in R&D, with one out of three employed this way. One out of every 10 life and physical scientists also teaches. One in four life scientists performs R&D.

Observations made by Atkinson also include the fact that over two-fifths of all physicists are engaged in R&D, as well as over 38% of all chemists. Life scientists indicated R&D in two out of 10 cases, with

TABLE 4

Scientists and Engineers Employed in Canada, 1972
By region of employment, by sector of employment, if 1967 trend continues

REGION OF EMPLOYMENT	SECTOR OF EMPLOYMENT			
	INDUSTRY	EDUCATION	GOVERNMENT	OTHER
Atlantic	51.5	14.5	32.9	1.1
Quebec	70.3	12.9	15.0	1.8
Ontario	63.3	14.8	20.3	1.6
Prairies	60.6	15.1	23.2	1.1
British Columbia	66.8	15.1	17.2	.9
All Regions	64.0 %	14.5 %	20.0 %	1.5%

TABLE 4.1

Scientists and Engineers Utilizing CAN/SDI, 1972
By region of employment, by sector of employment if 1971 subscription trend continues

REGION OF EMPLOYMENT	SECTOR OF EMPLOYMENT			
	INDUSTRY	EDUCATION	GOVERNMENT	OTHER
Atlantic	--	26.6	73.4	--
Quebec	18.9	78.9	--	2.2
Ontario	10.5	11.4	77.0	1.1
Prairies	6.2	66.4	26.7	.7
British Columbia	2.7	52.7	41.9	2.7
All Regions	9.1 %	31.8 %	57.8 %	1.3 %

TABLE 5

Scientists and Engineers Employed in Canada, 1972
By field of Employment, and By region if 1967 trend continues

FIELD OF EMPLOYMENT	REGIONS				
	ATLANTIC	QUEBEC	ONTARIO	PRAIRIES	BRITISH COLUMBIA
Engineering	5.6	25.8	45.0	13.8	9.9
Physical Sciences	5.1	21.8	48.4	17.1	7.5
Life Sciences	7.0	20.0	34.6	23.7	14.7
All Scientists & Engineers	5.7%	23.2%	45.3%	15.6%	10.2%

TABLE 5.1

Scientists and Engineers Subscribing to CAN/SDI, 1972
By field of Employment, and By Region if 1971 subscription trend continues.

FIELD OF EMPLOYMENT	REGIONS				
	ATLANTIC	QUEBEC	ONTARIO	PRAIRIES	BRITISH COLUMBIA
Engineering	15.9	16.7	42.9	22.2	2.3
Physical Sciences	20.0	13.2	44.2	17.7	4.9
Life Sciences	12.1	12.7	37.6	28.3	9.3
All Scientists & Engineers	12.0%	10.5%	57.4%	15.9%	4.2%

TABLE 6

Scientists and Engineers Employed in Canada, 1972
By Field of Employment and Work Function, if 1967 trend continues

FIELD OF PRINCIPAL EMPLOYMENT	ADMINISTRATION MANAGEMENT	SUPERVISION	R & D	TEACHING	DESIGN	PRODUCTION, OPERATION, MAINTENANCE	COUNSELLING PRACTICE, CASE WORK	OTHER
Engineering	29.6	14.3	10.3	2.1	12.7	5.0	0.8	25.2
Physical Sciences	18.4	6.6	32.9	13.6	1.0	2.5	0.4	24.6
Life Sciences	22.9	7.1	24.5	8.7	0.2	2.2	0.4	34.0

biologists accounting for over 50% of that activity and psychologists over 15%.

Atkinson was also able to identify patterns among the engineers who performed less R&D on a percentage basis than the other two principal fields. Fifty percent of the textile engineers, for example, perform research; this is one of the smallest areas in engineering. More than 30% of all ceramic and metallurgical engineers are engaged in research projects. Less than 6% of all civil, power, industrial, mining and surveying, which jointly comprise 44% of all engineers, perform R&D; this amounts to only 2,500 engineers.

Other characteristics observed by Atkinson under the area of WORK FUNCTION were that 75% of all R&D personnel are employed in industry, manufacturing and various government agencies; universities employed 13%, while construction firms, utilities, and professional services accounts for the remainder. If the research side of R&D is viewed, the government employes half of all the R&D personnel, industry 30%, and education 20%. With regard to the development side, 90% is employed in the industrial sector, 9% in the government, and 1% in education.

CONCLUSIONS

This study has briefly considered Canada's HQM resources, and has suggested some general growth and distribution patterns. It seems that at least 40,900 scientists and engineers in Canada are potential users of STI systems, such as CAN/SDI. The potential clientele is regarded as those actively engaged in R&D, i.e. 9,793 from engineering; 8,677 from the physical sciences; and 5,732 from the life sciences, with "others" making up 6,700. Teaching personnel are also included and constitute 10,018 persons.

Despite this large number of possible clients, not all can really be categorized as "potential" because of the variations in their information requirements. R&D requires mainly current awareness information, but it is assumed that the research activities of scientists and engineers, rather than the developmental activities, are of primary interest.

Thus, if research and teaching can be regarded as mainly representing the environment in which current awareness information is required, then the number of potential users of CAN/SDI is much smaller, approximately 28,000 persons. According to Atkinson, only 2,839 engineers, or 29% of R&D, conduct research, 5,553 physical scientists, or 64% of R&D conduct research, 4,356 life scientists, or 76% of all R&D conduct research, and 7,617 teach in all three areas. "Others" include an additional 7,292 scientists and engineers.

Of the 28,000 S&E representing about 16% of the total HQM resources in Canada 8.4% or 2,364 are already utilizing CAN/SDI. The question will,

however, be "How near to this ideal population can we get?"

REFERENCES

ATKINSON, A.G., BARNES, K.J., and RICHARDSON, E., 1970, Canada's Highly Qualified Manpower Resources: Ottawa, Information Canada.

BROWN, J.E., 1969, "The CAN/SDI Project: The SDI Program of Canada's National Science Library," Special Libraries, v. 60, no. 8, p. 501-9.

DOLAN, F.T. 1970, "Overview of COMPENDEX," Proceedings of the Annual Meeting of the Wes. Can. Chapter of the ASIS, v. 2, p. 5-9.

GAFFNEY, I., 1971, "CAN/SDI: Current Status and Future," Proceedings of the CLA Preconference Institute on Library Automation, Vancouver, in press.

LIPETZ, B.A., 1970, "Information Needs and Uses," Annual Review of Information Science and Technology, v. 5, p. 3-32.

MAUERHOFF, G.R., 1970, "NSL Profiling and Search Editing," Proceedings of the Annual Meeting of the Wes. Can. Chapter of the ASIS, v. 2, p. 32-53.

WOLTERS, P.H., and BROWN, J.E., 1971, "CAN/SDI System: User Reaction to a Computerized Information Retrieval System for Canadian Scientists and Technologists," Canadian Library Journal, v. 28, no. 1, p. 20-23.

Compendex Retrospective Searches

by

O. STANDERA

COMPENDEX RETROSPECTIVE SEARCHES

Oldrich Standera
The University of Calgary
Calgary, Alberta

ABSTRACT

A brief account is given of the retrospective searches conducted with COMPENDEX data-base tapes using IBM's TEXT-PAC system, at The University of Calgary. The required configuration and tape-format are briefly described, and the statistical option is outlined. Computer time and total cost involved in the searches are dealt with and a solution is suggested to the problem of the increasing data base. The optimum batch size is defined with the inherent limitations. The possibility of running the SDI service in the Retrospective Search module is considered.

INTRODUCTION

Retrospective Searching in the TEXT-PAC System could be defined as computer matching of a machine-readable data-base prepared as a result of manual (human, intellectual) abstracting and indexing, against one or more questions manually prepared and translated into the system language. The "hits" resulting from this matching are obtained in the form of a computer printout. Unlike some other systems, not only the title or key words (subject headings, descriptors, concepts) are searched. The entire record is scanned for the occurrence of the question words and their groupings as indicated by the logical connectors. The logic and search strategy are essentially the same as used for the TEXT-PAC Current Information Selection (see 3, 5).

TEXT-PAC RETRO-SEARCH MODULE AND COMPENDEX-TAPE SERVICE

The complete documentation of the TEXT-PAC software may be found in (1). The programs are in Basic Assembler Language (BAL) and are designed for the IBM's OS/360 (MVT or MFT). The required configuration comprises the system 360 and needs 180K core memory, a card reader, a

printer, four 9-track tape drives, and one DASD (e.g., scratch disk as temporary storage). The mode of computer processing is local batch.

COMPENDEX is supplied by the Engineering Index, Inc. on 9-track tapes 800 BPI in EBCDIC. Tape length is 1,200 feet. It is delivered monthly and contains some 6,000 records. Records are variable length, unblocked, maximum length 8,004 bytes. The input format is TEXT-PAC 360 Condensed Text. More information about the tapes may be obtained from (10).

Each record is classified by Main Subject Headings and Subheadings which are listed in (8). Another access point to the records represents the CAL (Card-A-Lert codes) described in (9).

Publications which are abstracted and indexed for COMPENDEX are listed in (7) together with the type of coverage: complete; partial; or monitored.

STATISTICAL OPTION

As we have already mentioned in our COMPENDEX Retro-Search Instructions (4) the user can obtain statistical data indicating which of the logic (words and logic connectors) has been responsible for the hits, if any were accomplished. This option is specified on the Header card (column 9) at the time a question is coded.

The statistical printout (or trigger cards) could be used, theoretically, to one or both of these objectives:

1. To decide what documents hit by the question should be printed. The trigger cards would make it possible. However, it seems to us that a responsible decision in this respect cannot be made with only trigger cards and/or statistical printout at hand. This would necessitate checking over the pertinent abstract in the Edit print which would have to be printed at an extra cost. Checking the printed answers is less time consuming and, therefore, the better alternative.

2. The statistical data about the hit logic provide the means for improving a profile. In this connection it should be stated that the statistical feature being described seems to be more appropriate in the CIS mode, where the profile is of a semi-permanent nature and thus has to be corrected continually on the basis of user's feedback. We

can, of course, modify a retrospective question in the event that there are either too many or too few answers.

Figure 1 illustrates what the programs do for the user depending on his option.

SEARCH TIME

In order to ascertain the effect of the number of questions, we have taken a data-base of 60,000 records which resulted from merging of individual monthly tapes, and determined the CPU times of the search programs for 1, 2, 3, 5, 10, 20, 30, 40, 50, 60, 70, 80, and 100 questions, with 12 hits per question.

To show the relationship between the CPU times and the number of records, we conducted a search for 10 questions against a data base consisting of 5,000; 10,000; 20,000, 40,000; 60,000; and 80,000 records.

It has been shown that the CPU time of the search programs is influenced by the number of questions (directly proportional), by the number of data-base records (directly proportional), and by the number of hits. We have not examined the impact of the number of hits as they can be monitored only indirectly and they vary from question to question. The relationship "CPU time to number of questions" is illustrated in Figure 2. The relationship "CPU time to number of records" is depicted in Figure 3 and Figure 4. In the former case, the number of hits per question was kept constant (12 hits per question); in the latter case, of course, the number of hits per question was increasing with the size of the data base. In the "CPU time per number of records" chart, the effect of looser questions on the search time is clear: the CPU time for 10 questions and 60,000 records equals 35.5 minutes, whereas in the chart "CPU time per number of questions" the CPU time for 10 questions and 60,000 records is less than 28 minutes. This difference reflects the different number of hits brought about by the looser question structure in the former case.

Cost of the Service

In calculating the cost of the service we adopted 60,000 records as a base of our calculations since this figure represented a yearly data-base increase at that time. The cost was computed for 5 and 50

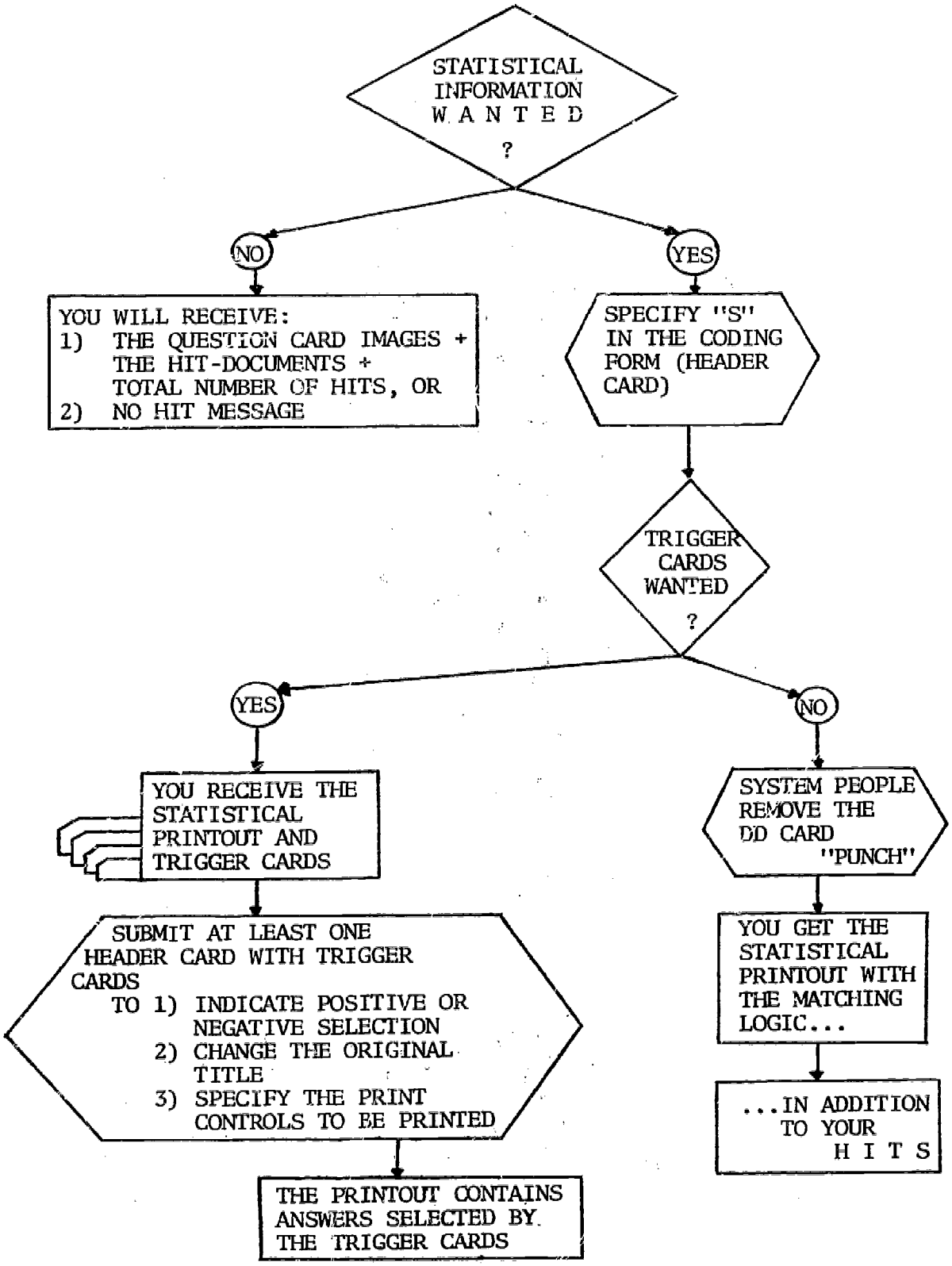


Fig. 1 Statistical Option: Decision making

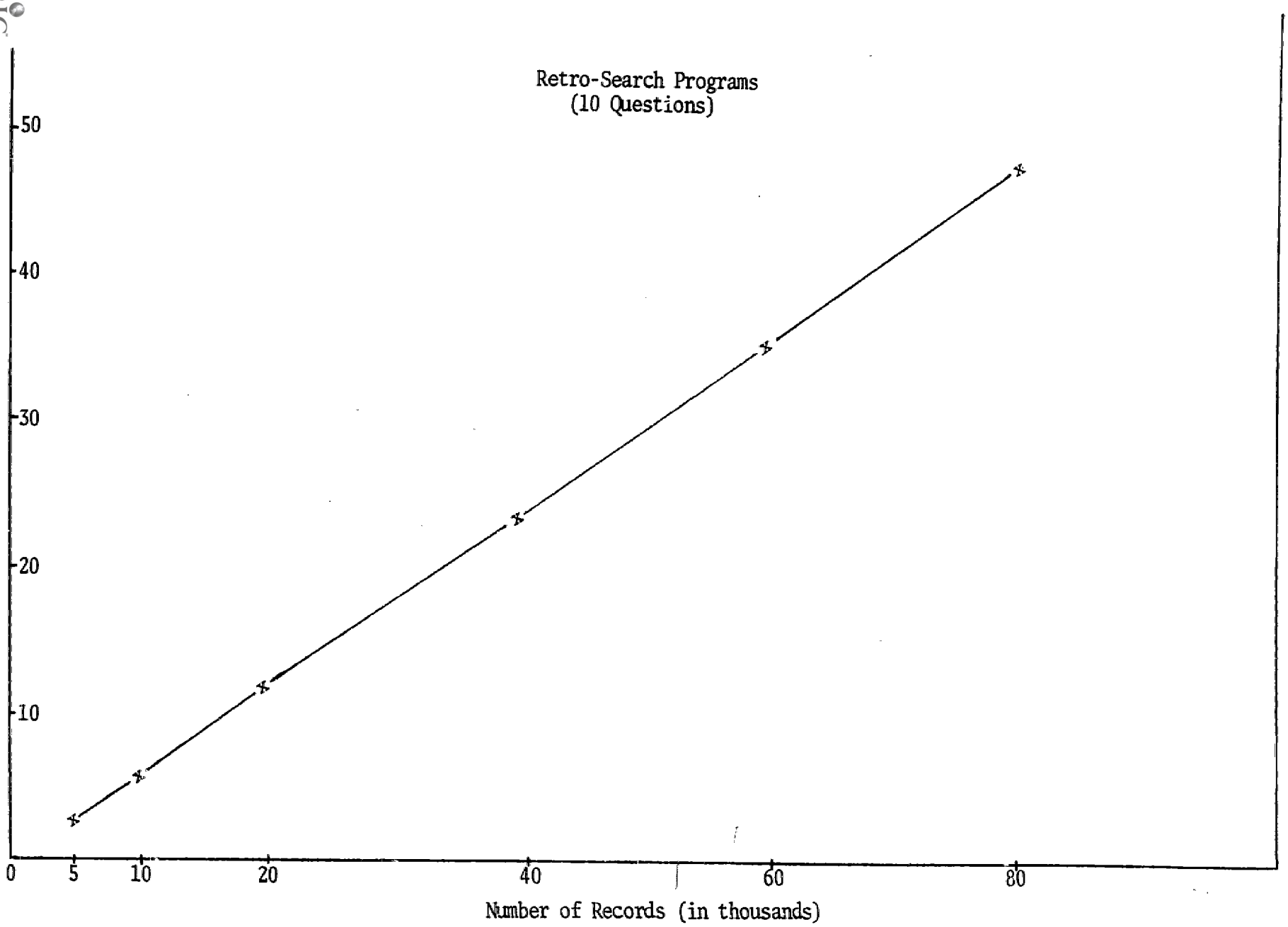


Fig. 3 Number of Records vs. CPU Time

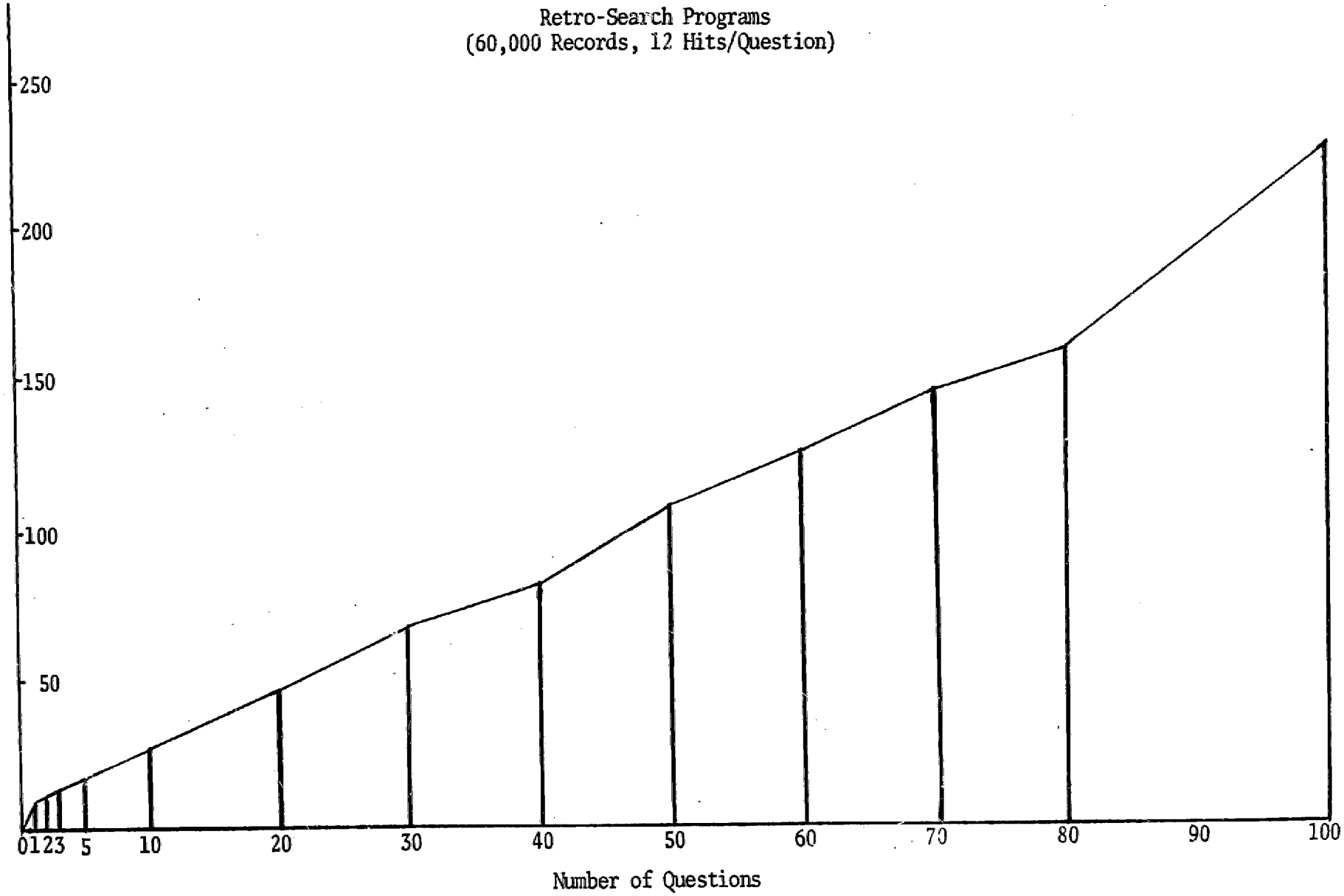


Fig. 2 Number of Questions vs. CPU Time

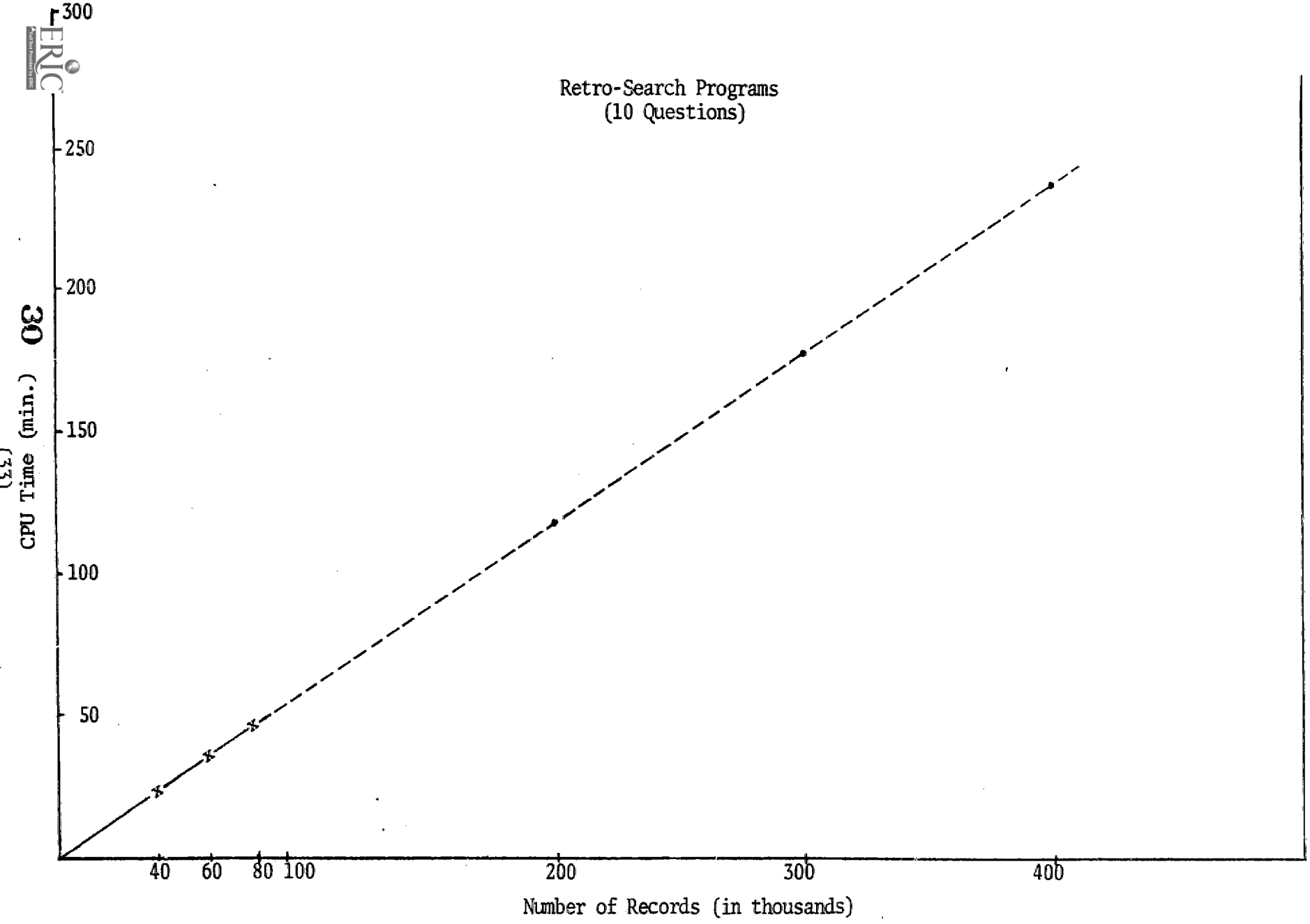


Fig. 4 Number of Records vs. CPU Time (extrapolated)

questions representing both a small and a large batch of questions. Both the "statistical" and "non-statistical" versions were examined.

The total cost encompasses the following components: Computer Costs, Cost of the System, Cost of Implementation, Search Editing etc., Keypunching-Verifying, Material, Handling-Mailing, other Overhead Cost.

Computer Costs (CC) is the sum of the component costs (CPU, Core and Input/Output) for each of the programs.

$CC = CPU + C + I$

CPU was calculated at \$85.00 per hour

$C = R \times (C_t + I_t) \times 0.20$

where R = Core requested (K)

C_t = CPU time (hours)

I_t = Input/Output time (hours)

\$0.20 is the cost of K/hour

$$I = \frac{(I_c \times 0.09 \text{ sec})}{3{,}600} \times 60$$

where I_c = Input/Output count

\$60.00 is the cost per hour

The resulting cost per question is as follows:

1 Question

Out of Five		Out of Fifty	
Non-statistical	Statistical	Non-statistical	Statistical
\$64.46	\$64.65	\$24.52	\$24.66

From the cost calculations several conclusions may be drawn. First of all, we can infer that the statistical option should be used wherever needed because of its merits and low additional cost.

Secondly, questions should be run in optimum batches. Whereas the size of a batch cannot influence the question-dependant costs e.g.

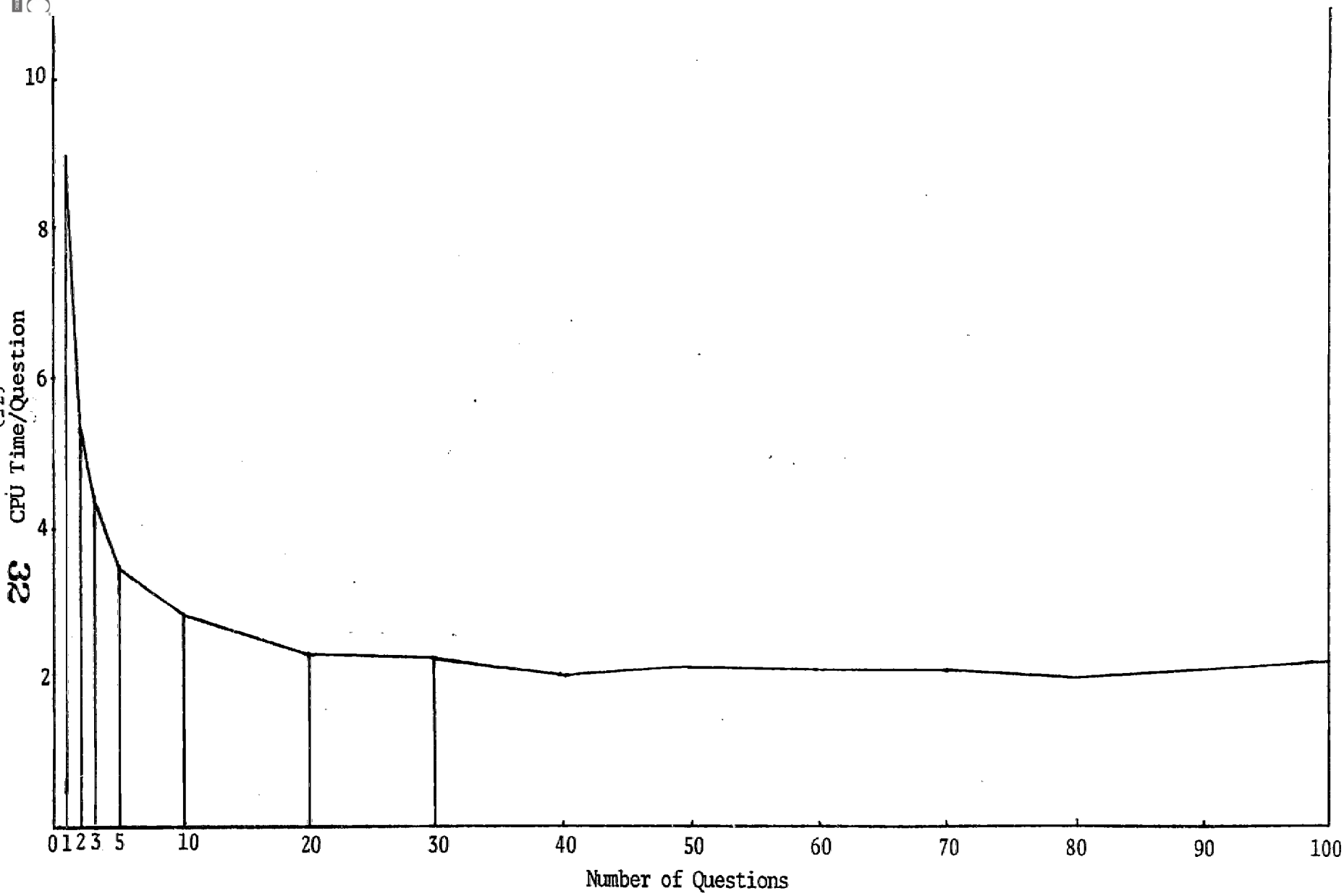


Fig. 5 CPU Time/Question (60,000 Records, 12 Hits per Question)

Search Editing, Keypunching, it will have a marked effect on the total and computer costs as may be seen from the table above. In our example (Fig. 5) the CPU time required to run 1 question is 8.5 minutes as compared with 2 minutes per question when processing a 40-question batch. The optimum search time sets in at 20 questions and extends up to the other limiting factor which is the capability to process one "memory load" of questions at one time: one memory load is approximately 100 questions (or slightly above, depending on the size of questions). If more than one memory load of questions are to be processed, two or more runs will be necessary.

Yet this optimum range of questions to be processed at one time (20 through 100) has another restrictive condition, namely the number of hits. The maximum number of hits which can be handled by the "Retrospective Text Sort" program is 6,000. A larger number of hits can be accommodated by using the IBM 360/OS Sort Program. An excessive amount of hits, however, prevents other users from running their jobs for hours.

Cost/Benefit

The question, which is always asked, is whether the cost of a service is justified by the benefits from the service.

Assume we have processed a question along with others in a batch of 50 against one year's data base of 60,000 records. The cost of this search has been $24.66 with the statistical option. Most of the information services are subsidized in some way or other, so the actual price to the user would be lower.

If our user has to cope with his information problem using hard copies of an abstract journal, he obviously does not have to scan all of the 60,000 abstracts, but rather approximately 1/10 of the abstracts, in some cases more, in others less. If he goes through 1,000 abstracts he probably would scan six of them in one minute. Getting through 6,000 abstracts would reduce the efficiency of scanning to four per minute. This literature search would take 25 hours and cost $250, if we charge only the research worker's salary and disregard the value he could generate if he were freed for his special work. This would represent a multiple of this amount. If he subscribes to some file card information service, his recall will be lower than in full text searching and the price is to be added to the cost of personal searching.

Frequently, however, a literature search is not done and this does not mean that the amount of $250 is saved. Rather, some work already done elsewhere is duplicated, other people's patent rights are infringed and the work itself is not done at the level it might have been had the literature been searched.

This once again substantiates the fact that experimenting in the

literature is cheaper than experimenting in a laboratory. It also proves that some organizations could increase their capacity by as much as one third by using professional information services.

DATA-BASE GROWTH

After a couple of years the size of the data base would make the search too lengthy and costly. As already mentioned the expected yearly growth is about 70,000 records. After five years the data base would represent 350,000 records on 35 tapes. As our graph (Fig. 4) indicates this would require 180-210 minutes of search time for 10 questions with a small number of hits. The most appropriate solution to this problem seems to be to subdivide the data base into a series of subject areas. This would enable us to confine the search to a data base of a limited size and obviate searching in its irrelevant regions.

The Card-Alert Codes of COMPENDEX would help in creating subsets. For example, after three years of operation, we would have over 200,000 records. At this time it would be practical to subdivide it into:

1. Civil--Environmental--Geological--Bioengineering
2. Mining--Metals--Petroleum--Fuel Engineering
3. Mechanical--Automotive--Nuclear--Aerospace Engineering
4. Electrical--Electronics--Control Engineering
5. Chemical--Agricultural--Food Engineering
6. Industrial Engineering--Management--Mathematics--Physics--Instruments

Instead of handling 20 tapes in a search, one would have to process approximately 3 of them, or 6 if the question would be expected to get response in two of the subsets specified above. After, say, two more years further splitting would take place separating e.g., aerospace engineering in a self-contained subject-field subset, and so on.

CIS IN RETRO-SEARCH MODULE

As the Retrospective-Search module has the "statistical option" indicating the matched words by a particular document, and the CIS module (2) does not, we have to solve the following dilemma: either (1) to "transplant" this option to the CIS section, or (2) to use the Retrospective-Search section to process the CIS profiles. The first alternative would entail study and reprogramming, but would not necessitate a change in the Header cards and would leave the output (the double cards) unchanged. The second alternative is more convenient because the profiles can be run after minor formal changes (see the CIS profile form and Retro-Search question form for more details) with the limitations as they were outlined for the Retro-Search: only one memory load of profiles can be run at one time; output is on printing paper, but can be easily reprogrammed on the feedback cards; maximum 6,000 hits are

recommended.

The costs of running 100 questions (in CIS called profiles) against 5,000 documents, with the statistical option, producing 5 hits per question, were analyzed. The cost per 100 profiles amounts to $1,186.51 or $11.87 per profile/month.

The most significant cost item is represented by the data-base tapes with reels which amount as high as 44.3 per cent of the total. This illustrates also the way to go if we plan to enhance the economy of the service: to process as many profiles as possible (with physical limitations in view) to keep the proportion of this cost per profile low. Further, the economy of the CIS service can be improved by retrospective searches which should be given wide publicity. Only the multiple use of this data-base can make it economically viable. As it is a fixed cost, its proportion per profile is decreasing with the rising number of profiles.

Search Editing (33.7%) represents a proportional cost which increases directly with the number of profiles. Seemingly, we can get more out of a monthly salary if we divide it by a higher number of profiles. This is a wrong approach, though, as it affects the quality. There is a certain limit imposed on the capacity of a search editor and after that we need additional search editors which, in turn, increases the costs.

The computer processing is a rather surprisingly low percentage of the total cost (18.51).

Summary

Retrospective searching in the COMPENDEX data base is now fully operational, in addition to the SDI mode (2) and indexes (6).

The COMPENDEX data base is available commencing January, 1969 and the yearly growth is expected to be 70,000 records, or seven tapes. The data base has proven to have a good mega-relevance to all of the areas of engineering. The system can operate over a wide range of relevance and recall values.

It has been shown that the CPU time of the search programs is influenced by the number of questions, by the number of data-base records and hits. We have found that one-question run requires as much as 8.5 minutes of the CPU time, whereas with a 40-question batch only two minutes per question are needed. The optimum search time sets in at 20 questions and extends up to the "memory load" or approximately 100 questions which can be processed in one run. The maximum number of matches processed in one run should be about 6,000, otherwise the standard utility sort program has to be used.

The statistical option should be used because of its merits and low additional cost. The cost of one question in a five-question batch is $64.46 (statistical $64.65), and it drops to $24.52 (statistical $24.66) for one question out of fifty; this applies to searching 60,000 records and 12 hits per question. These figures illustrate the effect of running the optimum size batches (20-100 questions).

It is suggested that the CIS service or SDI (Selective Dissemination of Information) be also run in the Retrospective Search module. This would enable us, with the statistical printout at hand, to adjust the profiles accordingly. We regard the statistical option as even more significant in the SDI service in view of the dynamic character of profiles. The costs of searching are reasonable. (One profile out of one hundred costs $11.87 per month, with five received answers.) Since the cost of the data base is the most expense, a better economy can be achieved by greater use of it.

In view of the substantial yearly data-base increase it is suggested that the last one or two years' data base be searched after simple merging, but the "historical" data base should be presorted to make up subject-area tapes. The Card-Alert Codes of Engineering Index would serve this purpose. Through this subsetting, the data base searched could be maintained at a reasonable size.

REFERENCES

(1) Kaufman, S. TEXT-PAC, S/360 Normal Text Information Processing, Retrieval and Current Information System (360.D.06.7.020). Armonk, N. Y., 10504: ITIRC, IBM Corporate Headquarters, 1968.

(2) Standera, O. R. COMPENDEX/TEXT-PAC(CIS) Project Report. Information Systems and Services Division. Calgary: The University of Calgary, 1970.

(3) _____. COMPENDEX Profiling Guide. Information Systems and Services Division. Calgary: The University of Calgary, 1970.

(4) _____. COMPENDEX Retrospective Search Instructions. Information Systems and Services Division. Calgary: The University of Calgary, 1970.

(5) _____. Profile Adjustment Manual. Information Systems and Services Division. Calgary: The University of Calgary, 1970.

(6) _____. TEXT-PAC Input, Bulletin and Indexes. Information Systems and Services Division. Calgary: The University of Calgary, 1971.

(7) PIE-Publications Indexed for Engineering. New York: Engineering Index, Inc., 1970.

(8) SHE-Subject Headings for Engineering. New York: Engineering Index, Inc., 1970.

(9) Card-A-Lert-Selective Information Service. New York: Engineering Index, Inc., 1970.

(10) COMPENDEX-Computerized Engineering Index. New York: Engineering Index, Inc., 1970.

Streamed Information System III

by

J.S. TRUSWELL

STREAMED INFORMATION SYSTEM III

J.S. Truswell
Imperial Oil Limited
Calgary, Alberta

ABSTRACT

An information-management system utilizing *COBOL* and *Informatics Mark IV* programs is described. The system was developed in and for an industrial information center. Design priorities were placed on modularity and project orientation. The system has processed data for up to 21 projects and data bases for 2 years. It has paid out its development costs through reduced operating costs.

INTRODUCTION

Imperial Oil Limited, Technical Information Services Department, first developed a computer-assisted, information-management system in 1963. The *COBOL* system, *Streamed Information System I (S.I.S. I)*, was to enable the Department to provide better information services than had been possible with traditional cataloguing and to cope with a rapidly growing volume of information.

S.I.S. I fulfilled its purpose from 1964-1969 operating on the IBM 1410 and producing inventory listings, current awareness bulletins, vocabulary control tools, keyword indexes, and a permanent, machine-readable data base (Cherry 1966, 1965, 1964). Although developed to process books and reports, the System showed its general value in the management of company and published map files, visual aids, computer programs and other media.

In 1968 a move to the IBM 360/, five years of experience with *S.I.S. I*, and the acquisition of the Informatics *Mark IV File Management System* led to the development of *Streamed Information System III (S.I.S. III)*. The new system was designed and programmed in-house and has been on production since July 1969. Some of the original *COBOL* programs were rewritten and others were replaced by *Mark IV* procedures. The economic success of *S.I.S. III* can be judged from the fact that it had paid out its development costs in two years through reduced operating costs (Truswell, 1971). The design of all aspects of the system have allowed it to function with little human intervention and hence no high personnel cost to the user.

The most significant advantage offered by *S.I.S. III*, however, has been the increase in file manipulation and retrieval capabilities. The

Mark IV system is primarily responsible for this, offering:

1) Boolean search capabilities;
2) search, manipulation and retrieval of any defined file element;
3) flexible output format specifications, independent from the query;
4) flexibility in file and transaction definitions, either of which may be easily modified at any time to respond to changing requirements.

In addition to the above features, a *Mark IV* text scanning and processing package now available will further improve *S.I.S. III* capabilities.

S.I.S. III is now also operating at Standard Oil of New Jersey where new programs to upgrade and extend it are being developed. Other Imperial affiliates are also evaluating the system.

DESIGN SPECIFICATIONS

The main design specifications illustrate some of the differences between *S.I.S. III* and *S.I.S. I* or *S.I.S. II* (University of Calgary, 1970):

1) A common input program for all data. This seems highly compatible with the concept of a "streamed" information system.
2) The Vocabulary/Document Edit or check of keywords to be kept independent of all other functions. This allows easy exclusion of the edit from indexing projects not using vocabulary control. Systems where the edit is "bound" to the update and maintenance routines suffer a loss of flexibility.
3) The Vocabulary/Document Edit to apply to transactions only. The editing of the entire master file during each run is unnecessary and expensive. A Master Edit program allows periodic editing of the document master file to resolve changes in the Vocabulary and indexing policy.
4) A single Document master file to be maintained in document number sequence. An inverted (or index-format) file becomes very bulky, presenting extreme cost and maintenance problems.
5) The system to be designed to process data on a project basis with each project having unique identification, data bases, flow charts, run procedures, report specifications and schedule.
6) The system to be highly modular with the possible use of various combinations of modules to suit specific requirements.

MODULES

The key and overriding objectives of the above specifications are modularity and flexibility. Experience has shown that these qualities

were the most valuable characteristics in a system which must support up to 20 separate indexing projects and data bases varying widely in type of document, indexing level and depth, input volume, update frequency, and report requirements.

Figure 1 shows the main *S.I.S. III* modules:

1) *Input* consists of card-to-tape and separation of Vocabulary from Document transactions. The input module is used in all projects.
2) *Housekeeping* produces inventory lists of Document transactions. Used in almost all projects where documents are processed.
3) *Current Awareness* produces various types of announcement bulletins. Used in only one project.
4) *Vocabulary Maintenance* updates and maintains a hierarchical Vocabulary file. Used in all projects having a dynamic or open Vocabulary.
5) *Vocabulary Edit* checks all keyword transactions against the Vocabulary. Used in most projects having both a Document and a Vocabulary file.
6) *Vocabulary Print* produces a Vocabulary report. Used with varying frequencies.
7) *Document Maintenance* allows update and correction to a Document file. Used in all projects having a Document file.
8) *Index Producer* selects records and produces reports in specified formats. Used with varying frequencies to produce various reports, not only keyword indexes.
9) *Search*: the *Mark IV* search utilizes the same procedures as the Index Producer module. The *Document Processing System* is available but no longer used (due partially to its output constraints, e.g. it can't produce an index). The *Inquire* system is available but not yet installed.

SYSTEM DESCRIPTION

General

S.I.S. III operates on OS360/ requiring a maximum of 128K bytes of core storage, three tape drives and 20 disk cylinders. The system is JCL controlled, has 20 job steps and disk spools all reports except the indexes and vocabularies which are tape spooled. Four tape files are produced from a complete job, with all transaction files being passed on disk and scratched after use. C.P.U. time for a complete job updating a master file of 2,000 documents with 200 new document transactions (approximately 2,000 cards) is approximately 4 minutes.

Input

The input is punched from handprinted or typed forms onto standard

STREAMED INFORMATION SYSTEM III

MODULES

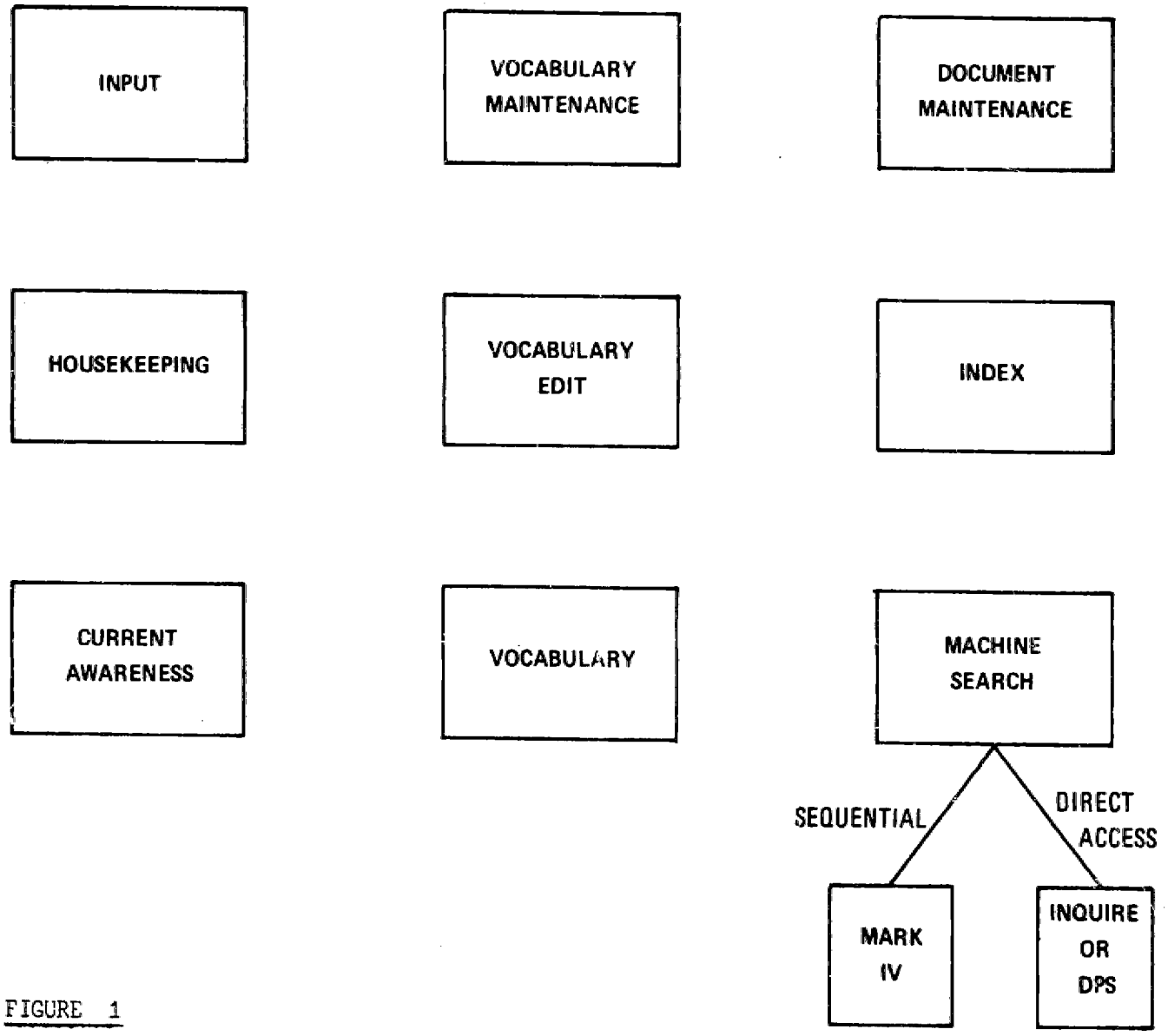


FIGURE 1

80 column cards. There are five card or transaction "types" involved in new transactions and one additional type for corrections to the master files. All other correction cards are variations of the six basic types.

The basic input or indexing form is illustrated in Figure 2. All cards begin with a six character, alphanumeric document number. This number is the "tie" between all cards related to a given document and later becomes the record key on the master file. Columns 7 and 8 bear the preprinted card type. The cards are as follows:

01 Card: contains fields which indicate the publication or indexing date, the document's general location (where there are multiple storage locations), and group and output codes.

The group code is generally used to distinguish broad categories of documents, e.g. book = 01, published report = 02, but may be used for other purposes. The output controls designate other broad categories which are useful in producing subfiles and special indexes. These include disciplines (01-15), corporate authors (20-36), geologic ages and geographical areas.

The output controls are extremely powerful and economical search tools and as such their proper coding is very important. Most of the elements on the 01 card are not printed in the routine indexes but, due to its importance in file manipulation, this card is used as the record creator key and documents lacking it are rejected.

02 Card: contains the first 72 characters of the title and bibliographic information.

03 Card: contains the second 72 characters of the title and bibliographic information.

04 Card*: contains the last 44 characters of the title and bibliographic information.

05 Card: contains the indexing keywords which include the authors. The use of "weight" indicators in columns 10 and 42 allows the grouping of concepts by importance or type. Numerical weighting for weighted searches is not practiced.

New terms may be added to the Vocabulary by placing an X in column 9 or 41.

Keywords are limited to 30 characters, 2 to 1 card.

*The limitation of the title block to 188 characters is based on experience.

S & CS—WR 554
4554

IMPERIAL OIL LIMITED – PRODUCING – WESTERN REGION

TECHNICAL INFORMATION SERVICES

1 6

INDEXING FORM

7 8 02 TITLE 9 80

7 8 03 9 80

7 8 04 9 80

01

DISCIPLINES	COMPANY & AFFILIATE	OTHERS	OTHERS

MAXIMUM OF TEN * COLUMNS 9 – 28

DISCIPLINES	COMPANY & AFFILIATE
01 CHEM. & PHYS. MATH	20 SONJ
02 GEOLOGY	21 EPRCo
03 GEOPHYSICS	22 HUMBLE
04 GEOCHEMISTRY	23 IOL REGIONAL
05 DRILLING	24 IOL PR. & TS.
06 WELL LOGGING	25 IOL EDMONTON DIV.
07 RESERV. ENG.	26 IOL REGINA DIV.
08 PROD. ENG.	27 IOL PEACE RIVER DIV.
09 COMPUTER USAGE	28 IOL FOOTHILLS
10 ECONOMICS	29 IO ENTERPRISES
11 STATISTICS	30 C.S. (A)
12 MGMT. & ER.	31 ESSO RESEARCH & ENG.
13 INFO. RET.	32 IOL TORONTO
14 ALLIED INT.	33 IOL CHATHAM
15 OFFSHORE OPERATIONS	34 IOL NWT EXPL. DIST.

GROUP CODE 29 30

LOCATION M 31 O 32 H 33

DATE INDEXED YR. 34 MO. 36 DAY 38

CONCEPT WEIGHTS: PRIMARY DESCRIPTOR – S; SECONDARY DESCRIPTOR – ; AUTHOR – A; CURRENT AWARENESS – C

VOCABULARY ADDITION – X

CARD TYPE	VOCAB	WEIGHT	CONCEPT	VOCAB	WEIGHT	CONCEPT
7 8	9	10	11 – 40	41	42	43 – 72
05						
05						
05						
05						
05						
05						
05						
05						
05						
05						
05						
05						
05						
05						
05						
05						

FIGURE 2

Correction Cards: for adding or deleting hierarchies to Vocabulary terms, deleting keywords or documents from the master files and revising keyword weights are coded 98 in columns 7 and 8. The Vocabulary corrections are distinguished by the literal VOCCOR in columns 1-6 while the document corrections bear the appropriate document number.

Corrections to 01, 02, 03 and 04 cards and the addition of 05 cards are accomplished via the same card types as on the indexing form.

Programs and Reports

Figure 3 lists the individual programs and their functions. The discrepancy between the 15 programs and the aforementioned 20 job steps arises from programs C3942 and C3910 requiring 3 and 4 steps respectively.

Figure 4 shows the organization of the programs, the flow of data through the system and the reports generated by each program.

C3906U: is the card-to-tape spool for all transactions.

C3906: is the input processor which accepts all transactions, checks their validity and creates transaction files for both the Vocabulary and Document update modules. The program also creates Vocabulary transactions from 05 cards bearing the appropriate flag.

C3906 generates two reports, *Invalid Transactions* and *Control Totals*. The former presents an image of all invalid cards while the latter reports the total cards read, total invalid cards rejected, total new documents entered and total new Vocabulary terms entered.

C3932: accepts the Vocabulary transactions from C3906 after they have been sorted by transaction number and type by C3932S. Hierarchical records are built by C3932 and reciprocal entries are created (these are reversed entries for related terms, broader term entries for narrower terms, etc.).

The C3932 *Control Totals* report gives Vocabulary transactions read, Vocabulary corrections written, new Vocabulary main terms written and the sum of corrections and new main terms.

C3941: updates and maintains the Vocabulary master file with transactions received from C3932, sorted alphabetically by C3941S. C3941 builds reciprocal entries for corrections involving the deletion of main terms. The program accomplishes this by reading the hierarchy accompanying the term from the master-in and writing the proper corrections out on another tape, the Vocabulary transcycle. The transcycle is read in at each update to resolve the cycled corrections from the previous run.

STREAMED INFORMATION SYSTEM III
PROGRAM DESCRIPTIONS

NUMBER	LANGUAGE	NAME	FUNCTION
C3906U	UTILITY	CARD-TO-TAPE SPOOL (All TIS Transactions)	
C3906	COBOL	INPUT PROCESSOR	SEPARATES DOC AND VOC STREAMS
C3932S	SORT	SORT VOCABULARY TRANSACTIONS	SORTS BY TYPE
C3932	COBOL	BUILD VOCABULARY RECORDS	BUILD HIERARCHIES
C3941S	SORT	SORT VOCABULARY RECORDS	SORTS ALPHABETICALLY
C3941	COBOL	VOCABULARY MASTER UPDATE	TRANS + RECYCLE ONTO MASTER
C3935	COBOL	VOCABULARY REPORT	
C3911S	SORT	SORT DOCUMENT TRANSACTIONS	SORTS BY DOCUMENT NUMBER & CD TYPE
C3911	COBOL	INVENTORY REPORT	INCLUDES ALL CARDS IN EACH DOCUMENT
C3934S	SORT	SORT DOCUMENT TRANSACTIONS	SORTS ALPHABETICALLY BY DESCRIPTOR
C3934	COBOL	VOCAB/DOC TRANSACTIONS EDIT	DOCUMENT TRANS vs VOCABULARY MASTER
C3909S	SORT	SORT DOCUMENT TRANSACTIONS	SORTS BY DOCUMENT NUMBER
C3909	MARK IV	DOCUMENT MASTER UPDATE	ALL NEW DOCS AND CORRECTIONS ONTO MSTR
C3910	MARK IV	CURRENT AWARENESS BULLETIN	OUTPUT CONTROLS AND KEYWORD SORTS
C3942	MARK IV	INDEX REPORT	SELECTION AND PRINT

FIGURE 3

STREAMED INFORMATION SYSTEM III

FLOWCHART SHOWING ALL REPORTS GENERATED

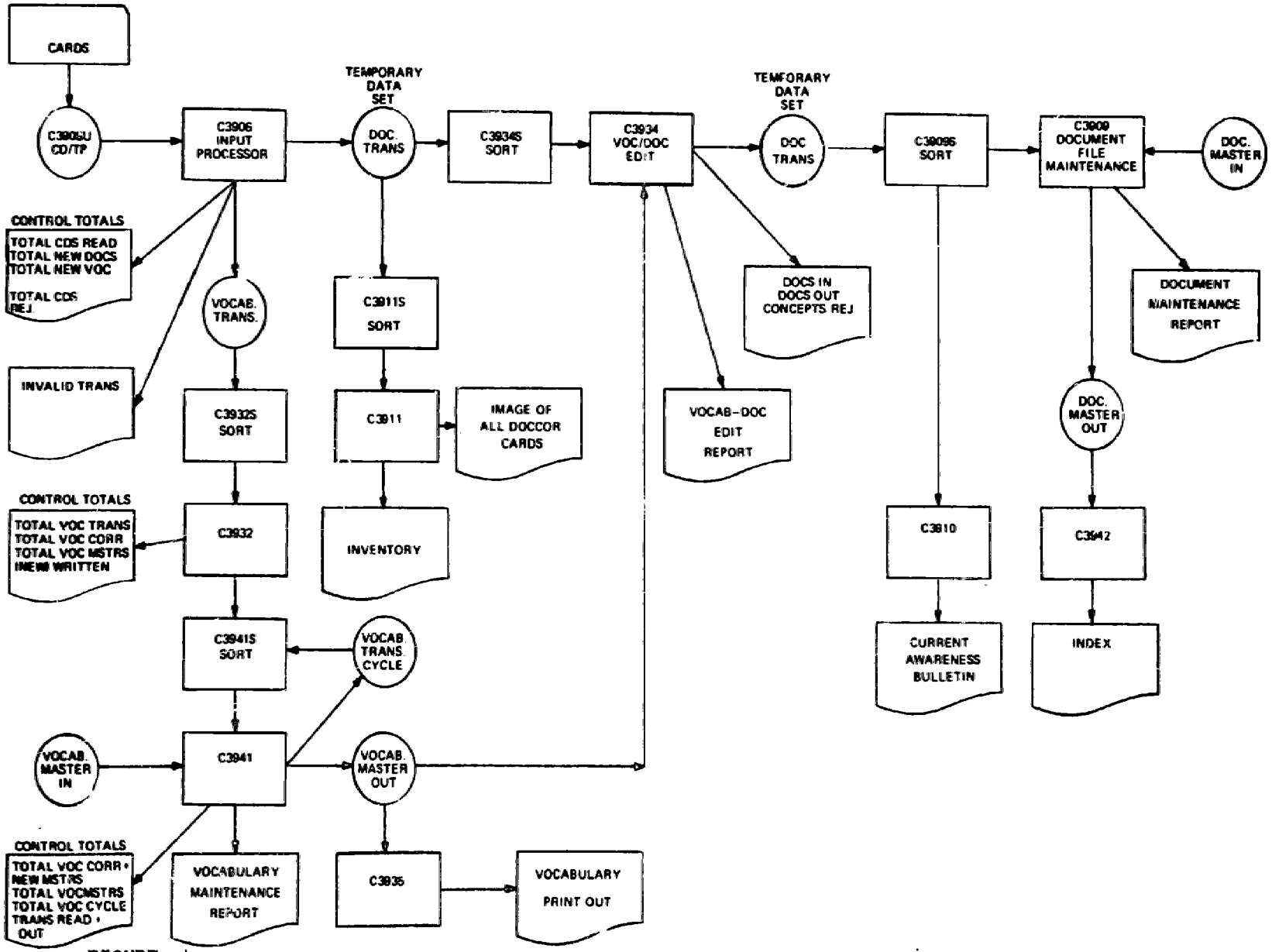


FIGURE 4

C3941 generates two reports, *Control Totals* and *Vocabulary Maintenance*. The *Control Totals* are total main terms (on the Vocabulary master) read and written, new entries and cycled transactions read, and transactions cycled out. The *Vocabulary Maintenance* report shows images of all correction cards submitted and flags those which were rejected due to input error.

C3935: generates the *Vocabulary* report from the updated master file received from C3941.

C3911: accepts the Document transactions file from C3906, sorted by document number and card type in C3911S. The main purpose of C3911 is to generate the *Inventory* report of new documents entering the system. This report lists all information for all documents in document number order.

A second report from C3911, *Document Corrections*, presents an image of all Document correction cards.

C3934: receives an inverted Document transactions file, sorted by keyword in C3934S, and the updated Vocabulary master file from C3941*. All keywords on the Document transactions file are checked against the Vocabulary master and those not finding a match are rejected.

Rejected keywords are listed by the *Vocabulary/Document Edit* report. The C3934 *Control Totals* report lists total documents read, total documents written, and total concepts rejected.

C3910: receives the edited Document transactions, sorted by document number in 3909S and generates 12 *Current Awareness* reports by selecting documents bearing the appropriate Output Controls. The reports are in index format.

C3909: updates and maintains the Document master file with the transactions from C3909S. New documents are added to the master in sequence and corrections are executed. A *Document Maintenance* report shows all additions and corrections which are invalid.

C3942: generates *Index* reports from the updated Document master file. Parameters for the contents of the index are extremely flexible.

E2401, E2402, E2403 (not illustrated in Fig. 4): are the programs which edit both the Vocabulary and Document masters. They are run infrequently, e.g. once per year on a project which has scheduled runs bi weekly.

*C3934 cannot be run until all new Vocabulary terms have been added to the Vocabulary master by C3941. For this reason all Vocabulary programs are run first.

STREAMED INFORMATION SYSTEM III

PROJECTS

Figure 5 illustrates the Projects which have used *S.I.S. III* and the general flow of data from the client department through the system.

STREAMED INFORMATION SYSTEM III
PROJECTS AND DATA FLOW

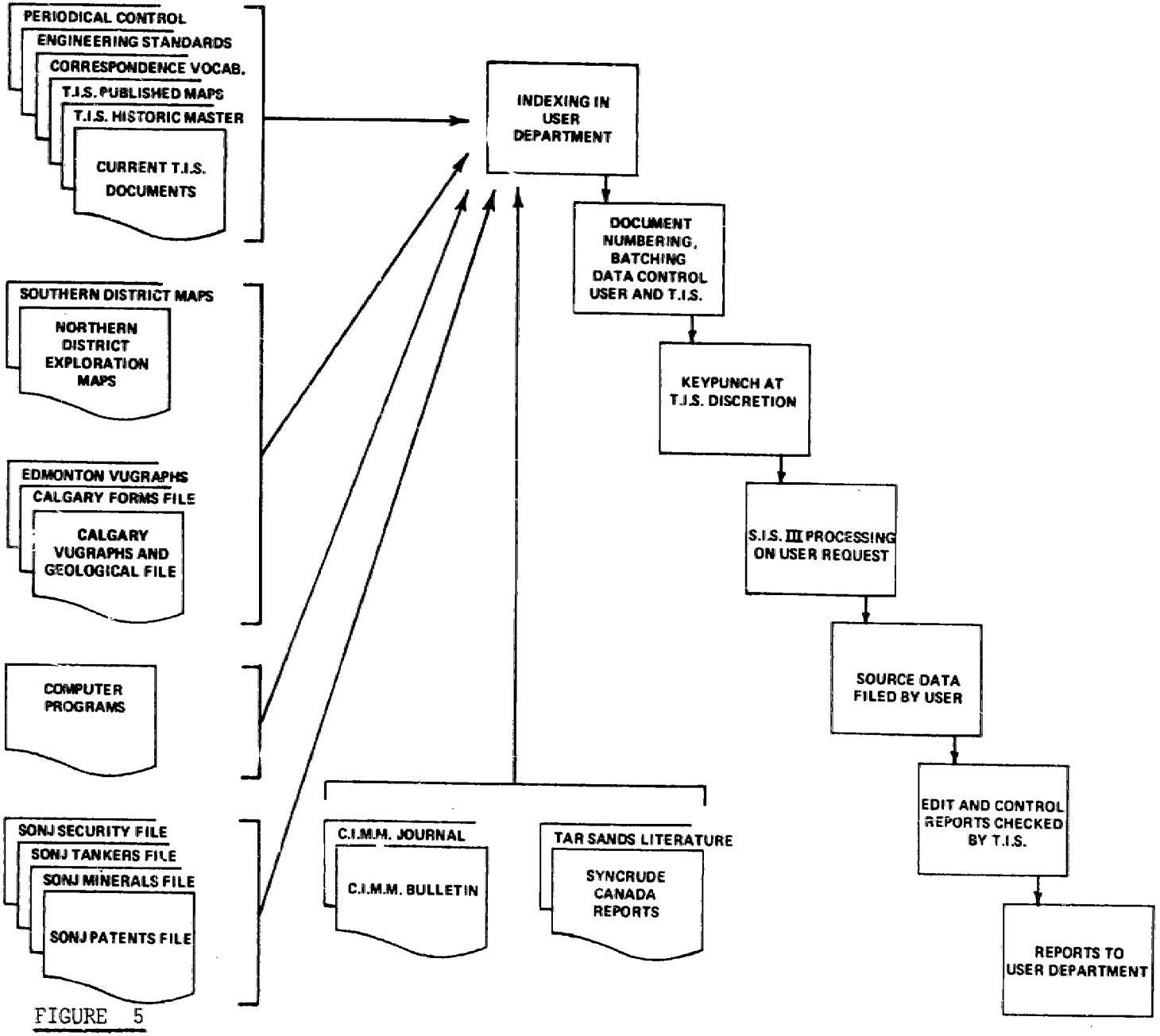


FIGURE 5

REFERENCES

CHERRY, J.W., 1966, "Computer-Produced Indexes in a Double Dictionary Format," *Special Libraries*, p. 107-110, Feb.

_____, 1965, "A Computer-Assisted, Industry-Oriented Information Retrieval System," *Canadian Library Association, Occasional Paper 48*, June.

_____, 1964, "Automation and Information Retrieval," *Proceedings Computing and Data Processing Society of Canada 4th Conference*.

INFORMATICS INCORPORATED, 1970, "Mark IV User's Manual," 5430 Van Nuys Blvd., Sherman Oaks, Calif. 91401.

TRUSWELL, J.S., 1971, "Streamed Information System III, Development and Operating Costs," Imperial Oil Ltd. report, Calgary.

UNIVERSITY OF CALGARY INFORMATION SYSTEMS DIVISION, 1970, "S.I.S. II - Streamed Information System-Conference, Nov. 13-14, 1969," *Information Systems Report No. 3*, Jan.

An Information Retrieval Laboratory for Computing Science Students

by

J. HEYWORTH

AN INFORMATION RETRIEVAL LABORATORY FOR COMPUTING SCIENCE STUDENTS

Janice Heyworth
Department of Computing Science, University of Alberta
Edmonton, Alberta

ABSTRACT

Computing science students educated in information storage and retrieval methods are need d for the implementation and maintenance of automated library and information systems. An option in information storage and retrieval offered within the Department of Computing Science at the University of Alberta, Edmonton, Canada, provides laboratory facilities which give the students practical experience in designing and implementing automated procedures for information handling.

The automation of procedures connected with library or information centre methods has resulted in the need for trained students of computing science who have specialized in this area. Library management has often considered the implementation of automated procedures to be a technician's job and the systems design and definition as being the province of the librarian, but this division of responsibilities has sometimes resulted in unfortunate examples of library automation. Costly and inefficient file structures have been used and inappropriate programming languages chosen because the librarian did not know the implications of computer implementation and the technician did not understand the library problem or have enough advanced knowledge of computers to automate complex library procedures. A partial solution to these problems clearly lies in the training of computing science students in the design of library and information systems and in placing special emphasis on the computing techniques necessary for the efficient and advanced implementation of these designs. This type of education must be reinforced by complementary instruction for librarians to effect a complete solution.

To accomplish the computing science objective it is necessary for students to be aware of certain traditional library procedures and to be expert in automation techniques. The information storage and retrieval option offered within the Department of Computing Science at the University of Alberta attempts to provide this type of training. When the option was first developed in 1967 it was realized that the courses would remain too theoretical unless laboratory facilities were provided that offered the students practical experience in combining those traditional library techniques with the computerized handling of information. An information retrieval laboratory, therefore, was established as an integral part of this option and the Computing Science subset of the University

Library collection was loaned on a long-term basis to form the core of an information centre laboratory. The laboratory was set up to cater to the needs of third and fourth year computing science students entering the retrieval option. It was designed both to give practical experience in library automation and information centre work to students studying for degrees in computing science and to duplicate the types of problems they could expect to encounter working with the dissemination, storage, and retrieval of information.

The students must first understand the principles involved in the conventional manual handling of information before they can begin adapting these procedures for automation or designing new information handling methods. The laboratory is used to complement course instruction in traditional aspects of librarianship, such as the theory of classification and indexing, study of information sources, and identification of their characteristics. The emphasis is placed on science and technology. In acquiring some knowledge of librarianship and its terminology computing science students are better equipped to communicate with librarians; and, if the interaction is reciprocal, close cooperation will be assured.

Initially manual and computer batch procedures only were illustrated in the laboratory. However, the laboratory was designed to be used in conjunction with courses discussing on-line control of library operations with off-line batch processing used where this is more feasible. An on-line terminal was installed to allow students to gain first-hand experience in experimenting with and in implementing an on-line circulation system and an on-line catalog. For both batch and on-line use machine-readable data bases are necessary. These have been prepared; programs have been written to generate batch computer-produced book catalogs and KWIC indexes of the information centre holdings. The same data bases are available to test and improve file structures designed for efficient on-line manipulation and fast access. An on-line thesaurus linked to a classification scheme is being developed as an automated aid to classification. In addition to projects such as these, students also work with commercially available tapes to examine format structure and to develop search strategies.

It is realized that knowledge of and improvement in computer operating systems and support will have to advance further before widespread fully automated on-line library or information systems are economically feasible or operationally possible. Yet it is also acknowledged that the more students are exposed to the problems connected with the traditional manual handling of information and to the advanced on-line manipulation of information the better prepared they are to cope with the problems that arise in converting from manual to automated systems or in combining both. These dual requirements were taken into consideration when the laboratory was designed. With the computing science students in mind, it serves both to give practical experience in the pertinent aspects of librarianship necessary for information centre work and to provide a

testing ground for new ideas in automated and manual methods of information storage and retrieval.

Since the information retrieval option was established in 1967 both the courses and the laboratory have developed. The enrolment in the option has risen from 6 to 63 in four years. Thirteen of these students are registered in post-graduate programs; 10 are masters candidates and 3 doctoral candidates. On graduation the students are immediately employable either in specialized information centres doing current awareness or retrospective data base searching or in conventional libraries that have circulation or similar problems of a magnitude that requires computer assistance.

The key courses are CMPUT 560 - Information Storage and Retrieval, CMPUT 670 - Coding and Storage in Information Retrieval, and CMPUT 671 - Classification Problems in Information Retrieval. During the past year the 50 undergraduate students used the laboratory for conventional literature searching of indexes and abstracts, performed assignments in indexing and classification, used a manual coordinate index, wrote KWIC programs, and developed searching programs using boolean and weighted search logic for a computing science literature collection, the data base of which they had helped prepare using the journals available in the laboratory. Commercial machine-readable data bases, such as MARC, the AIP/SPINO, and Clearinghouse tapes, are also available for testing both for automated indexing and automated classification. The graduate students also use the laboratory for research in fields related to the development of a total integrated information system.

The on-line terminal, installed as part of the laboratory equipment, is an IBM 2741. It is connected to the University of Alberta's IBM 360/67 computer. All students taking the information retrieval option have sessions at the terminal which demonstrate the current capabilities of the projects which are being developed by the information retrieval group. Through the terminal, students are able to investigate query languages and to gain experience in what information is necessary for storage, accessibility, and retrieval in an on-line situation. It is planned to install an IBM 2260 CRT display device later this year. The CRT display will be used as an aid in developing design specifications and programs that will result in a proposal for a special-purpose terminal specifically planned for integrated on-line library automation.

The laboratory's library collection is made up of monographs, journals, reports and theses, manufacturers' manuals, and computer tapes. It now numbers 1700 monographs, 150 journal subscriptions, 500 reports and theses, 350 manuals, and tapes of 15 data bases. These may be accessed variously through an on-line catalog, computer-produced book catalogs, a KWIC index, manual coordinate indexes, or on-line and batch search profiles. Because it houses the Computing Science library collection, the laboratory also serves all computing science faculty and

students in addition to fulfilling its primary purpose as a basic element of the information retrieval teaching and research program. An information specialist is employed to oversee and coordinate aspects of various research projects, to interact with the students and teaching specialists, and to serve in a supervisory capacity for the day-to-day operation of the laboratory.

Several of the research projects will be described in some detail to illustrate the scope of the laboratory. One of these is the design of an automated on-line total library system, encompassing acquisitions, catalog, and circulation. The monograph records in the laboratory are maintained in machine-readable form and make up the data base used for the project. These records at present contain most of the information found on the LC cards except for subject headings, which will be added as the system progresses; provision has been made for the inclusion of acquisition details such as price and ordering information.

A real-time circulation system has been designed, tested over an eight-month period in a real-life situation, and modified for full-scale implementation in a special purpose library of about 10,000 titles with the present computer operating system. Later this year the circulation system will be used regularly in the laboratory for controlling the Computing Science monograph collection. The catalog may be searched on-line for entries that contain a specific author, author/title, or title words. The first two types of searches are very fast and efficient. A doctoral student is concerned with the development of more efficient file structures that, among other things, will make all three types of catalog searches economically feasible for a large academic library. Students have also begun to consider the problems of designing an on-line acquisitions system and integrating it with the catalog and circulation systems; this system is still in the design stage. All orders at present are channelled through the central University Library,which also allocates the book funds and controls the cataloging.

The data base used in developing the automated library system is stored both on disk and on tape. The disk files are updated either on-line or by off-line batch as necessary. At present the tape is updated monthly and is used to generate off-line searching tools, such as computer-produced book catalogs ordered both by author and shelf list number, and a KWIC index.

As stated, the laboratory accepts the basic bibliographic descriptions and classification for its monograph collection as assigned by the central University Library. At the same time, however, a group is investigating classification aided by large classified data bases and an on-line thesaurus. This thesaurus was developed initially for teaching and illustration purposes but, now linked to a classification scheme, is also to be used with a fully integrated special-purpose information system. The integrated system is to be operational in a governmental/industrial

environment.

Programs have been written to link any thesaurus by on-line computer to the UDC classification scheme. Information storage and retrieval of a classified data base can be handled making use of the thesaurus and classification scheme linkage. The system operates in both on-line and batch modes. The thesaurus/classification programs, as stated, are concerned with the development of a fully integrated storage and retrieval system. This system will be used initially to assist in the management of water resources. Indexing and searching will normally be carried out at the terminal through access to the thesaurus linked to a MARC-like data base, with UDC numbers in one of the fields. In the classification investigation associated with this research various machine-readable data bases were manipulated to test and compare the appropriateness of the UDC, the LC, and the United States Water Resources Thesaurus as indexing languages for the control of water resources planning literature.

The Computing Science Department information retrieval group has developed efficient current awareness and retrospective search programs for large data bases. These programs were designed initially for searching Chemical Titles tapes in cooperation with the Alberta Information Retrieval Association, but are now applied to searching tapes such as one containing a computing science journal collection, the data base of which, as mentioned earlier, was prepared with the help of the information retrieval students.

Search, statistical analysis, and reformatting and printing programs have also been developed for the MARC tapes and students either use these to search the tapes, to develop modification to the tapes, or to organize MARC searches for representative faculty members across campus. This gives the students the opportunity to experience the difficulties of question formulation in both scientific and non-scientific fields and to write suitable user guides. The information specialist assists in the preparation of these guides and in formulating and running various profiles.

The research projects currently under development require the services of the information specialist to coordinate efforts and to maintain the various data bases since some projects are contingent upon the successful completion or upkeeping of others. The continuing development of the overall plan to study certain important aspects of information storage and retrieval is very dependent upon continuity of research. To this end programs must be documented, progress reports written, and manuals prepared. The documentation of programs and their maintenance is one of the duties of a full-time analyst attached to the staff of the Computing Science Department. The faculty specialists, the students, and the information specialist issue reports, write papers, and prepare manuals documenting the research progress and aiding the teaching function. The necessary

coordination and control of certain basic aspects of the research and teaching program can only be accomplished through the provision of a full-time information specialist.

Similarly, as has been shown, the laboratory is an essential unit of the information retrieval option. The practical experience acquired in using its facilities makes computing science students fully conscious of the difficulties of identification of information content and makes them familiar with the problems of indexing and classification, both manual and automated. This training also includes the organizational aspects of library and information systems design, such as knowledge of the funds and the organization necessary for the efficient computerized handling of information, and an understanding of how these requirements compare with traditional needs. Higher costs in one sphere of the operations must be counterbalanced by reductions in another, or must be justified by increased benefits to the user community. To achieve these objectives the students must be thoroughly competent in the computing techniques necessary for implementing advanced and complex information systems.

In addition to furnishing a training ground for the practical applications of computing techniques in information handling, the laboratory also serves to expose the students to the problems of human interaction. The real-life environment that it provides teaches the students the need for close cooperation and understanding between librarian and computing scientist. It also gives them first-hand knowledge of user demands and of user reaction to automation. By interacting with the users the students learn what services the user expects the information centre or library to provide and these can then be incorporated into the total system design. The laboratory, therefore, performs the task of training the computing scientist in the essential facets of information handling in a practical environment that mirrors the technical and human problems to be overcome when successfully implementing and maintaining automated library or information centre procedures.

The author wishes to thank Professor Doreen Heaps and Mr. James Dimsdale for their assistance and suggestions. The Department of Computing Science will furnish upon request a list of publications by members of the information retrieval group.

Operations Research in the Information Science

by

Dr. F. LEIMKUHLER

OPERATIONS RESEARCH IN THE INFORMATION SCIENCE

Dr. Ferdinand Leimkuhler
Purdue University
Lafayette, Indiana

ABSTRACT

To contrast the two fields of operations research and Information Science with respect to their history and development. The nature of operations research methods and approval to problem solving is described. The function of models is analysed, the development of a model being contrasted with its formal presentation. Criteria for good models are suggested. Attention is then focused on the application of operations research techniques to problems in the analysis, design, and management of information systems. Examples of the research carried out under the direction of the author are given.

Business Meeting

Unfortunately the business reports were not in on time to be included in Proceedings and we will attempt to make them available as a separate reprint to all attendants of the Annual Meeting.

The Evolution of a Storage and Retrieval System for Indexed and Annotated Bibliographic References

by

K.E. MARSHALL

THE EVOLUTION OF A STORAGE AND RETRIEVAL SYSTEM FOR INDEXED AND ANNOTATED BIBLIOGRAPHIC REFERENCES

K. E. Marshall
Freshwater Institute
Fisheries Research Board of Canada
Winnipeg, Manitoba

ABSTRACT

The way in which the storage and retrieval system using the University of Manitoba IBM 360/65 computer has developed is outlined. Some of the limitations of the present programs are discussed and a possible line of future development is suggested.

INTRODUCTION

Almost all research scientists maintain files of indexed and annotated references to papers which they consider to be potentially relevant to their line of research. These files are organized (or in some cases disorganized) in different ways. The file may simply be a collection of reprints and photocopies arranged in a subject sequence. The commonest form that the file takes is that of a card file. The complexity of the organization of the card file may range from a simple alphabetical author index to complex classified sequences with multiple entries for references dealing with more than one subject. These classified schemes are usually based on the users own concept of subject terms relevant to his own interests.

With any system using plain cards in a classified or subject index more than one card (or a cross reference) has to be made unless the reference fits neatly into a single category. As time passes the file becomes larger and the problem of filing new material into the system becomes greater.

Quite a number of scientists are overcoming the problem of having to make multiple copies of their entries by using edge-notched cards. These are very satisfactory until the file size reaches the 2000 mark and then searching becomes tedious as many passes with the sorting needles have to be made to find the required items every time a search is made.

Very few scientists, in my experience, make use of unit-entry or peek-a-boo cards. The former, in particular, are a cheap and efficient way of retrieving relevant references.

More scientists are aware that the computer can be used as a tool for retrieving references from a file which is maintained in a

machine readable form.

At the Freshwater Institute we were fortunate in having on staff a statistician who developed a computerized storage and retrieval system primarily for his own reference file using the University of Manitoba IBM 360/65 computer. The initial scheme was modified and gradually evolved to have more general applications and it is this evolution that I propose to outline.

INPUT

The exact format for input was decided upon after a study of the type of information that potential users might wish to store. The type of format used is illustrated in Fig. 1, which shows a typical reference which has been split into the different sections which the computer is instructed to recognize. The actual input is made using standard IBM 80 column punched cards. The references are input and stored sequentially on tape in the order that they are input. Additional references can be added using the same program and these are added to the tape following the last reference already input.

SEARCHING

The initial programs made by our statistician were for on-line searching, using the remote terminal in our Institute. The first program searched for specific index terms, author, date of publication and/or journal. Any one of these characters, singly or in combination, could be used to retrieve references on file. For example: a request could be made for references by SOAP,J published in 1967 and 1968; or papers indexed under the heading RESPIRATION which were published in J.EXP.BIOL. The references requested would be printed out in the sequence that they occurred on the tape.

A separate program was made to enable searches to be made of the titles of papers on file. It was felt that this might have applications if the index terms chosen by the user did not prove adequate for any particular reason.

COMMENT

It rapidly became clear that this on-line system did not attract any users other than the originator of the scheme. We have only one terminal in our building and it may be in use just when a potential user may wish to make a search. Our statistician had his office next door to the terminal but other users would not be so

FIG. 1. Reference format as input on IBM 80 column cards

```
2, 'FACTOR' 'TIMESERIES' 'ANDERSON,T.W.' '1963' 'THE USE OF FACTOR ANALYSIS IN T
HE STATISTICAL ANALYSIS OF MULTIPLE TIME SERIES' 'PSYCHOMETRIKA' 'V.28' '' 'P.1-
25' 'THIS PAPER DISCUSSES SOME DIFFICULTIES IN APPLYING FACTOR ANALYSIS TO MULTI
PLE TIME SERIES AND ATTEMPTS TO INDICATE TO WHAT EXTENT SUCH METHODS CAN ACCOMPL
ISH THE GOALS OF THE TIME SERIES ANALYSIS.  SOME OTHER METHODS ARE SUGGESTED.'
```

Notes: The initial number indicates the number of index terms used - two in this example. These must be followed in sequence by strings of characters between '', representing in sequence author, date of publication, title of paper, title of journal or other source, volume number, part number, page numbers, and, abstract or other annotation. If any of the above items are lacking a null string '' must be inserted - part number is lacking in above example. The end of the abstract (i.e. end of reference) is indicated to the computer by two blank spaces, it is therefore essential that two consecutive blanks are not left in any other place during key punching. Each string is limited in size: e.g. 200 spaces for title, 10 for volume etc. If these are not suited to individual needs minor changes in the programs would allow for this type of modification.

favourably placed. Following discussions with interested scientists it seemed that if a printout of the reference file was available for desk use this would be attractive. This printout would, in effect, replace the old card file and could be updated as necessary. The form that the printed listings should take was discussed and when a Career Development Student became available she was given the task of making programs to produce these listings.

FURTHER DEVELOPMENT

Programs were made which instructed the computer to sort out the references already on tape in the following ways:

1. A complete listing of the whole file arranged alphabetically by author.

2. A listing of all references indexed under specific index headings (this listing to be sorted into alphabetical order of index terms and under each heading alphabetically by author).

3. A subject index to list No. 1. (i.e. an alphabetical list of the index terms with, under each index term, a list giving only author and date of publication of references on file indexed under that heading).

Examples of the format of these listings are given in Fig. 2 and Fig. 3.

Details of the programs which produce these listings are given in Fisheries Research Board Technical Report No. 209. I would draw your attention to the fact that the listings which can be made are in effect bibliographies. Using listing 2 or 1 and 3 together we have a ready means of printing out a bibliography. Thus, given an input of checked, carefully checked, suitably indexed references, with or without annotations, with no further human intervention and consequent clerical errors, it is possible to have a printout suitable for distribution. Those of us who have had the job of preparing bibliographies know that sometimes almost as much time is spent checking the typist's copy for errors as was spent in searching for the original references.

POSSIBLE FUTURE DEVELOPMENT

The present arrangements for adding new references to the file are very simple and it is not easy to add additional information to a reference already on file. Thus, if a user adds a reference to his file without any annotation and subsequently wishes to insert an abstract, this cannot be done unless you return to the original deck of punched cards used to make the tape, insert the new information in the appropriate place and input the whole amended deck of cards, discarding

FIG. 2. Format of printout

A. Alphabetical Listing:

BERGLUND,F. AND NILSSON,G. 1967
MERCURY IN FOOD PRODUCTS (2): METHYL MERCURY POISONING FROM EATING GAME BIRDS?

LAKARTIDNINGEN V.64 NO.37 P.3621-3627
FISHERIES RESEARCH BOARD, TRANSLATION SERIES, NO.1360. AVAILABLE FROM FRESHWATER INST
ITUTE, WINNIPEG

BERGLUND,F. AND WRETLIND,A. 1967
TOXICOLOGICAL EVALUATION OF MERCURY IN SWEDISH FISH

VAR FODA NO.1 P.9-11
TRANSLATED BY U.S. BUREAU OF COMMERCIAL FISHERIES

BERLIN,M. AND ULLBERG,S. 1963
ACCUMULATION AND RETENTION OF MERCURY IN THE MOUSE

ARCH ENVIRONM HEALTH V.6 P.589-

(NOTE: If the title of the reference takes more than one line the blank line illustrated above will be used and a further blank line will not be left.)

FIG. 3. Format of printout

B. Subject Index:

LIVER	
WESTOO,G.	1967
MEAT	
WESTOO,G.	1967
MERCURY	
ABERG,B., EKMAN,L., FALK,R., ET AL	1969
AKITATAKE,T.	1968
ALEKSEEVA,A.A.	1969
BERG,W., JOHNELS,A., ET AL	1966
BERGLUND,F. AND NILSSON,G.	1967
BERGLUND,F. AND WRETLIND,A.	1967
BERLIN,M. AND ULLBERG,S.	1963

the old tape.

I would suggest that a more sophisticated arrangement such as the following will offer many advantages:

1. Input unsorted references on to Tape A.

2. Sort this tape into the sequence meeting the requirements of the user (say alphabetical by author) and store in this form on Tape B making a printout for desk use if required.

3. Additional references are input on to Tape C.

4. Sort Tape C into the same sequence as Tape B.

5. Merge Tapes B and C to produce Tape D from which an updated printout can be made.

Provision could be made at the merge stage to allow the replacement of incomplete references by more complete data and to allow the removal of an unwanted entry.

These modifications would, I think, overcome the more obvious shortcomings of the present scheme.

Our simple scheme has attracted a certain amount of attention and I have supplied copies of our program card decks to three organizations in recent months but I have not heard to what extent they have been able to make use of them as yet.

REFERENCES

LEE, P.J., HANSON, J.A. and MARSHALL, K.E., 1971, "Storage and retrieval of indexed and annotated bibliographic references," Fisheries Research Board of Canada Technical Report No. 209, 43pp. (This gives full details of our programs.)

Details of more sophisticated systems are to be found in:

BURTON, Hilary D., RUSSELL, Robert M. and YERKE, Theodor B., 1969, "FAMULUS: a computer-based system for augmenting personal documentation efforts," U.S. Department of Agriculture, Forest Service, Research Note PSW-193, 5pp. (Fuller details in PSW Computer Services Library Operating Guide No. 3, 1968.)

NAMKOONG, Gene and GRAHAM, John, 1970, "A machine storage and retrieval system for scientific literature in genetics research laboratories," (Un-numbered report issued by North Carolina State University, Institute of Statistics) 80pp.

Retrospective Search Services at the Whiteshell Nuclear Research Establishment

by

H.B. NICHOLLS

RETROSPECTIVE SEARCH SERVICES AT THE WHITESHELL NUCLEAR RESEARCH ESTABLISHMENT

H.B. Nicholls
Atomic Energy of Canada Limited
Whiteshell Nuclear Research Establishment
Pinawa, Manitoba

ABSTRACT

The development and operation of a retrospective search service for scientists and engineers at a new nuclear research institute is described. Statistics are presented showing the growth of the service, subject content and type of questions received, sources of information consulted to answer the requests, form of the answers, time taken by information officers on this type of work, and utilization of the service.

INTRODUCTION

The information services of a research establishment must be prepared to undertake work which shows a remarkable diversity in both subject content and type of information needed. The scientist or engineer, when faced with a specific problem, often finds that his fund of knowledge is not adequate, and therefore needs to find data, a technique, a process, a method, a theory, to aid his solution. Occasionally he requires to make a thorough survey of a subject that is new to him, and needs to extract from the huge mass of documentation a high proportion of the available information on this subject. This paper provides statistics on the development and operation of a technical information service which meets these kinds of requests from scientists and engineers in the nuclear field. The service was developed at a new research laboratory when it was growing and new programs were being introduced.

SETTING THE SCENE

The Whiteshell Nuclear Research Establishment (WNRE) of Atomic Energy of Canada Limited (AECL) is situated some 75 miles from Winnipeg, Manitoba, and was brought into being in 1963. The other major Canadian nuclear research establishment is AECL's Chalk River Nuclear Laboratories, Ontario, which has been in operation much longer than WNRE.

AECL is a Crown company which is responsible for research into and development of peaceful uses of atomic energy. In particular it is concerned with the development of nuclear power systems that will help meet near- and long-term Canadian needs for low cost energy and the

extension and improvement of the uses of radiation and radioisotopes.

WNRE has as its major research facility the WR-1 nuclear reactor which is unique in that it is cooled by an organic, oil-like, liquid. Priority in the research programs of the establishment, which involve a variety of scientific disciplines, is given to serving the current- and near-term needs of nuclear power, but fundamental research in such fields as medical biophysics and radiation chemistry is also undertaken. An important part of the program is the science and engineering of high-temperature nuclear materials.

The Information Services Branch has responsibility for a variety of functions including information retrieval (the subject of this paper), library services, registry services, and report editing. The organization and growth of WNRE Information Services, as well as the Chalk River Technical Information Services, has been described in a recent paper by Williams, Brandreth and Baines (1970).

INFORMATION RETRIEVAL AT WNRE

The information retrieval service at WNRE really got underway early in 1968. From the outset priority was given to retrospective literature searches, and a group of technical information officers with scientific or engineering training and wide interests was gradually built up to do this work. An equal balance has been maintained between people with previous experience in research or industry, including two with experience in information work, and graduates straight from university. There are several chemists, a physicist, a mechanical engineer, a mining engineer, and an electrical power engineer. In addition, during the summer months two or three university students are employed on this type of work. A current-awareness service (not covered in this paper) is operated which complements the retrospective search service. All information retrieval staff are involved in both services and, in addition, the majority have other duties such as administering research contracts and acting as secretaries for technical committees. This "other" work provides information retrieval staff with variety and stimulation; it also enables them to keep informed of some of the work of the research staff. Their offices are located in the Scientific Information Centre which also houses the library. The Centre is situated between the research and development building and the main research tool, the reactor, and is physically linked to both. This central location facilitates good communication between information and research staff, a factor which has contributed towards the successful operation of the service.

The success of a technical information service of the type described in this paper depends to a very large extent on the quality of the library it uses. At WNRE the library is considered to be one of the major research tools. It has been given sufficient priority and

budgetary support to allow for the rapid accumulation of a comprehensive collection of books and journals. Currently it has a stock of some 23,000 bound volumes and 125,000 reports, and subscribes to 750 different journal titles.

Vickery (1970) states that the purpose of any retrieval system is to deliver output to users - ideally, just the information they need, in the form and at the time they need it. This has been the philosophy at WNRE. Each answer is an individually prepared package, and there has been little standardization in the format of the output supplied. During the course of the search the information officer keeps in close touch with the "customer", which ensures that the results have as high a degree of relevance as possible.

The information retrieval group at WNRE do not maintain their own catalogues and indexes, with the exception of an index of searches undertaken in the section, but rely on available tools such as abstract journals, handbooks, encyclopaedias, standard treatises and monographs. The nuclear field is served particularly well in this respect with well established abstract journals such as Nuclear Science Abstracts. Some small use has been made of computerized information systems run by outside organizations such as the Euratom Nuclear Documentation System (ENDS) as described in a paper by Rolling (1966), but it has been found that carrying out retrospective searches on a computer "by letter" is not without its trials! Direct access to the computer appears to be virtually essential for retrospective searching. Ziman (1969) in an address to the British Association makes the observation that "the conventional system (the abstract journal) is not without its resources against this suffering foe (the information explosion)", and the experiences at WNRE support this.

STATISTICS

The following statistics relate to searches requiring at least 0.5 man-days work by an information officer. No records are maintained for jobs taking less time.

Growth of the Service

The following table illustrates the growth of the service during the three year period from 1 April 1968 to 31 March 1971. Some searches were undertaken prior to April 1968, but incomplete records are available, and the three years shown can be considered as the main growth period. Mooradian, Perryman and Kennett (1971) note that whereas academically oriented research can be carried out with each scientist supported by one or two technicians, nuclear programs of technological significance require a broad base of both technical and administrative support. For

this reason numbers of professional staff are shown in the table below in addition to the total staff. 47% of these professionals are technical- or administration-support staff.

Year	1967/68	1968/69	1969/70	1970/71
No. of searches completed	–	97	124	153
Total WNRE staff as at 31 March	678	754	780	784
Total WNRE professional staff as at 31 March	118	145	155	157
No. of technical information officers	2	5	8	8

Subject and Type of Information Requested

The following table shows the number of searches requested by main subject field during each of the years from 1968/69 - 1970/71, inclusive.

Subject field \ Year	1968/69	1969/70	1970/71	TOTAL
Chemistry	33	41	52	126 (34%)
Engineering & technology	23	24	37	84 (22%)
Materials	28	26	28	82 (22%)
Life sciences & environmental	7	15	18	40 (11%)
Physics	3	4	6	13 (3%)
Economics	–	4	8	12 (3%)
Other	3	10	4	17 (5%)
TOTAL	97	124	153	374

The following table shows the type of information requested. A similar analysis was carried out by S and M Herner (1958), based on over 4000 questions put to the libraries of the United States Atomic Energy Commission (see Column 2). As far as is known the U.S. figures include "quick" questions, requiring less than 0.5 man-days to answer, which may account for some of the differences between the two sets of figures below.

Types of questions	Col. 1 WNRE %	Col. 2 USAEC %
Description of a process or method	24	20
Properties of a substance	24	20
Description of apparatus or equipment	5	14
Description of plant or systems	13	
Physical and chemical data	5	13
Biological effects	5	5
Radiation effects	9	2
Commercial, economic statistics	5	6
Information about institutions and people	2	3
Other	8	17
	100	100

Source of Information

Abstract journals are the most common tools used in the searches. Based on use by information staff during the years 1969/70 and 1970/71 they were consulted in some 78% of the searches. (Of course in some of these searches other sources were consulted as well). Details are given below. Some 40 abstract journals, including journals containing both technical articles and abstracts, are taken at the establishment.

Average Annual Use of Abstract Journals	
Journal	No. of searches in which consulted
Nuclear Science Abs.	51 - 60
Chemical Abs.	41 - 50
Engineering Index, Metals Abs.	11 - 20
Biological Abs.	6 - 10
Ceramic Abs. (J. Amer. Cer. Soc.) Physics Abs.	5
Scientific & Technical Aerospace Reports	4
Analytical Abstracts, Applied Science & Technology Index, Digest (U.K. Central Elec. Generating Bd.) Electrical and Electronics Abs., Index Medicus	3
British Ceramic Abs., British Technology Index Fuel Abs.	2
Applied Mechanics Reviews, Computer and Control Abs.,Corrosion Abs., Dissertation Abs., Gas Chromatography Abs., J. Iron & Steel Institute (Abs), Nuc. Science Abs. of Japan, Pollution Abs., Spectrochemical Abs., Solid State Abs.	1

Other major sources of information are primary journals and people. In some 15% of the searches a third person (on- or off-site) is consulted for information.

Type of Answer

A bibliography with subject index is supplied with 28% of the answers. Indexing is done either manually, using edge-notched cards or by computer (KWIC index). When a subject index is not supplied the references are usually arranged under broad subject categories. Numerical data are supplied with the answers to 12% of the jobs, and in some 25% of the jobs some form of review or survey, often quite brief, is written.

Time Involved

On an average each information officer spends 70% of his time on retrospective searches. The remainder of his time, as previously stated, is devoted to current-awareness and other duties.

The following table, based on the 277 jobs completed during 1969/70 and 1970/71 shows the wide variation in the times spent on searches by information officers i.e. excluding clerical duties such as typing and photocopying which are undertaken by support staff:

Man-days spent on search	% No. of searches
0.5 - 2	26
3 - 5	30
6 - 10	16
11 - 15	14
16 - 25	8
26+	6

The average time spent by information officers on a search was 9 days.

Utilization of the Service

Some 50% of all professional staff, with the exception of information and other administrative support staff, are currently requesting at least one search per year, either directly or through their technicians.

The figure is higher (65%) for staff engaged in research and development, i.e. excluding technical support staff such as those engaged in reactor operations, maintenance, project engineering, radiation hazards control, environmental control, and industrial safety. On an average each of the users requests two searches per year.

CONCLUSION

The statistics illustrate the development and operation of a conventional retrospective information retrieval service which is much appreciated by research and technical support staff. It is difficult to evaluate the performance of the service, but an indication of its effectiveness is demonstrated by the ability to compete successfully for establishment staff positions. There has been no fall-off in demand for searches as scientists have become established in their research programs, and it is concluded that the present level of use of the service (equivalent to one search per professional per year) can be regarded as an indicator of the minimum demand on the service in the future. A mechanized current-awareness service is to be introduced at the establishment shortly, based on the computer searching (on-site) of Nuclear Science Abstracts tapes every two weeks. The availability of this service to all staff could slightly reduce the demands on the retrospective search service. In the longer term the on-site computer searching of the International Nuclear Information System (INIS) tapes will most probably be introduced, both on a retrospective and current-awareness basis. Information retrieval staff at WNRE provide part of the Canadian input to this system which was described in a recent paper by Woolston et al. (1970).

REFERENCES

HERNER, S. and M., 1958, "Determining Requirements for Atomic Energy Information from Reference Questions", Proc. of Internat. Conf. on Sci. Inf., Washington, v. 1, p. 181-8.

MOORADIAN, A.J., PERRYMAN, E.C.W. and KENNETT, T.J., 1971, "Irradiation Facilities: Their Importance in Developing Canadian Nuclear Competence". Paper A/CONF.49/A/152, Fourth United Nations Conference on the Peaceful Uses of Atomic Energy, Geneva, IAEA, September 1971, p. 9.

ROLLING, L.N., 1966, "A Computer-Aided Information Service for Nuclear Science and Technology". Journal of Documentation, v. 22 no. 2, p. 93-115.

VICKERY, B.C., 1970, Techniques of Information Retrieval: Connecticut, Archon Books, p. 48.

WILLIAMS, G.P.L., BRANDRETH, M. and BAINES, G.O., 1970, "Nuclear Information in Canada - An Approach to a National Service". Proc.

of the Symposium on the Handling of Nuclear Information, IAEA, Vienna, STI/PUB/254, p. 9-20.

WOOLSTON, J.E., ISSAEV, L.L., IVANOV, M.V. and DEL BIGIO, G., 1970, "The Design and Implementation of an International Nuclear Information System". Proc. of the Symposium on the Handling of Nuclear Information, IAEA, Vienna, STI/PUB/254, p. 607-617.

ZIMAN, J.M., 1969, "Information, Communication, Knowledge", Nature, v. 224 (25 Oct.) p. 318-324.

A Modified KWOC System for Small Specialized Libraries

by

K.E. KOOLE

A MODIFIED KWOC SYSTEM FOR SMALL SPECIALIZED LIBRARIES

Katherine E. Koole
University of Calgary
Calgary, Alberta

ABSTRACT

The Information Services and Systems Division of the University of Calgary has modified a KWIC-KWOC system so that it can be used to generate a large variety of indexes. One of the most successful of these applications has been the creation of indexes for small very specialized reference collections.

INTRODUCTION

This system is designed to produce indexes to the reference literature that university departments and individuals collect. These collections are not large enough to warrant the hiring of a librarian or even a full-time clerk. They do, however, contain valuable information that should be retrievable.

The programs are based on a KWIC-KWOC system written by P.L. White of IBM. Originally there were four PL/1 programs and two Assembler subroutines. Only one program has been added for producing library indexes but the first and main program has undergone several major changes. All the features of the original system have been kept in the modified version.

BASIC FEATURES

The Input

All input is handled by one program. This program is controlled by switches set at execution time so that only the data sets or files that are needed are generated. The files are of three different formats depending on the type of index they are to be used to produce. Each file is sorted and passed to one of the other programs which print the actual indexes.

The modified system can use the same input format as the original programs. Switches have been added which make it possible to change some of the coding procedures, such as omitting the sequence numbers, but these

are seldom used in the production of library indexes.

See Figure 1.

The Output

The system produces four types of indexes KWIC (Keyword in Context), KWOC (Keyword out of Context), TITLE (an alphabetic listing of titles), and LISTING (a list of the card input in the order of reference number).

See Figure 2 and 3.

The KWIC index is seldom used for library indexes. The format is not popular with users and the enriched KWOC is much more useful.

The keywords for the KWOC indexes are taken from three different sources:

(1) from author or type "1" data where the keyword is simply the sixty characters of text on each card;
(2) from title or type "2" data where the keyword is extracted from the text;
(3) from the descriptor or type "4" data where the keyword is again extracted from the text.

When the keyword is extracted from text, certain rules are used to determine what is a keyword. These rules can be changed and are different for type "2" and type "4" data, but all rules work on the idea that a keyword is any series of alphabetic or numeric characters bounded by a space or punctuation symbols. Certain binders can be used to connect what would otherwise be separate keywords into one.

This system also provides a list of up to 256 stopwords. If a selected keyword matches one of these stopwords, the keyword is not used to create an index entry. This prevents such words as "THE", "AND", etc., from appearing in indexes. The system can also eliminate selected keywords that are less than three characters in length.

See Figure 4.

The system was originally designed to produce a separate physical KWOC format file for each type of KWOC keyword. If an index made up of any combination of these types was wanted, either the combination had to be permanent or a special additional file had to be made. This problem was solved by adding a field so that each KWOC entry was flagged to show the source of the keyword such as OA for author and then putting all KWOC entries on the same physical file. What the extra testing costs, in computer time, is saved in the cost of providing, only one storage

T = "1" AUTHOR

T = "2" TITLE

T = "3" PUBLISHER

T = "4" DESCRIPTORS

1 60	72	75 80
LAPP PHILIP A	RDC 637011	
UNDERGRADUATE ENGINEERING ENROLMENT PROJECTIONS FOR ONTARIO	RDC 637012	
1970-1980	RDC 637022	
SECRETARIAT OF THE COMMITTEE OF PRESIDENTS 230 BLOOR STREET	RDC 637013	
WEST TORONTO 181 1970	RDC 637023	
CUPO STATISTICS	RDC 637014	
SKOLNIK ML	RDC 638011	
MCMULLEN WF	RDC 638021	
AN ANALYSIS OF PROJECTIONS OF THE DEMAND FOR ENGINEERS IN	RDC 638012	
CANADA AND ONTARIO AND AN INQUIRY INTO SUBSTITUTION BETWEEN	RDC 638022	
ENGINEERS AND TECHNOLOGISTS	RDC 638032	
SECRETARIAT OF THE COMMITTEE OF PRESIDENTS 230 BLOOR STREET	RDC 638013	
WEST TORONTO 181 1970	RDC 638023	
CUPO STATISTICS	RDC 638014	
FARMER JAMES	RDC 639011	
AN APPROACH TO PLANNING AND MANAGEMENT SYSTEMS	RDC 639012	
IMPLEMENTATION	RDC 639022	
CALIFORNIA STATE COLLEGES 5670 WILSHIRE BLVD LOS ANGELES	RDC 639013	
CALIFORNIA 90036 1971	RDC 639023	
WICHE PMS PROGRAM	RDC 639014	

SAMPLE INPUT

FIGURE 1

AND ONTARIO AND AN	INQUIRY INTO SUBSTITUTION BETW	RDC	638
CH TO PLANNING AND	MANAGEMENT SYSTEMS IMPLEMENTAT	RDC	639
OLMENT PROJECTIONS	ONTARIO 1970 *UNDERGRADUATE EN	RDC	637
EERS IN CANADA AND	ONTARIO AND AN INQUIRY INTO SU	RDC	638
GN *AN APPROACH TO	PLANNING AND MANAGEMENT SYSTEM	RDC	639
INEERING ENROLMENT	PROJECTIONS FOR ONTARIO 1970-1	RDC	637
GISTS *ANALYSIS OF	PROJECTIONS OF THE DEMAND FOR	RDC	638

KWIC INDEX

MANAGEMENT		
AN APPROACH TO PLANNING AND MANAGEMENT SYSTEMS IMPLEMENTATION	RDC	639
ONTARIO		
UNDERGRADUATE ENGINEERING ENROLMENT PROJECTIONS FOR ONTARIO 1970-1980	RDC	637
AN ANALYSIS OF PROJECTIONS OF THE DEMAND FOR ENGINEERS IN CANADA AND ONTARIO AND AN INQUIRY INTO SUBSTITUTION BETWEEN ENGINEERS AND TECHNOLOGISTS	RDC	638
PLANNING		
AN APPROACH TO PLANNING AND MANAGEMENT SYSTEMS IMPLEMENTATION	RDC	639

KWOC INDEX

FIGURE 2

RDC 637 LAPP PHILIP A
UNDERGRAUDATE ENGINEERING ENROLMENT PROJECTIONS FOR ONTARIO 1970-1980
SECRETARIAT OF THE COMMITTEE OF PRESIDENTS 230 BLOOR STREET WEST TORONTO 181 1970
CUPO STATISTICS

RDC 638 SKOLNIK ML
MCMULLEN WF
AN ANALYSIS OF PROJECTIONS OF THE DEMAND FOR ENGINEERS IN CANADA AND ONTARIO

RDC 639 FARMER JAMES
AN APPROACH TO PLANNING AND MANAGEMENT SYSTEMS IMPLEMENTATION
CALIFORNIA STATE COLLEGES 5670 WILSHIRE BLVD LOS ANGELES CALIFORNIA 90036 1971
WICHE PMS PROGRAM

LISTING INDEX

TITLE INDEX

FIGURE 3

AN APPROACH TO ANALYSING HIGHER EDUCATION

PROJECTIONS FOR 1970-1980

ACCEPTABLE AS INDEX TERMS	NOT ACCEPTABLE
ANALYSING HIGHER EDUCATION	AN (Too Short) APPROACH (Stopword) TO (Too Short) FOR (Stopword) 1970-1980 (Keywords beginning with a numeric character are usually stopped)

EXTRACTION OF KEYWORDS

FIGURE 4

volume (usually tape) rather than three or four.

The single biggest reason for wanting to combine the KWOC format files was to produce the Enriched KWOC index. The KWOC index of title keywords operates on the belief that titles contain informative enough keywords to describe the publication. This is a great deal to expect of any title but the fact that it does not work as well as hoped does not mean that an index based on title keywords is useless. It is the quickest and easiest way of producing a subject index and is better than nothing. Then if one gives such an index a little help by adding a few well chosen descriptors, a really useful index can be the result.

A title index is an important feature of most libraries and should be no less so of small specialized subject collections. To keep files and data set descriptions to a minimum, it was decided to use the KWOC format to store the title data even though the actual index is not a KWOC index. The format has all the necessary fields.

A switch and a small subroutine were added to the first program so that KWOC entries of type "OT" for title could be created. These entries have a blank keyword and modified titles where articles("THE","AN","A") are removed from the first of the title and appended with a comma at the end. Then a sort procedure, a short PL/1 program were added to print the title index.

The LISTING index is useful as the other indexes only give the title and reference number. This is the only index that prints out such information as the publisher which is usually put on a tape "3" card. It also gives a list of all the descriptors and authors for each publication.

All files are sequential. This means that for the KWOC format data sets,the title and reference number have to be repeated for every selected keyword.

Corrections

The only provision the original system made for correcting the existing files was the ability to delete an entry. This is extremely impractical for a tape orientated system as the data cards are usually destroyed once the data is entered. This means that to add or correct a keyword, all keyword entries for that publication have to be deleted and then the data has to be re-keypunched and re-entered.

A program has been written that allows corrections or additions to be made to the keywords. It does not allow changes to be made in the actual title data that is printed in the indexes. The program is cumbersome but provides a needed feature and can be used to provide a type of "post" vocabulary control.

APPLYING THE SYSTEM

Reference Numbers

The collections are usually highly specialized making it very difficult to define standardized subject divisions. As the collection should be small enough to be in one room, there is no real need to attempt to make the publication's location dependent on its subject matter. Everything is simply assigned a location and corresponding reference number as they are added to the collection. Therefore, the reference or ID number only gives the location and has no relationship to subject.

As the collections vary from reprints to hardback reference books, there is sometimes a need to keep material in different ways such as filing small reports and reprints and shelving only hardback publications. Another possibility is such mutually exclusive subdivisions as IBM computer manuals and standard Computing Science reference books.

For these groupings the twelve-character reference number can accommodate a prefix such as "PER" for periodicals. Usually for inventory purposes and convenience, the prefix indicates a separate location and a separate numbering system.

See Figure 5.

Using Binders

The system has two types of binders for joining separate keywords into one. One is the dash which appears in the final indexes. The second is the underscore character which is removed from the data and the keyword once the keyword has been selected. This is useful for joining words such as "INFORMATION" and "SCIENCE" where a dash would appear out of place.

The use of binders requires discretion. The best rule is to let the system work by default unless one is sure they are making a worthwhile contribution with binders or descriptors.

For example, someone might want to combine the keywords "HIGHER" and "EDUCATION". Unless the combined keyword is accepted and common among users, it is better to leave them separate when the second word is a useful keyword by itself, such as "EDUCATION" is.

Assigning Descriptors

Descriptors can make a title keyword index very useful. They can be used to add keywords the title missed, create large subject divisions (STATISTICS,APPLIED MATHEMATICS), characterize a publication according to some physical characteristic (REPRINT,PERIODICAL) and finally to make an entry for a keyword that is bound to another in the title (EDUCATION

PER 129 COLLEGE & UNIVERSITY BUSINESS
MCGRAW-HILL, NEW JERSEY

PER 130 CIPS
CANADIAN INFORMATION PROCESSING SOCIETY
MONTREAL

REF 1 POLLACK SEYMOUR V
A GUIDE TO FORTRAN IV
COLUMBIA UNIVERSITY PRESS, 1965

REF 2 CHAPMAN EDWARD A
ST. PIERRE PAUL L
LUBANS JOHN JR.
LIBRARY SYSTEMS ANALYSIS GUIDELINES
JOHN WILEY & SONS, NEW YORK, 1970

REF 3 STABLEY DON H
LOGICAL PROGRAMMING WITH SYSTEM /360
JOHN WILEY & SONS, NEW YORK, 1970

REF 4 CARNAHAN BRICE
LUTHER H A
WILKES JAMES O
APPLIED NUMERICAL METHODS
JOHN WILEY & SONS, NEW YORK, 1969

USING REFERENCE NUMBERS

FIGURE 5

in HIGHER EDUCATION). Less caution is needed in using descriptors then in using binders, but misuse of descriptors can create alot of garbage out of what would have been a good index.

The best procedure has been to spend a little time with the users and arrive at a short list (20 or 30) of keywords which best describe their interests. No attempt should be made to try to enter a descriptor for every subject in a publication, just worry about the users' interests. A clerk can then take a quick look at a publication and the list and add any keywords that are not provided by the title.

The clerk should do this quickly as the most important thing is to get the publication into the index. Once it is in, the clerk can circulate the publication to certain people to see what descriptors they would like to add or the users can simply make suggestions to the clerk after they have used the publication or the index. The next time the clerk updates the index, these suggestions can be added to the existing index.

THE FUTURE

Some excellent features can be added to the system to increase its sophistication and to make it possible to handle larger collections.

(1) Add the optional use of a sophisticated thesaurus.
(2) Add the ability to create and use random access and cross reference files. The cost of such files is justified when the size of the collection or the number of index terms per publication reach a certain level. These files would make it possible to print other data elements than just the keyword, title and reference number in KWOC indexes.
(3) Finally, the ability to use on-line terminals for entering, correcting and retrieving indexes.

The system, as it is, can cheaply and quickly produce author, subject and title indexes to a small reference collection. It can vary in sophistication and usefulness according to the time and effort that a user wants to put into it, but even with the minimum of work, useful indexes are possible.

REFERENCES

Reports

IBM,1962, Keyword In Context (KWIC) Indexing, Form No. E20-8091-0.

KOOLE, K.E., 1971, OIR Library Indexes, Data Centre, University of Calgary, Alberta.

WHITE, P.L., 1968, KWIC-KWOC For 360 (360D 06.7.014), IBM (U.K.), Wilmslow, Cheshire, U.K.

Classifying, Indexing, and Searching Resource Management Information Via an On-Line Thesaurus

by

F. ALBER AND D. HEAPS

CLASSIFYING, INDEXING, AND SEARCHING RESOURCE MANAGEMENT INFORMATION VIA AN ON-LINE THESAURUS

F. Alber and Doreen Heaps
Department of Computing Science
University of Alberta
Edmonton, Alberta

ABSTRACT

Information systems of the future should offer facilities to access data bases from many entry points and to allow for on-line indexing, classification and searching. A project which uses an on-line thesaurus as a central tool for classifying, indexing and searching is described. A natural resource data base and thesaurus form the test environment.

INTRODUCTION

The basic purpose of any information system is to place the user in more efficient and direct contact with the data bases of concern to him, and thus to enable him to make more efficient decisions. Various tools have been developed to improve efficiency both at input and output. The techniques of coordinate indexing, the employment of classification schemes, the development of thesauri, the batching of computer profiles, and on-line query languages are all aids of this kind. In general, however, one type of tool has been used with one kind of data base and with one form of computer or library application. Similarly, much of the literature on the improvement of relevance and recall deals with attempts to increase efficiency by deciding on which particular technique is the best and with measurement and comparison of efficiencies achieved by one or the other of various methods.

If the user's interests range widely then such restrictions limit his access to information. Some information needs are satisfied only by access to a non-homogeneous data base or to several linked data bases. Some material should be indexed and some classified, some accessed immediately, some available, with delay, in off-line batch. This diversity of needs and material is most evident in interdisciplinary fields, such as those that deal with urban dwelling problems, control of pollution, or conservation of natural resources.

There have been limited attempts to combine techniques, one of the best known being the Thesaurofacet of the English Electric Company. The project described here, which combines techniques, uses concepts similar to those of the Thesaurofacet. It was, however, developed

independently and is basically concerned with interactive computer applications.

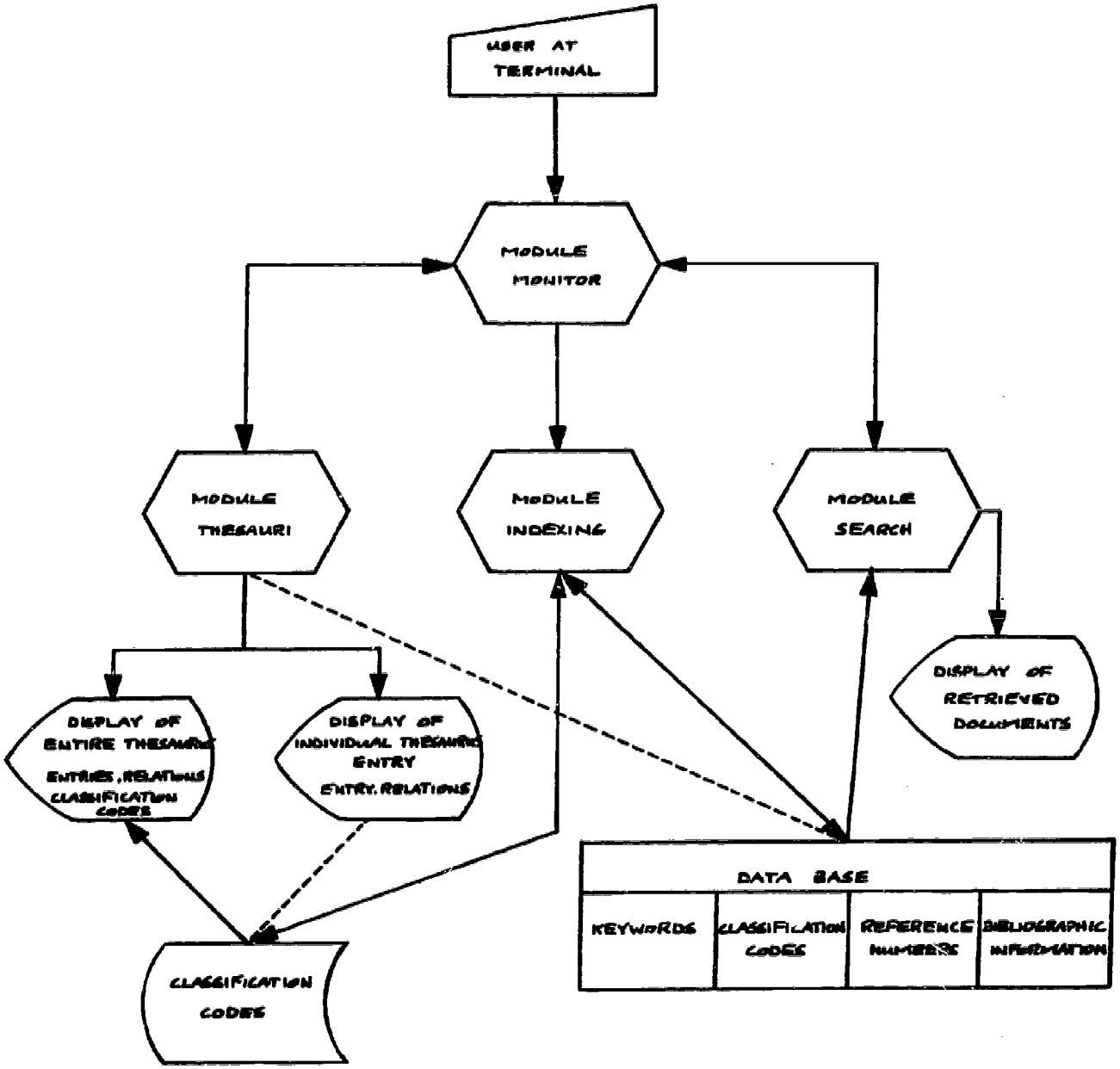
The project focuses on some of the problems of natural resource information management, of particular interest to Canadians; these involve the classification, indexing and searching of material from a wide variety of sources and a wide variety of fields. This paper deals with the development of an on-line thesaurus as a primary aid in classifying, indexing, and searching a specific water resource data base. Users of this data base are persons responsible for water resource management decisions.

The data base contains material in bibliographic format of non-standard type. The documents include research project descriptions, research grant applications, monographs, journal articles, abstracts of statutes, entire statutes, and so forth. In general the material is of the type necessary for administrative rather than research decisions. The documents are indexed and/or classified by L.C., Dewey, or U.D.C. The computer record includes, therefore, standard bibliographic elements, such as author, title, publication data, and is augmented by keywords, either accession or location number, or classification numbers or both. For the purpose of this study it is assumed that both keywords and class numbers are available on every document surrogate. At present no abstracts or continuous text are included although these could be.

This material will be accessed and controlled and new documents indexed and/or classified through an on-line thesaurus. The searcher or the indexer sits at a terminal and uses the thesaurus as the initial entry point (Fig. 1). The system has a thesaurus, a data base (document surrogates), and a class structure (schedules). All of these exist in the computer. The thesaurus is the heart of the system.

DESIGN CONSIDERATIONS

In the design of a computer program to facilitate on-line manipulation of the thesaurus many problems required careful consideration. Attention must be given to the complexity of the query language that accompanies the man-machine interaction in such an on-line project. How much information must the computer convey to the user before expecting a correct response from him? Must the user be prompted before each reply? It is reasonable to assume that after the user becomes familiar with the requirements of the program regarding command specification, the query language should be able to undergo vast simplification without endangering the correct functioning of the man-machine interaction involved in accessing and modifying the given thesaurus.



Thesaurus-Centered System

Figure 1

ON-LINE CLASSIFYING AND SEARCHING

The thesaurus must somehow be linked to the data base of mixed bibliographic format to facilitate both the indexing of new data base entries and the searching of existing data base entries for information. The thesaurus must also be linked to classification codes and schedules in order to allow for searching of the data base with class numbers as well as keywords. The class numbers will be linked to schedules. In regard to the thesaurus itself questions arise about what relationships should be permitted and how many entries should be allowed for a specified relationship under a given entry.

As stated previously the programming system is centered on a computer program which allows a user to create, modify, or display the thesaurus or parts thereof. The key requirements for the on-line thesaurus program are the following: (1) fast access to terms and associated relationships; (2) an unlimited number of relationships of a certain type associated with a particular term (within the limits of computer memory); (3) generality, in that a user is not tied to specific relationships; and (4) brevity of actual program so as to allow as much computer memory as possible to be devoted to the required accounting and storage of data.

DESIGN IMPLEMENTATION

To obtain fast access to thesaurus entries and their associated relationships a binary search was adopted. Searches for terms are conducted on a variable length list using a fixed key size. Two bytes are used for the address of the term in a table containing all the thesaurus entries, two bytes for the length of the associated term, and two bytes for a pointer to a disk record containing relationship information for the term (Fig. 5). The storage area for the keys is appropriately called TERMLIST while the storage area for the actual thesaurus entries is called TERMSTOR.

The key to the thesaurus program is the file structure associated with the handling of relationships for a given thesaurus entry. The file structure employed efficiently handles accounting for the thesaurus entries and their relationships and yet is on a general enough basis so that with very minor changes a different thesaurus with totally different relationships can be handled.

An article written by E. H. Sussenguth Jr. (1963) provided major leads for the choice of file structure to handle the relationship entries. G. Salton (1968) adopted Sussenguth's techniques without modification for storing thesaurus data in his SMART system. In the program developed at the University of Alberta Sussenguth's theory was adopted and then significantly modified in its application.

In his paper, Sussenguth has, for the most part, adopted Iverson's definitions. The most important definition is of what he calls a 'filial set'. He states: 'The set of nodes which lie at the end of a path of length one from node x comprises the filial set of node x and x is the parent node of that set'. Figure 2 should convey the meaning of most of the terms that are defined.

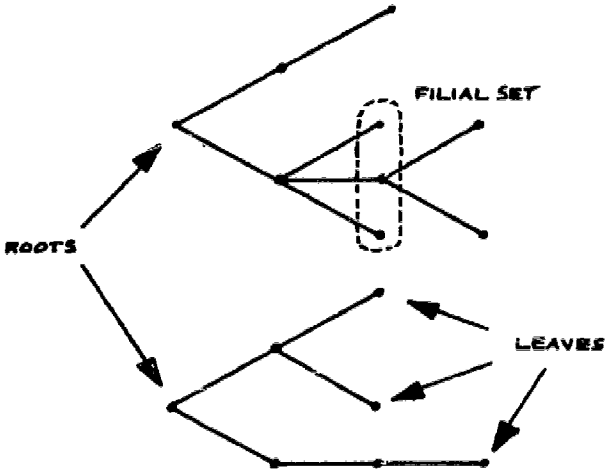


Figure 2 (after Sussenguth)

He states further that for any root the associated filial set is made up of nodes which are second key elements used with the key element. For example, with English words as keys the nodes might correspond to letters. Then the filial set of the letter B would be all the letter which can be used with B to start an English word. Figure 3 illustrates this concept.

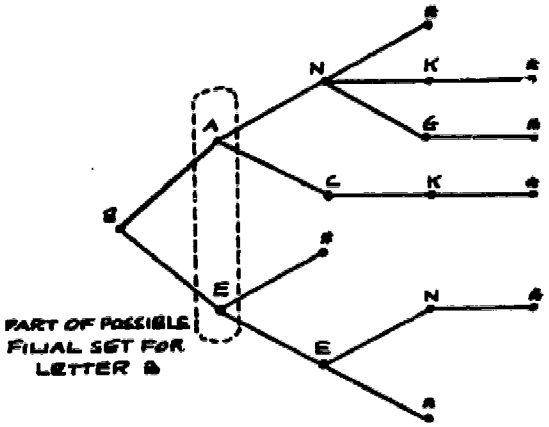


Figure 3 (after Sussenguth)

ON-LINE CLASSIFYING AND SEARCHING

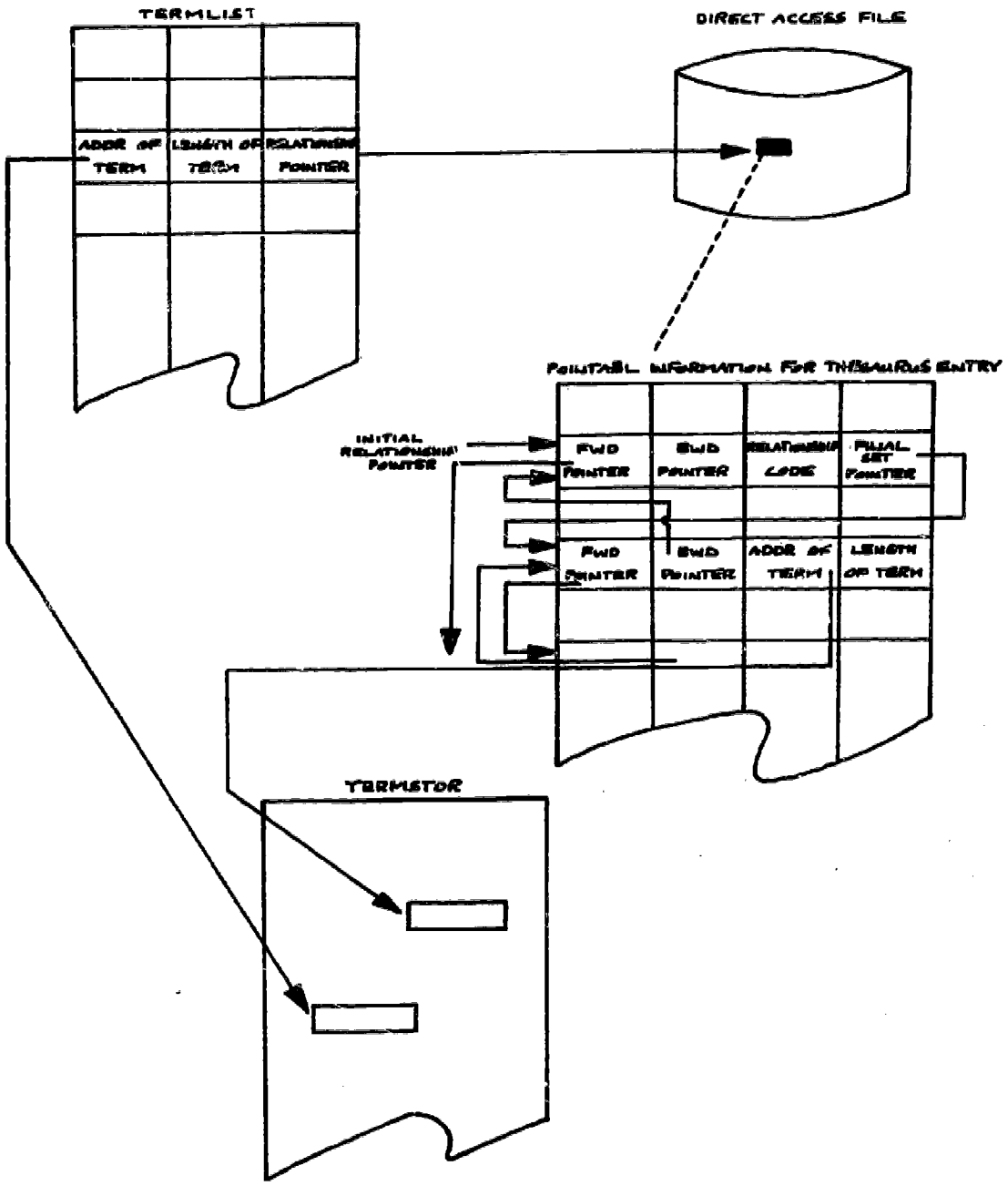
In using a computer to represent a tree structure each node might be chained to its filial set and the nodes within a filial set might be chained together. Figure 4 gives a possible memory configuration for the tree of Figure 3.

1	B		2
2	A	12	3
3	N	5	4
4	*	8	
5	C		6
6	K		7
7	*		
8	K	10	9
9	*		

10	G		11
11	*		
12	E		13
13	*	14	
14	E		15
15	N	17	16
16	*		
17	*		

Figure 4 (after Sussenguth)

In the thesaurus program developed at the University of Alberta Sussenguth's basic idea of chaining each node to its filial set and having the nodes within a filial set chained together is retained as a step in the program. However, rather than having each key value correspond to an English letter as Salton does in his SMART system and Sussenguth does in his examples, one of the entries of the key value corresponds to a code which indicates the relationships that other thesaurus entries have to the thesaurus entry or TERM being considered. The filial set of each key value is made up of pointers to the entries which bear the relationship to the main thesaurus entry that is indicated by the code. For example, assume that CODE3 indicates the relationship 'related term' and that TERM4, TERM5, and TERM6 are related to TERM1. Then in the relationship entries for TERM1 a key is set up containing among other things CODE3. The filial set for this entry key will contain pointers (i.e. address of entry and length of entry) to TERM4, TERM5, and TERM6. The table containing all of the relationship information for a particular entry is called POINTABL. The disk address of the relationship information for a term is contained in the 'relationship pointer' mentioned previously (Fig. 5).



File Structure

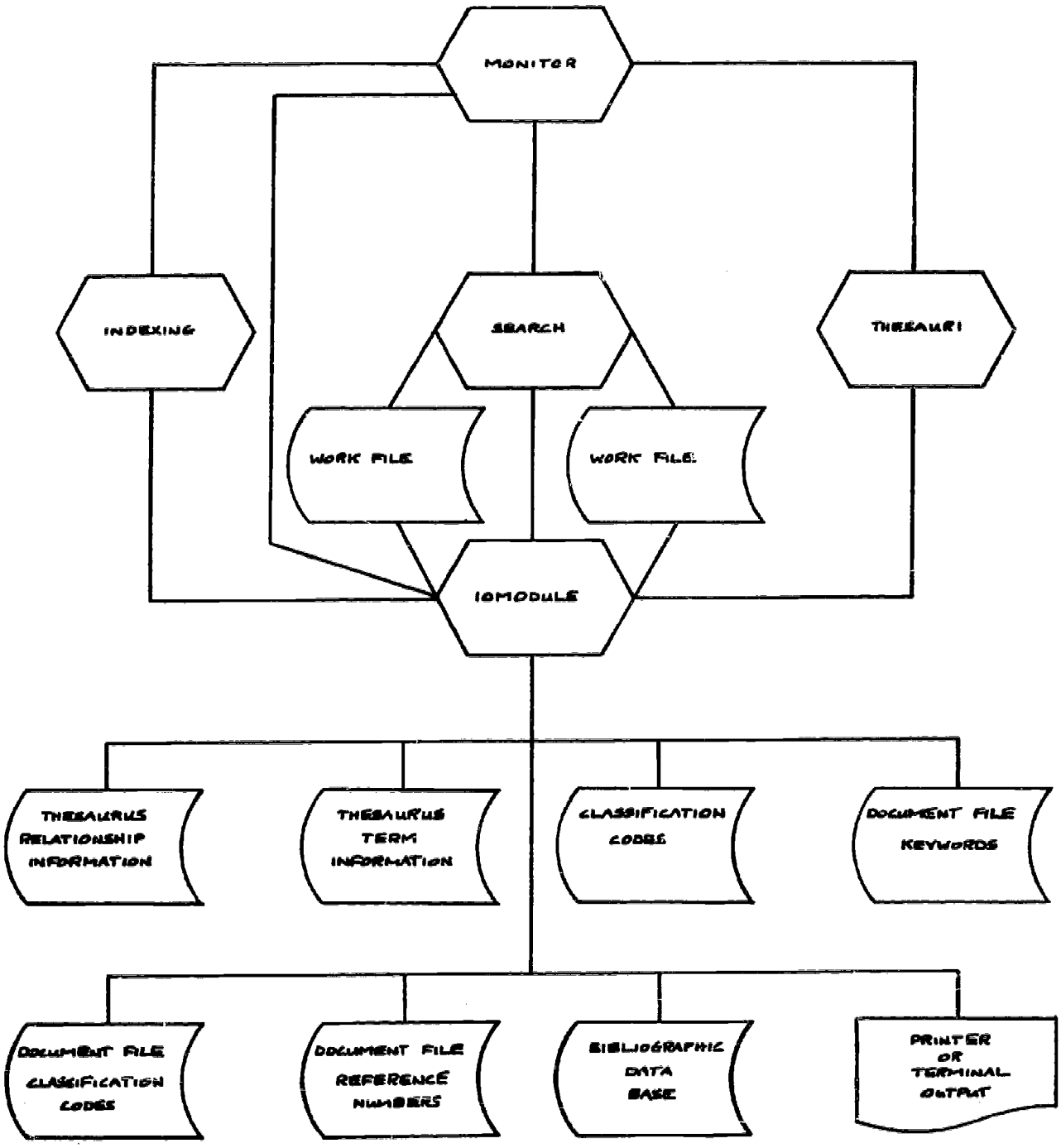
Figure 5

ON-LINE CLASSIFYING AND SEARCHING

A table in the program contains entries which represent the meanings of the codes explained above. For example, if CODE2 indicates the relationship 'broader term' the corresponding table entry might be 'BT'. This facility allows the user to change the meaning that the codes have by altering the table entries. The user can enter as many entries for a specified relationship as he wishes. Reciprocal relationship entries are made automatically by the computer program. By altering a translate table the automatic entering of reciprocal relationships can be suppressed or initiated depending on the user's wants.

Access to classification markers (U.D.C., L.C., etc.) is provided by assigning numbers to entries in the thesaurus that are used in indexing and searching and a link is thus established with the data base containing classification numbers corresponding to the thesaurus entries. Thus, if the number 10 is assigned to a particular thesaurus entry, record number 10 in the classed data base contains classification codes which correspond to this entry. In indexing and searching the numbers assigned to the thesaurus entries can be used both as indexing entries and search terms. In indexing or searching via the classed data base the number assigned to the thesaurus entry serves as the index into the classed data base. The corresponding classification numbers may be used both to class incoming documents or prepare search questions, whatever the case may be. Programs have been written that allow a user to update or retrieve from the classed data base, index incoming documents via keywording, classification, or numbering, and search via keywords, classification codes, or the numbers automatically assigned to terms used in indexing and searching (Fig. 6). If the system were to operate with a very large data base a different searching scheme than is presently employed would be more efficient. At the University of Alberta search techniques that operate with compressed data have been developed and implemented for efficiently and economically searching large data bases of this type (Heaps and Thiel (1970)).

The thesaurus program and all other programs associated with the programming system are written in 360-Assembler and operate under the MTS time-sharing system on the IBM 360/67 at the University of Alberta. The thesaurus program takes up approximately 1000 4-byte computer words of storage. The system is operative in batch as well as on-line mode.



Overall System

Figure 6

REFERENCES

Aitchison, J., 1970, "The Thesaurofacet: A Multipurpose Retrieval Language Tool", Journal of Documentation, v. 26, no. 3, p. 187-203.

Blagden, J.F., 1968, "Thesaurus Compilation Methods: A Literature Review", ASLIB Proceedings, v. 20, no. 8, p. 345-359.

Bulletin De L'Association Internationale Des Documentalistes, 1968, v. V, no. 4.

Heaps, H.S., and Thiel, L.H., 1970, "Optimum Procedures for Economic Information Retrieval", Information Storage and Retrieval, v. 6, no. 2, p. 137-153.

London, G., 1966, A Classed Thesaurus as an Aid to Indexing Classifying and Searching.

Salton, G., 1968, Automatic Information Organization and Retrieval: New York, McGraw-Hill Book Company.

Sussenguth, E.H. Jr., 1963, "Use of Tree Structures for Processing Files", Communications of the ACM, v. 6, no. 5, p. 272-279.

Selective Dissemination of Marc:
A User Evaluation

by

L.R. BUHR

Unfortunately this paper was not in on time to be included in Proceedings and we will attempt to make it available as a separate reprint to all attendants of the Annual Meeting.

Computer Indexing with S.I.S. II and Famulus

by

A. McPHERSON

COMPUTER INDEXING WITH S.I.S. II AND FAMULUS

Arlean McPherson
Canada Agriculture Research Station
Saskatoon, Saskatchewan

ABSTRACT

A pilot project involving the indexing of two special subject collections, and using two computer systems, was inaugurated at the Canada Agriculture Research Station in Saskatoon in 1968. Papers on Acrididae were indexed using S.I.S. II, and papers on Brassica using FAMULUS. As work on the project progressed, experience applicable to the organization of central library collections and to personal document collections was gained. This is summarized in the following paper.

INTRODUCTION

Early in the development of library services at the Canada Agriculture Research Station in Saskatoon, it was realized that scientists retrieve information not only from the central library, but also from document collections located in offices or laboratories. An appreciation of the value of both types of information files to the total information picture at the station, led to the initiation of a small-scale pilot project to investigate means of facilitating the organization and maintenance of both kinds of document collections. When the project was first organized, construction of indexes to particular subject collections was of paramount importance, but as work progressed, emphasis shifted to the more general consideration of computer application to information retri al in the local research environment.

MATERIALS AND METHODS

To date, the indexing of two collections of reprints and other papers has been undertaken at the Canada Agriculture Research Station in Saskatoon. In each case, the collection of reprints covers a limited subject field in depth, and has been assembled by laboratory staff to serve the information requirements of a team of researchers. In each case, also, the reprints had not previously been indexed, and library personnel were asked to organize the material for efficient retrieval.

The pilot phase of the project involves the organization of papers pertaining to Acrididae, and approximately 2,500 have been indexed since its inception two years ago. Bibliographic data identifying each item, its location, and the descriptors by which it is being indexed for retrieval are entered on a printed form. This information is key-

punched, submitted to a 360/50 computer and processed by a Streamed Information System, S.I.S. II.

This system, developed by Imperial Oil in Calgary, and revised by Peardon (1968) at the Computation Centre at the University of Saskatchewan, Saskatoon Campus in 1968, comprises six programs. EDIT detects format errors and produces a basic input tape. FILE LOAD - 1 converts inventory and source master files to disk work files for use by the UPDATE program, which, as its name implies, handles additions and revisions to the inventory and source files, and produces updated thesaurus and inventory master files for use in subsequent runs. FILE LOAD - 2 converts the updated inventory master into an inventory work file on disk, preparatory to an index print/purge run. The SORT program sorts the cross-reference master into alphabetic order by descriptor concept or index term. And the INDEX PRINT/PURGE program produces print-outs of full or selected indexes as required.

The second phase of the project involves the indexing of papers dealing with Brassica, and approximately 700 items have been indexed since July, 1971. Again, bibliographic, location and descriptive data for each document is key-punched. Hardware is an IMB 360/50. Software is FAMULUS.

FAMULUS is a personal documentation system for research scientists, designed by T.B. Yerke, R.M. Russell and H.D. Burton of the Pacific Southwest Forest and Range Experiment Station in Berkeley, California, (1966). It consists of eight main sub-systems. EDIT writes punched card input onto tape, and permits the user to make corrections, additions, and deletions. SORT rearranges the file order by changing the order of fields within records so the file can be realphabetized. MERGE provides updating facilities and permits enlargement of the files through merging two individual files into one master file. GALLEY prints the file in any of several formats. VOCAB prints in alphabetic order the words in any given field of a file, making lists of index terms, or keywords in title, etc. INDEX lists keywords and indicates in what records they may be found, providing an index to the file. SEARCH scans designated fields in the records of a file, matching them against user-prepared search profiles for print-out. OSSIFY punches card deck equivalents of tape files, for use as safety decks or for massive correction operations.

DISCUSSION

The pilot indexing project was undertaken with two objectives in mind - to produce a useful guide to a specific collection of papers, and to gain practical local experience in the small-scale application of library-computer technology. In the beginning, the practical production of an index to papers on Acrididae took precedence; and, as has been indicated, one covering 2,500 items was produced. As the

project progressed, however, major emphasis gradually shifted to the second objective which was extended to include more general aspects of computer applications to information retrieval and library services in our particular research environment.

A review of the literature confirmed our tentative findings that two basic types of information files, distinct in character and function, are to be found in research establishments. Jahoda, Hutchins and Galford (1966) quote statistics indicating that between 45% and 66% of scientists maintain office or laboratory files of reprints and other papers pertaining to a narrowly specialized subject field, to supplement the more comprehensive and formally-organized library collections to which they have access. Burton (1970) estimates that between 40% and 60% of all information needs in research establishments are filled by the user himself - either through recall or the consultation of personal files. Engelbart (1961) defines the particular quality of each type of file, describing the general store of information housed in libraries as "macrodocumentation", and the smaller packages of information kept by the scientist in the relatively closed domain of his personal files as "microdocumentation". Yerke (1966) discusses subject control of personal collections and how it differs from traditional concepts of bibliographic control, while Wallace (1966) gives a superficial description of the requirements of scientists as documentalists. Both conclude that services should be provided by information specialists to facilitate not only library operations, but also the individual's information handling practices, and have devised computer systems for this latter purpose. Burton (1970), in analyzing personal documentation methods and practices, theorizes that such systems provide a unique contribution towards the satisfaction of information needs, and concludes that even if this contribution can be approximated elsewhere, it can only be done at higher cost to the user. As our work with computer-produced indexes developed, therefore, we attempted to assess the relevancy of our experience to the organization both of a station library, and to personal document collections.

The deficiencies of the index produced on Acrididae are, unfortunately, painfully evident. These derive in large measure from our inexperience and subsequent mistakes in using S.I.S. II, and to some extent from the limitations of the system. As documented by Cherry (1965) and Goodman (1967) the Streamed Information System was originally developed to catalogue the holdings of the Imperial Oil Library in Calgary, and to provide an "in-house" information service for that firm's personnel. It produces an author-subject catalogue in book format; and possesses the convenient feature of reproducing bibliographic and location data for each item whenever a descriptor refers to it in the print-out. This contrasts with many computer-produced catalogues and indexes, in which the user is directed by numerical reference to a second volume to find title and location

of relevant papers; and has particular value in so far as the version used by us did not have machine-search capacity. (Since spring, 1971, however, this feature is available). The catalogues produced by S.I.S. II include a master index containing a complete listing of all holdings in an alphabetically-sorted author-subject file, from which separate author and subject catalogues, partial subject indexes, current acquisitions lists or branch library catalogues may be generated.

As the preceding summary indicates, the Streamed Information System was designed for information retrieval from a discipline-oriented, comprehensive library collection, assembled for permanent retention; or, in other words, for the control of "macrodocumentation". In using S.I.S. II to index what had at one time been a personal document collection, but which was later converted to an information file for use by a team of researchers, a lack of flexibility in the system became apparent. The fixed length of 87 characters allocated for title exemplifies this; and necessitated the butchering of lengthy titles, particularly when translations from foreign languages were required.

On the other hand, advantageous features of S.I.S. II, such as its capacity for generating a thesaurus, were equally apparent. This asset proved invaluable in indexing the Acridid collection, as keywords to control the subject vocabulary had not previously been compiled; and standardization is highly desirable if more than one researcher is to use the file. Herein, however, we erred in failing to invest sufficient time in preliminary compilation and definition of keywords, and this mistake necessitated not only considerable revision, but also affects the quality of the index which was produced. As a result, basic thesauri pertaining to special subjects are to be constructed for all indexing projects at our station, if the index is being produced for the use of more than one scientist.

The second indexing project, like the first, was intended to produce by computer an index to a special subject. During the past several years, research in oilseeds, particularly rapeseed and mustard, has expanded rapidly at our station, and five scientists with support staff are now working as a team on Brassica. Because they encountered increasing problems in the control of office documentation, and because they found substantial areas of overlap in their respective subject interests, a co-operative scheme was suggested, and the library approached for assistance. Previous experience with S.I.S. II indicated that its limited space for title, lack of capacity for inclusion of abstracts, and restricted sorting options, made it somewhat less than ideal for this undertaking, while literature on FAMULUS suggested its greater suitability.

As described by Burton, Russell and Yerke (1969), FAMULUS is a computer-based system designed to support the documentation activies

of the individual scientist with minimum interference in his information-organizing habits and instincts. The user is offered editing, sorting, indexing, vocabulary-building, searching and file-revision features in a package which leaves him free to structure his input according to his idiosyncratic needs. It would seem to be one answer to Engelbart's earlier plea for "a way to store, retrieve, and manipulate the information within the individual's private domain, with information-packet sizes that match his actual needs"; or, in other words, a means to control "microdocumentation".

Again, our adaptation of FAMULUS to the indexing of papers for team use, is not the purpose for which it was designed. But the very options built into the system to satisfy individualized requirements are meeting the needs of the Brassica project. From the point of view of the user, the FAMULUS system is broken down into ten fields, which may, or may not, be labelled for use. In the Brassica project, seven fields are currently being used - author, title, date, publication, descriptor, abstract and location. The other three have been left but may be labelled, if and when required. Information may be retrieved from each field, and the material within each sorted as desired. Each field is open-ended, so that length of title, number of descriptors, length of abstract, etc., may vary. This is limited only in so far as total input for one item may not exceed 4,000 characters.

In the print-out versions of its subject index, FAMULUS follows the conventional format of referring by number to the items listed under keyword. This involves checking the title and location of specific references in a second volume for manual search. However, the inconvenience caused by this formating is minimized, in that FAMULUS has machine-search capability. It also has the capacity to create Keyword-In-Title indexes, in which the necessity of manual indexing may be bypassed, or which may serve as a check on the keywords assigned.

One feature included in S.I.S. II, the absence of which has been felt in our use of FAMULUS, is the capability to generate a thesaurus. FAMULUS compiles a list of keywords used, but does not permit the assigning of hierarchical relationships, or check the use of indexing terminology against previously authorized terms.

In any comparison of FAMULUS with S.I.S. II one other aspect should be mentioned. It was pointed out earlier that the Streamed Information System was designed for the production of a computer-produced library book-catalogue, from which a variety of specific indexes or catalogues might be excerpted. Conversely, FAMULUS was designed to index the files of individual scientists, and has the capacity to merge several smaller collections into a larger unit. At the U.S. Forest Station in Berkeley, for instance, at least one multi-user collection has been so indexed. Each staff member works on a different aspect of an over-all problem, each indexes individually, and the separate

personal files are merged for joint use. To be effective, this presupposes that each researcher discuss his individual indexing habits with the others involved, so that before a search question is submitted to the multi-user file, the various and different approaches of the several indexers are mutually understood. This particular use of FAMULUS differs strategically in design to our multi-user Acridid and Brassica projects, in which vocabulary input is standardized as in conventional library practice.

Indexing of papers on Acridids and Brassicas fortunately coincided with the Selective Dissemination of Information (S.D.I.) service being made available by the National Science Library in Ottawa. Both research teams are among personnel who submitted profiles to this current awareness service; in fact, Brassica researchers have two profiles, and these have facilitated the acquisition of current material. They have also stimulated the co-operation of researchers and librarians, in that all profiles are submitted to, and all alterations handled by, library personnel. In so far as S.D.I. directly affects user requirements for information retrieval in the station, it is relevant to this discussion of the computer-indexing project.

CONCLUSION

In adapting the two systems to projects for which neither was originally designed, we are aware, more particularly in the case of FAMULUS, but also to some extent in our use of S.I.S. II, that their respective potentials relative to "microdocumentation" and "macrodocumentation" are not being fully utilized. However, factors of cost and time, and the priority of team requirements for information retrieval, dictated a procedure which was not only within our means, but which has also clarified our thinking on local requirements, and which has, hopefully, indicated paths which may be explored. Future assessment of station library needs and personal documentation requirements, as well as the preparation of software specifications, will presumably benefit from this learning experience. In addition to the indexes which have been, or are being produced, two thesauri pertaining to agricultural subjects have been compiled. In this latter exercise, procedures for the standardization of subject terminology have been delineated for future application.

CITED REFERENCES

BURTON, H.D., 1970, Thesis Proposal: Personal Documentation Methods and Practices, with Analysis of their Relation to Formal Bibliographical Systems and Theory. Typed manuscript, 27 pp.

_____, RUSSELL, R.M., and YERKE, T.B., 1969, FAMULUS: A Computer-Based System for Augmenting Personal Documentation Efforts. U.S.D.A., Forest Service, Research Note PSW-193.

COMPUTER INDEXING

CHERRY, J.W., 1965, A Computer-Assisted, Industry-Oriented Information-Retrieval System. Ottawa, Canadian Library Association, Occasional Paper No. 48.

ENGELBART, D.C., 1961, "Special Considerations of the Individual as a User, Generator, and Retriever of Information." American Documentation, 12(2): 121-125.

GOODMAN, E.V., 1967, S.I.S. II. Calgary, Alb rta, Imperial Oil Limited, Systems and Computer Services Department, IPRC-2MU-67.

JAHODA, G., HUTCHINS, R.D., and GALFORD, R.R., 1966, "Characteristics and Use of Personal Indexes Maintained by Scientists and Engineers in One University." American Documentation, 17(2): 71-75.

PEARDON, G.W., 1968, Information Retrieval System Specifications. Saskatoon, University of Saskatchewan, Computation Centre.

U.S. FOREST SERVICE, PACIFIC SOUTHWEST FOREST AND RANGE EXPERIMENT STATION, Berkeley, California, 1969, FAMULUS: A Personal Documentation System ... Users' Manual. 39 pp.

WALLACE, E., 1966, A User's Guide to SURF-Support of User Records and Files. Santa Monica, California, System Development Corporation, Memo 2913/0000/00.

YERKE, T.B., 1966, "Adapting Library Computer Programs to Individual Documentation." Proceedings of the American Water Resources Research Association, 2: 432-440.

OTHER RELEVANT REFERENCES

ARNETT, R.H., 1970, Entomological Information Storage and Retrieval. Baltimore, Bio-Rand Foundation, 210 pp.

CHERRY, J.W., 1966, "Computer-Produced Indexes in a Double Dictionary Format." Special Libraries, 5(2): 107-110.

HANSON, C.W., 1964, "Research on Users' Needs: Where is it Getting Us?" ASSIB Proceedings, 16(2): 64-78.

HENLEY, J.P., 1970, Computer-Based Library and Information Systems. New York, Elsevier, 84 pp.

HOLM, B.E., 1968, How to Manage Your Information. New York, Reinhold, 292 pp.

COMPUTER INDEXING

I.B.M., 1969, Index Organization for Information Retrieval. .ite Plains, New York, IBM, Technical Publications Department, GC20-8062-1.

INSTITUT DES CORPS GRAS, SERVICE DE LA DOCUMENTATION, Paris, 1968, Le Colza; ses Produits et ses Derives, Bibliographie Analytique Decennale Indexee, 1957-1966. 99 pp.

INSTITUT DES CORPS GRAS, SERVICE DE LA DOCUMENTATION, Paris, 1970, Thesaurus ITERG; Termes des Industries de Corps Gras et Produits Apparentes. 90 pp.

ITERG, Paris, 1969, Index de la Revue Francaise des Corps Gras. 85 pp.

JAHODA, G., 1970, Information Storage and Retrieval Systems for Individual Researchers. New York, Wiley, 135 pp.

_____, HUTCHINS, R.D., and MILLER, D.M., 1967, "Analysis of Case Histories of Personal Index Use." Proceedings of the American Documentation Institute, 4: 245-254.

LEE, P.J., HANSEN, J.A., and MARSHALL, K.E., 1971, Storage and Retrieval of Indexed and Annotated Bibliographic References. Winnipeg, Fisheries Research Board of Canada, Freshwater Institute, Technical Report No. 209.

MAIZELL, R.E., 1960, "Information Gathering Patterns and Creativity; a Study of Research Chemists in an Industrial Research Laboratory." American Documentation, 11(1): 9-17.

SLATER, M., and FISHER, P., 1969, Use Made of Technical Libraries. London, ASLIB, Occasional Publication No. 2.

WALLACE, E.M., 1964, "Experience with EDP Support of Individuals' File Maintenance." Proceedings of the American Documentation Institute, 1: 259-261.

_____, 1966, "User Requirements, Personal Indexes, and Computer Support." Proceedings of the American Documentation Institute, 3: 73-80.

YERKE, T.B., 1970, "Computer Support of the Researcher's Own Documentation." Datamation, 16(2): 75-77.

Coding Questionnaires

by

S. BERTRAM

CODING QUESTIONNAIRES

Sheila Bertram and Marian Richeson
University of Alberta
Edmonton, Alberta

ABSTRACT

A brief discussion of some of the problems and considerations in coding questionnaires.

Some of the advantages of using a questionnaire to collect data include:

1. large samples, wide range of phenomena
2. can be used to collect different types of data; factual, opinions, attitudes, interests, standards of behaviour, etc.
3. uniformity; standardization of questions asked, wording, order of questions, instructions, etc.
4. objectivity
5. elimination of interviewer bias
6. anonymity may promote honesty
7. respondent has more time to consider answer
8. ease of preparation and distribution

However, a carefully constructed, carefully administered questionnaire is not always simple, quick, and inexpensive but requires time, patience, ingenuity, skill, and funds. It is not difficult to collect data using a questionnaire, the problem is to collect meaningful data.

Some of the common abuses, errors, and imperfections in questionnaires include:

1. lack of clarity in definitions
2. ambiguous or inappropriate wording of questions
3. varying meanings of words to different people
4. unwarranted assumptions
5. double questions
6. leading questions, emotional words
7. use of generalities, too many or weak modifiers
8. unnecessary questions
9. inadequate categories for responses
10. too long
11. inappropriate question sequence
12. embarrassing or objectionable questions
13. seeking opinions and then using them as facts
14. using an existing questionnaire which does not fit the study
15. devising a questionnaire when a suitable one already exists
16. using a questionnaire when the information is readily available

Many of these problems can be identified and eliminated with adequate pre-testing of the questionnaire. Pre-testing can also give some idea of costs and whether the demands are reasonable, some indication of the categories of answers which might be expected, and whether or not these answers are going to produce the desired results.

In addition, many researchers seem to overlook the cooperative nature of the questionnaire, the fact that they are asking a busy person to do something for them. This leads to one major problem of using the questionnaire as a research method, that of non-response; a 60-70% response is usually considered good. Not only do people not respond, leading to inconclusive results, the people who do respond may be statistically different from those who do not, thus ruining the sample and introducing uncontrollable variables.

Some of the procedures which have been suggested to motivate people to return questionnaires include:

1. With the questionnaire. A covering letter should always be sent with the questionnaire describing the purpose of the study, indicating how the end results will be used, stating the sponsoring agency, explaining why the respondent was selected, promising anonymity, and sometimes asking for permission to quote. The more personal the letter, the more likely the response. A self-addressed, stamped envelope and a second copy of the questionnaire for the respondent's files should also be included. In addition, there should be somewhere on the questionnaire for the respondent to check if a summary of the final report is desired. A well designed, short questionnaire with an attractive format is more likely to be answered than a long, carelessly designed, formidable one. Some of the questions that should be asked before sending the questionnaire to an individual are; does the respondent have the information? is he willing to give it? will he give the correct information? is he allowed to give it? is the questionnaire relevant to the respondent and his situation? One suggestion is that the questionnaire be sent to someone in authority who will then delegate the responsibility to the person who knows the answers.
2. Follow-up. A follow-up letter is usually sent after a given length of time to non-respondents. If this does not promote a response, perhaps a post card reminder and then eventually a second copy of the questionnaire along with an appropriate letter is sent. If a very high response rate is required, further procedures such as personal letters, telegrams, telephone calls, a shortened questionnaire with the essential questions may be used.

There are two major types of questions which can be used, open-ended, where the respondent is free to answer in his own words, and closed or fixed-alternative, where the respondent has to choose from the answers provided. The purpose of the study and the

type of response required to provide the necessary data should determine which type of question is used. A better depth of response may be obtained with the open-ended question but it is more difficult to interpret, code, and analyze. The closed question has the advantage of pre-coding possibilities, standardization of answers in the direction required by the study, it is quick and easy to code and analyze, and the keypunching can be done directly from the questionnaire. The closed question is usually easier to answer and therefore more likely to be returned. However, it does force the respondent to make a choice which may not adequately describe his situation and he may, as a result, omit the question. This may in turn invalidate the whole questionnaire. Closed questions therefore require more preliminary work and pre-testing to ensure that all of the reasonable alternative answers have been included

CODING

Coding consists of assigning a number or symbol to each answer according to a predetermined categorization. It is the classification process required for subsequent tabulation and analysis of the data. Before the questionnaire is sent out, the researcher should have drawn up preliminary tables, outlining the category sets necessary for the desired correlations to ensure that the questionnaire will provide these answers. These same preliminary tables can then be used to determine the category sets which will then need to be coded. The respondent himself may do the coding when he answers the questions (closed questions) or the coder may code the responses on receipt of the questionnaire.

The actual codes are usually entered in one of four places.

1. Questionnaire form - by respondent. The code numbers can be printed opposite the possible responses on the questionnaire, thus when the respondent selects his answer, the code is automatically selected at the same time. This procedure saves both time and money since the intermediate coding and transcription processes are not required. Many clerical and accompanying supervisory routines are avoided as are possibilities for errors in transcription. However, more work must go into pre-testing closed questionnaires so that the categorization accurately reflects the probable answers before the questionnaire is sent out. The time saved on receipt of the questionnaire may well have been already consumed at the pre-testing stage. In addition, if hand tabulation is to be used, the questionnaire form is usually not suitable for repeated handlings and so a special code sheet is likely to be prepared anyways. If machine tabulation is to be used, one step is eliminated since the keypunchers can work directly from the questionnaire form. Mark-sense and pre-punched cards may be used directly, thus eliminating even the keypunching.

2. Questionnaire form - by coder. The code numbers are often placed on the questionnaire form by the coder after its receipt. The major advantage of this method compared with the above is that the categorization can be done after scanning the actual responses and coding on the questionnaire form is no longer restricted to closed questions.

3. On transcription sheets. In some cases, the information or raw data is transcribed from the responses onto transcription sheets from which the codes can be readily assigned. This is frequently done when the arrangement of the questions is not the best arrangement for coding and tabulating the data. The arrangement of the transcription sheet should be carefully thought out in advance and related to the arrangement of the codes on the punched cards. One large, unwieldy sheet should be avoided since it takes up a lot of space, only one person can work on it at a time, and there is a greater possibility of putting the information in the wrong place. Once the data has been transcribed, the appropriate codes are then usually assigned right on the sheet.

4. On code sheets. The fourth method is to put the code only on special code sheets directly from the response given on the questionnaire. These sheets can then be used directly in hand tabulation or arranged so as to facilitate keypunching. The major disadvantage is that it is difficult to check for accuracy when the codes have been separated from the data.

Exactly where the code is to be placed, either directly on the questionnaire form or on a separate sheet, often depends on whether or not machine tabulation is going to be used, since keypunchers prefer to work from code sheets. However, this extra transcription step is a tedious operation which can produce errors and is usually wasteful if the questionnaire is primarily made up of closed questions. If there is a combination of closed and open-ended questions, it may be best to provide for a double keypunch, one for the closed questions, to be done directly from the questionnaire, the second for the open-ended questions, to be done from code sheets. The decision as to where the codes are to be placed and how the keypunching is to be done must be made before the layout of the questionnaire can be determined.

CODES

The categories into which the raw data are to be placed should be made with the distribution of possible answers, ease of tabulation, and eventual correlations in mind. Each category within a given set must be choosen using the same classification principles, the set must provide for all possible answers, and the categories within the set must be mutually exclusive. The code for closed questions can be prepared in advance and the questionnaires coded upon receipt.

CODING QUESTIONNAIRES

While the possible answers for open-ended questions should also be considered in advance, the final code should not be drawn up until a large proportion of the questionnaires have been returned and scanned so that the code encompasses the actual answers. A more detailed classification should be used than will appear in the final report since it is easier to combine categories than it is to go back and separate categories. On the other hand, the code itself should be kept as simple as possible and as uniform as possible throughout the questionnaire so that it is easily learned and remembered. If hand tabulation is to be used, alphabetic, mneumonic notations are often prefered while numeric notations are preferable for keypunching. If punched cards are to be used, it is desirable to use 12 or fewer categories per item so that each question will correspond to one column on the card. Since some equipment can not handle multiple punches, it is best to avoid them. Therefore, even if there are only a few categories per question, each question should still have a separate column. It is always useful if standard classifications or classifications used in similar or related studies are used so that valid comparisons can be made. This also might mean that certain background data may be already available in a useable form.

In addition to the codes for the data obtained from the questionnaire, other information often needs to be supplied by the coder. Each questionnaire must have an identification number. This might be an accession number assigned in order of receipt, or a control number assigned before mailing to be used as a check on non-receipts. Other information about the respondent not supplied directly by answers to the questionnaire may be required by the study, such as sex, address, occupation, number in the listing used for sampling, etc. and a combination of the codes for this information can provide the identification number. This information is often coded directly onto the questionnaire upon receipt (since some of the information may be available only from the return address) and later transcribed onto the code sheet.

CODERS

A training period for the coders is usually necessary. This provides an opportunity not only to see how well the coders perform, but also to see how well the codes and the coding procedures work. Usually all the coders practice coding the same sample with a supervisor reviewing the work until satisfied that the coder understands what is to be done and will do it consistently and accurately. Detailed coding instructions must be written down so that uniform procedures are being followed and any changes must be immediately brought to the attention of all the coders. The coding should be a routine operation with it understood that any problems are to be brought to

the attention of the supervisor. In addition, it is helpful, and perhaps provides for better performance, if the coders are familiar with the study, its purpose and methodology, and are made to feel that they are an essential part.

The coder can work through the questionnaires, coding one or two questions at a time. Similarily, if there is more than one coder, each is assigned a number of questions rather than a number of questionnaires. This method is recommended if the codes are complex and difficult to remember. Greater consistency in coding individual questions is usually obtained with this method. Others argue that it is best to let the coder work through the questionnaire, question by question, so that he gains an overall picture of the respondent and can spot inconsistencies in responses. Either way, the coder must be required to sign by the question to indicate that it has been coded and on the code sheet to indicate responsibility.

Verification is usually done in the form of spot checks by the supervisor or by having questionnaires coded by two different coders working independently and then reviewing differences.

When all the codes and formats have been determined, it is extremely useful to have everything combined and written down in one place; the questionnaire, the categories and the corresponding codes, the code sheet layout, and the punched card layout. Such a document will more than repay its cost of preparation when it is then distributed to everyone concerned with coding and analyzing the data: coders, keypunchers, programmers, various advisors, and those involved with interpretation and writing the report. For one of the biggest problems with using the questionnaire is that of communication, and when the transcribing of information is considered as well, the opportunities for misunderstandings and mistakes are myriad indeed.

REFERENCES

Doby, J.T., ed., 1967, An Introduction to Social Research: New York, Appleton-Century-Crofts.

Goldhor, H., 1969, An Introduction to Scientific Research in Librarianship: Washington, U.S.O.E.

Oppenheim, A.N., 1966, Questionnaire Design and Attitude Measurement: New York, Basic Books.

Parten, M., 1950, Surveys, Polls, and Samples: New York, Harper.

Simon, J.L., 1969, Basic Research Methods in Social Science: New York, Random House.

University of Calgary Library:
Computer-Based Circulation Control System

by

R. MURCH

UNIVERSITY OF CALGARY LIBRARY: COMPUTER-BASED CIRCULATION CONTROL SYSTEM

Ron Murch
University of Calgary
Calgary, Alberta

ABSTRACT

Because of the physical requirements of a new addition as well as the demands for increased control, better reader service, and greater information availability, the University of Calgary Library has developed a computer-based circulation control system. The complete system, organized as four (4) interacting subsystems (daily processing, reader identification, accounts, statistics) provides information processing support and procedural and reporting flexibility.

The data collection functions are handled primarily by an off-line terminal system (C-DEK) supplemented by a manual coding effort. The C-DEK system is upgradable to an on-line environment, if such is desirable in the future.

In designing this system, the emphasis has been on flexibility and expandability. Upon completion, implementation will have been in three (3) phases. - the last phase, the statistics subsystem, will not be totally operational until a reliable data base has been accumulated.

COMPUTER-BASED CIRCULATION CONTROL

INTRODUCTION

The concept of automated support for library operations perhaps began in 1930 when Ralph Parker at the University of Texas started experimenting with punched card equipment for use in circulation. Since then, many libraries have developed various areas of automated support, with the circulation functions usually enjoying a high priority.

Since receiving its charter in 1966, the University of Calgary has expanded quite rapidly. Necessarily, the increase in scope and size of the university community has accentuated the need for better information

to be available in a shorter period of time. The implementation of a computer-based circulation system is going to assist the Library in satisfying that need.

REQUIREMENTS AND OBJECTIVES

There were five (5) main requirements of this system:

1) Primarily, the Library must be able to meet the coming demands in the areas of reader service.
2) Automated support must be able to assist in relieving pressures and conditions internally within the Circulation Area.
3) The Library must be able to improve the controls exercised on circulation functions.
4) It must be able to meet the coming demands for management-type information, both in terms of availability and timeliness.
5) The construction of an addition to the Library has created a physical need for automated support by separating the charging areas from the discharging area.

The use of computer support greatly increases the storage and handling capabilities for information relating to the circulation processes. In providing this support, there were three (3) main objectives:

1) It is necessary to develop and retain both procedural and reporting flexibility.
2) Provision for co-ordinating expansion of the existing system to benefit other areas of Library operations (cataloguing, acquisitions, etc.) must be retained.
3) The main areas of support should be in the areas of data collection and reporting.

SYSTEM DESCRIPTION

The system is designed as four (4) subsystems, which interface together, to provide the necessary data. These subsystems are responsible for processing the data, maintaining the files, and producing the reports which relate to each particular area. There are three (3) types of files used - master files, work files, history files - and five (5) types of reports - system monitoring, circulation oriented, internal working papers, reader oriented and management oriented. The interaction of these subsystems is shown in figure 1.

Daily Processing

This subsystem carries the major load of the processing. It feeds information to all the other subsystems and uses information maintained by the Reader Identification Subsystem. The major functions satisfied by this area are those of file maintenance on both the Book Identification

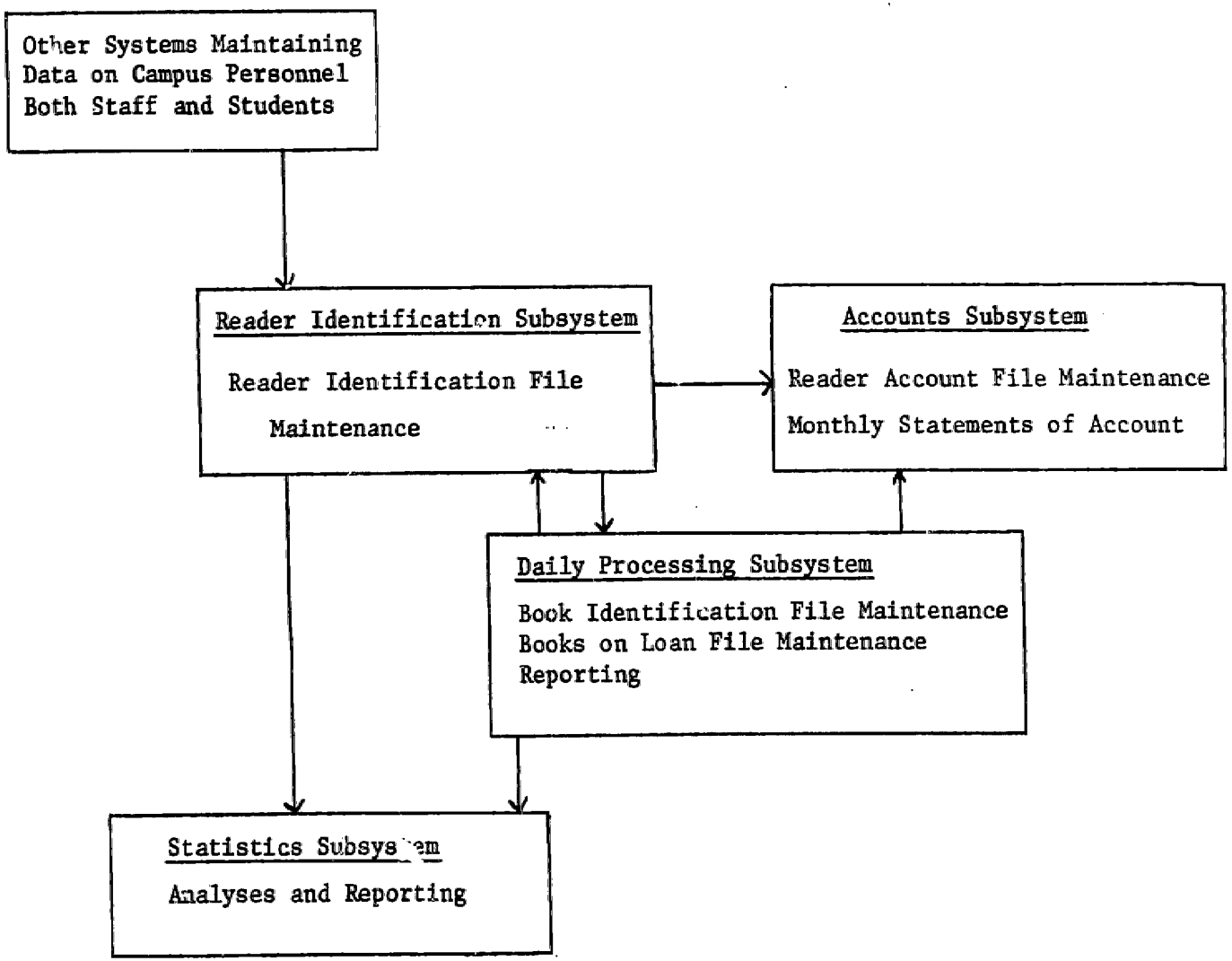


Figure 1.

The interaction of the four subsystems.

File and the Books on Loan File. Products of this maintenance include several reports, notices, and internal working papers.

Reader Identification

This subsystem performs file maintenance functions on the Reader Identification File. In accordance with a general University desire to collect data on campus personnel (staff and students) in a uniform manner, the information maintained by other systems is used to keep our data consistent with official campus records. The numbering scheme for assigning Reader I.D. Numbers is quite flexible and allows for maintaining information on readers who may not be associated with the University community.

Accounts

This subsystem receives data on fine and charge assessments as well as account payments and adjustments and maintains it for reporting on the Monthly Statement of Account. Each reader will receive a statement detailing the activity in his account since the last time the balance was zero. If there is no activity and his account balance is zero, he will not receive a statement.

Statistics

This subsystem utilizes information gathered in other systems to produce statistics relating to circulation functions and collection usage. A history file is retained to allow analyses of past use. This area is only partially operational at present and as future requirements become apparent will be fully developed to provide statistical information relating to management needs.

DATA COLLECTION

There are two methods of data collection being used by this system:

1) Colorado Instruments Inc. C-DEK system records all information relating to circulation - charges, discharges, renewals, tracers, fine payments.

2) Other information is coded onto special forms and keyed onto magnetic tape.

C-DEK

The C-DEK system is a totally integrated data collection system which records on magnetic tape information transmitted from the terminals to the central controller (multiplexor). Our system has six (6) identical terminals (with a capacity for adding ten more) transmitting over a dedicated cable to the central controller which records the information simultaneously on twin tape recorders (800 bpi, 9 track, 500 cps). The terminals are all identical and mounted in pairs at three (3) locations. Thus, in the event of a hardware malfunction, a particular location is not totally disabled.

The terminals (composed of an 80-column card reader, a 22-column badge reader, and an 11-column keyboard) record information identifying the transaction, the book involved (if any), the reader involved (if any),

and other related data (loan period, receipt number, amount paid). The Human Error Control Logic feature forces an operator to enter the items of information required for the transaction selected. The terminal will not transmit if a required item of data is missing.

Book identifying information is entered using an 80-column card. These cards contain the call number, copy number, book I.D. number and material category code for each physical volume being circulated within the system. They are stored in the book pockets of the matching books. Our C-DEK terminals read only the 7-digit book I.D. numbers from the cards entered. Should a book I.D. number be required for a transaction (i.e. charging or discharging) and the book card is either not available or not usable, the number may be entered via the keyboard.

Reader identifying information is entered using a 22-column badge which is punched with a 10-digit reader I.D. number. The University Student I.D. card is being used for students' badges and special plastic badges are being used for all others. There is no provision for keyboard entry of reader I.D. numbers. Thus, should a reader's badge not be available or usable, he must obtain a temporary cardboard badge before any transactions using his number may be entered.

Control punches are used in both the book card and reader badge to insure that they are entered correctly. If the terminal cannot read the control punch properly, it will not transmit the data.

Coded Data Entry

The data entered into the system through coding and subsequent keying includes the following:

1) All data for maintenance of the Book Identification File (additions, changes, deletions, inactivations).
2) Some data for maintenance of the Reader Identification File (non-campus readers).
3) Some data for maintenance of the Accounts Master File (account adjustments, book replacement charge assessments).
4) Transactions which would normally be entered via a terminal, but for some reason, must be entered manually.

Processing

The data recorded by the C-DEK terminals and coding operations are edited and processed by the Daily Processing Subsystem every evening after the Library closes. The reports are then available for distribution when the Library opens the following morning. The Book Identification File and Books on Loan File are updated daily, with the other master files in the system being updated on a less frequent schedule.

REPORTS

As mentioned previously, there are five (5) types of reports produced by this system - system monitoring reports, circulation oriented reports, internal working papers, reader oriented reports and management oriented reports.

System Monitoring Reports

These reports include an extensive analysis of terminal transactions relating location, throughput, transmission errors, and read errors. By monitoring the performance of our data collection equipment, we are able to anticipate problems and use the information for preventive and corrective maintenance. We are also monitoring file and report sizes and usage.

Circulation Oriented Reports

The only major reports produced strictly for use by the circulation areas are the Reader Identification Report and the Index to Reader I.D. Numbers. There are some smaller reports which provide information to the circulation areas and also serve as management information. The system is also capable of reporting on the internal location of volumes (i.e. Undergraduate Reading Room, Bindery, Cataloguing, etc.).

Internal Working Papers

The major report in this category is the Index to Book Identification Numbers which is used to determine the book I.D. number of a volume for which the call number is known. Other reports contain information on irregularities detected by the system (charge with no previous discharge, attempted renewal with the readers not the same, etc.), on information processed by the system (daily accounting transactions, notices produced, etc.) and various errors detected which require attention. The information is used to monitor the actions of the system and to respond quickly to situations requiring attention.

Reader Oriented Reports

The major report in this category is the Books on Loan Report which lists all books presently out in circulation, as well as those returned during the last reporting period. Other reports include the notices (recall, first overdue, second overdue and charge) and Monthly Statements of Account.

Management Oriented Reports

This category includes summary reports of data processed by the system (accounts outstanding by reader category and balance, notices produced, etc.), as well as statistical reports on system performance and data processed. Most of the reports produced by the Statistics Subsystem are of this type.

CONCLUSION

Although the original data conversion effort to create the Book Identification File began in October, 1970, the implementation of the major portions of this system is just now taking place. It will require a reasonable period of operation to build a reliable data base for the statistics subsystem to access. Consequently, the implementation of the complete system will have been in three (3) phases with the major statistical portions not being totally operational until early 1972.

The benefits derived from the increased capabilities of this system for information handling and massaging will be evidenced in the future. At present, while building a data base, the system will be used mainly for executing and controlling the circulation functions; however, the basic design has allowed for expansion of the capabilities of this system and also, for the eventual interfacing/integrating of this system with other areas of automated support in library operations.

I.B.M. Information Retrieval Packages from LUHN to Now

by

S.E. FURTH

Unfortunely this paper was not in on time to be included in Proceedings and we will attempt to make it available as a separate reprint to all attendants of the Annual Meeting.

SELECTIVE DISSEMINATION OF MARC: A USER EVALUATION

Lorne R. Buhr
University of Saskatchewan
Saskatoon, Saskatchewan

ABSTRACT

After outlining the terms of reference of an investigation of user reaction to the selective dissemination of MARC records, a summary of the types of users is given. User response is analysed and interpreted in the light of recent developments at Library of Congress. Implications for the future of SDI of MARC in a university setting conclude the paper.

INTRODUCTION

F. W. Lancaster (1968) in his detailed study of MEDLARS makes the following statement, which has application to all SDI work: "In order to survive, a system must monitor itself, evaluate its performance, and upgrade it wherever possible." Since SELDOM operates in a fairly new field, SDI for current monographs, an evaluation is most important. To a great extent it must be made without reference to other systems since most of the operational SDI services deal with tape services in various fields of scientific journals, and although there are some parallels, there are numerous differences. Whereas, services such as Can/SDI cater primarily to the natural and applied sciences, SELDOM opens up the possibilities for SDI in the humanities and social sciences.

The background to the SELDOM Project at the University of Saskatchewan has been outlined earlier by Smith and Mauerhoff (1971) and will not be repeated here. After five months of operation a major questionnaire was sent out to each of 121 participants in the experimental SELDOM service. This questionnaire was based almost entirely on the one used by Studer (1968) in his dissertation at Indiana State University. The general purpose of the study was to elicit user reaction to SELDOM, their evaluation of its usefulness, time necessary to scan the weekly output, suggestions regarding continuance of the service, etc. Besides this general purpose, the gathering and analyzing of data on SELDOM will be useful to the Library Administration in determining the future of an SDI service of this nature. A separate cost study is being prepared in this connection.

Several factors prompt a cautionary stance in assessing the value of an SDI system on the basis of one questionnaire: (1) There is no control situation to which we can compare SELDOM, ie., there was no

systematic service for current awareness in the field prior to the advent of SELDOM. Faculty and researchers were dependent on their ingenuity to ferret out information on new books which were pertinent to their field of research and instruction. SELDOM, is therefore being compared to a conglomeration of ad hoc methods which may be as numerous as the individuals using them. Therefore, we must be cautious or we will tend to say, "Something in the field of current awareness is better than nothing," when we really do not know what that "nothing" is. (2) Although SELDOM had been operational for some twenty weeks when evaluation began, this is a relatively short period on which to base an assessment. On the other hand Studer's evaluation was based on the experiences of thirty-nine users and covered only eight weekly runs against the MARC tapes scheduled on an every other week basis. (3) SELDOM was implemented without any study to determine the adequacy of the ad hoc approaches, to which I have already referred, nor to assess the patterns of recommendation for purchase. It was assumed that there was a need for SELDOM and some of the response would indicate that this is a fairly valid assumption, since almost 90 per cent of the respondents wanted the service continued. A random investigation in mid-August of 748 current orders in the Acquisitions Department for books with imprint of 1969 or later revealed that 95 or 12 1/2 per cent referred to SELDOM as the source of information for a particular recommendation to purchase. This may or may not be significant since there is no way of assessing whether these items would have been recommended anyway, only later perhaps.

One by-product of orders based on SELDOM information is that correct LC and ISBN numbers are given and with the capabilities of the TESA-1 cataloguing/acquisitions system such orders can be expedited more quickly and can also be catalogued sooner than non-MARC materials, thus ostensibly getting the desired item to the requestor in less time than previously. SELDOM is valuable in our University setting, therefore, not only as a means of awareness of new items, but also in the actual retrieval of the item for the user, in this case through acquisition. Our analysis, however, must be directed to the effectiveness of SELDOM as an awareness service, vis a vis the ad hoc approach.

USER GROUP

Of 121 questionnaires sent out, 77 or 63.5 per cent were returned. Six of these had to be rejected for the purposes of this study since either only a few questions had been answered or a general letter had been sent instead of answering the questionnaire. Thus, the data presented in this study will be based on 71 completed questionnaires or 58.6 per cent return. Three additional verbal comments were made to the writer and thus we in fact heard from 80 or 66 per cent of the users. The term "users" will designate the 71 who completed their questionnaires, although comments from the other nine individuals will also be referred to.

The users have been grouped into three categories according to Table 1.

TABLE 1

I.	Library and Information Science		
	A. On campus	12	
	B. Off campus	17	29
II.	Social Sciences and Humanities		
	A. On campus	15	
	B. Off campus	2	17
III.	Natural and Applied Sciences		
	A. On campus	23	
	B. Off campus	2	25

Categorization was along fairly traditional lines, with category I being necessary because of the large number of people falling into this area. The 17 off campus users coming under designation (I) represent the library schools in Canada as well as librarians/information scientists in Canada and the United States. The on campus users are library department heads and heads of branch libraries.

Included in the Social Sciences and Humanities are the fields of Psychology, Sociology, History, Economics, English, Commerce, Classics, etc. The Natural and Applied Sciences include all the Health Sciences plus Physical Education since the two profiles in that area are tending toward the Health Sciences. Engineering, Poultry Science, Physics, Chemistry, Biology, etc. are represented here.

OBSERVATIONS

A sample of the questionnaire used appears at the end of this paper and includes a tally of the number of responses for each possible alternative answer to each question. In some cases the total number of replies for a question is less than 71. This is explained by the fact that some questions on some questionnaires were not answered or were answered ambiguously so they could not be tallied.

Generally speaking, users found SELDOM to be good to very good in providing SDI for new English monographs. Twenty-five point eight of the users found the lists very useful while 48.5 per cent said they were useful. Six users said the listings were inconsequential for their purposes; in several instances this may be due to poor profiling or profiling for a subject area in which little would appear on the MARC data base.

Twenty-three point six per cent of the users indicated that in most cases items of interest found on the SELDOM lists were previously not known to them. Forty-five point eight per cent said that "of interest" items were frequently new. Seventy-six per cent of the group believed that the proportion of "of interest" items which also were new was satisfactory, a percentage which speaks well for the currency and effectiveness of an SDI capability.

One of the chief drawbacks for which SDI services are often cited is the absence of evaluative commentary or abstract material to accompany the citations. Some tape services do provide either an abstract or a good number of descriptors, and this has pro to be an asset in helping the subscriber. SELDOM is based on the MARC tapes which provide complete cataloguing data but do not give either evaluations or a multiplicity of descriptors. (Some indications are that the information now available in Publishers Weekly might at some time in the future be added to the MARC tapes.) Interestingly enough, 83.5 per cent of the users said the information included in the entries was adequate to determine whether an item was of interest or not. Predictably, title, author/editor and subject headings were the three indicators, in that order, which were found most useful in making evaluations. This is significant since titles in the Humanities and some of the Social Sciences, particularly, are often not as specific in describing the contents of a work as are titles in the physical sciences.

Sixty-three point five per cent of the users indicate that SELDOM information is used for recommending titles for acquisition by the Library. As a result it is quite possible that purchasing in the areas covered by SELDOM profiles may increase and the tendency to broaden the collection should increase. Unfortunately no pattern of pre-SELDOM recommending for purchase is known. Some instructors use the weekly printouts to keep current bibliographies on hand both for teaching purposes and for research purposes. Since over half the users (55.8 per cent) needed no more than ten minutes per week to scan the printouts, there is no indication that excessive time is taken up in the use of such an SDI service.

In reply to the question, "Would you be willing to increase the number of irrelevant notices received in order to maximize the number of relevant one?" opinions were nearly balanced with 58 per cent replying in the affirmative and 42 per cent answering negatively. On the other hand, increases in the MARC data base expected some time in 1972 when other Roman alphabet language imprints and records for motion pictures and filmstrips are added, did not seem problematic with only 25 per cent of users asking that an upper limit be placed on the quantity of material retrieved by their profiles. Numerous individuals (30) responded favorably to the prospect of wider language coverage by MARC. On the other hand, several individuals commented that non-English output on SELDOM would not enhance

the service for them, and this likely reflects language capabilities more than a lack of non-English material in their subject area.

The question regarding format brought interesting comments, especially from library personnel and off campus librarians: "Computer type format is often confusing." "A book designer should be consulted to improve the format." "Spacing could be improved to separate title and imprint information from subject headings and notes at foot of entry. Would make scanning easier."

Questions fourteen, nineteen and twenty-one provide an over-all summary of user reaction. Eighty-eight point six per cent of users want the service to continue. Overall value of SELDOM was rated "very high" by 11.3 per cent, "high" by 33.8 per cent, "medium" by 42.2 per cent and "low" by 12.7 per cent. SELDOM served to demonstrate the possibility of SDI for monographs "amply" according to 36.6 per cent of users, "adequately" to 50.6 per cent of users, and "poorly" to 12.65 per cent of users. There was less certainty on how such a program should be administered or costed particularly since a long range cost study was not yet available. Clearly those who were impressed with SELDOM's effectiveness and future possibilities wanted other faculty to have the same opportunities, yet they cautioned against a blanket service. One comment sums this up best, "It should be available to anyone who has a perceived need for it -- but require them to at least make the effort of setting up the profiles, etc." Many of the less than enthusiastic comments about SELDOM could be correlated with little or no user feedback to the search editor in order to improve relevancy and recall. User education in this regard is crucial in order that all users fully understand the possibilities and limitations of the SDI service. The success of any exisiting SDI service in the periodical literature has hinged on a good data base and up-to-date, specific profiling according to Smith and Lynch (1971). The effectiveness of the profiling is a direct function of the ingenuity and persistence of the user and the profile editor.

DISCUSSION

This study has attempted to weigh the usefulness of an SDI service primarily with regard to its utility as a current awareness service. SELDOM, in order to be worthwhile must either be faster or broader in its coverage than exisiting services. Two comparisons readily arise out of the commentary of the users. Some library science professors felt that the LC proofslip service was just as fast as SELDOM and thus there was no advantage in having the latter when the former was available. A study done at the University of Chicago by Payne and McGee (1970) repudiates this argument fairly effectively. Findings at Chicago show that MARC is faster than the corresponding proofslips. A number of users rely heavily on publishers blurbs and pre-publication notices and find that often books for which

records appear on SELDOM are already on the library shelves. This observation is not altogether an indictment of SELDOM since another user observed that he appreciated being able to have the hard copy immediately; and in some cases he might not even have known about the item except for SELDOM. Some users mentioned that waiting for evaluative reviews could put one at least a year behind just in placing the order for the book, let alone receiving it.

SELDOM has the virtue of informing individuals of the existence of new books, but the delay in having the actual item might be problematic, so one question was directed to this consideration. Some people felt that it was at least worth something to know that a book existed even if one could not consult it immediately. Numerous complaints were aired regarding the slowness of obtaining items ordered through a library's acquisitions' department. In fact one user said this slowness meant he had to purchase personal copies of items he wanted/needed. As indicated earlier in the introduction, the TESA-1 acquisitions-cataloguing routine at the University of Saskatchewan Library does have the capability to speed up actual receipt of books by the patron.

A recent development at the Library of Congress has definite implications for the future of SELDOM and any other MARC-based SDI programs. The CIP (Cataloguing in Publication) / program initiated this summer means that LC will now be able to make available cataloguing information, except for collation, for books about to be published, at a time factor of up to six weeks before publication. Such MARC records will have a special tag designating them as CIP material. Furthermore, CIP records will appear only on MARC, the number predicted is 10,000 for the first year and 30,000 by the third year, a figure which would include all American imprints. MARC-OKLAHOMA // has already surveyed the subscribers to its SDI Project to determine whether they would prefer to receive both CIP MARC records and regular MARC records or only one of the two categories. Users preferred to receive both types of information and appropriate changes have been made to the Oklahoma SDI programs. Beginning with September MARC CIP records will appear and present information on books 30 to 45 days before they are published.

/ Library of Congress, Information Bulletin, v. 30, no. 29, p. 426-427 (July 22, 1971) and v. 30, no. 32, p. 463 (August 12, 1971)

// Oklahoma. Department of Libraries, Automation Newsletter, v. 3, no. 4, p. 12-13 (1971)

Several library personnel appreciated the usefulness of SELDOM as an outreach service of the university library into the academic community. They see SELDOM as a public relations tool. Numerous efforts are at the present time being made by librarians to alert individuals to materials in their several fields of interest, and SELDOM can play an important role in providing an active dissemination of information on a systematic basis. This is the direction in which we need to move so that our role becomes both that of a collector of information and a disseminator of information. Special librarians have been doing this kind of thing for years and SELDOM allows for specialized service to a larger user group.

IMPLICATIONS AND CONCLUSIONS

1. An SDI service based on MARC can be helpful in building a balanced library collection depending on the efforts of faculty and/or bibliographers in setting up their profiles and maintaining them. The article by Ayres (1971) is particularly good on this aspect. The parameters of the MARC data base must constantly be kept in mind, just as must the constraints of the ad hoc methods be considered in any comparisons. Publishers' blurbs in journals have the limitation of not systematically covering all the publications in a given subject area; book reviews tend to appear too late to allow users to receive current information on new books; SELDOM corrects the first shortcoming at the expense of not having the evaluations appearing in book reviews. On the other hand MARC tapes do represent the cataloguing of books in the English language by one of the largest national libraries in the world, and thus provide a coverage which is hard to duplicate by any one other alerting service.

2. Comments, especially from users in the Social Sciences and Humanities indicate that an SDI system for new monographs has greater pertinence in their area than perhaps in the Natural and Applied Sciences simply because of the nature of research done in the two areas. A recent study by J. L. Stewart (1970) substantiates this factor for the field of political science. His detailed analysis of the patterns of citing in the writings appearing in a collective work in political science indicated that 75 per cent of such citations were from monographs leading him to the obvious conslusion that "monographs provide three times as much material as do journals" in the field of political science. By contrast, journals are likely more crucial for the fields of natural and applied science, and provide the key access point for vital information.

3. SDI of MARC, most users felt, should demand a fair amount of effort on the part of users to assure that the service would obtain optimum return for money invested. A blanket service to all faculty would be wasteful since many faculty would not have a perceived need for it and others would not use it enough if it was simply offered free to everyone. Comments tended to favor making contact through the departmental

library representative and channel weekly printouts through this individual. A cost study will help determine whether it is economically feasible to operate SELDOM in an academic setting with at least 100 users. If current subscription costs for SDI services such as those offered by Can/SDI of the National Science Library, Ottawa, can be maintained and early indications are that they can, a cost of $100 per profile per year may be feasible bringing the annual expenditure for 100 users to $10,000. A chief variable which makes effective costing difficult is the variation in the number of records appearing on each weekly tape and this is a variable which can only be dealt with by prediction on the basis of the number of records on past tapes.

4. SELDOM has the virtue of adding a major role of dissemination of information to libraries which up until now have primarily operated as storers of information.

ACKNOWLEDGEMENT

The author thanks William J. Studer, Director, Indiana University Regional Campus Library Distribution Center, for granting permission to use one of the questionnaires used in his dissertation.

REFERENCES

AYRES, F. H., 1971, "The Case Against Marc: How Strong Is It?" Library Association Record, v. 73, no. 7, p. 131.

LANCASTER, F. W., 1968, Evaluation of the Medlars Demand Search Service; Bethesda, Md., National Library of Medicine.

Library of Congress, Information Bulletin, v. 30, no. 29, p. 426-427 (July 22, 1971) and v. 30, no. 32, p. 463 (August 12, 1971)

LYNCH, J. T. and SMITH, G. D. W., 1971 "Information by the Yard," New Scientist and Science Journal, v. 50, no. 745, p. 32-41.

_____, 1971, "Scientific Information by Computer," Nature, v. 230, p. 153-156.

MAUERYOFF, G. R. and SMITH, R. G., 1971, "A Marc II-Based Program for Retrieval and Dissemination," Journal of Library Automation, v. 4, no. 3, (in press).

Oklahoma. Department of Libraries, Automation Newsletter, v. 3, no. 4, p. 12-13 (1971)

PAYNE, C. T. and McGEE, R. S., 1970, "Comparisons of LC Proofslip and Marc Tape Arrival Dates at the University of Chicago Library," Journal of Library Automation, v. 3, no. 2, p. 115-121.

STEWART, J. L., 1970, "Literature of Politics: A Citation Analysis," International Library Review, v. 2, p. 329-353.

STUDER, W. J., 1968, Computer-Based Selective Dissemination of Information (SDI) for Faculty Using Library of Congress Machine-Readable (MARC) Records, Ph.D. Disseration, Graduate Library School, Indiana University.

SAMPLE PROFILE NOTICES

INFORMATION , STORAGE , RETRIEVAL

COOPER, MICHAEL DAVID.

EVALUATION OF INFORMATION RETRIEVAL SYSTEMS: A SIMULATION AN D COST APPROACH.

BERKELEY,SCHOOL OF LIBRARIANSHIP, UNIVERSITY OF CALIFORNIA,1 971.

XIII, 209 L. ILLUS. 28 CM.

INFORMATION STORAGE AND RETRIEVAL SYSTEMS COSTS. **INFORMATI ON STORAGE AND RETRIEVAL SYSTEMS EVALUATION.

LC 72-026658 P1364 EN 01 TW 000 WT 000 S R0318 FC BIBL LENG Z699 029.7

822 33, SHAKESPEARE WILLIAM

SECCOMBE, THOMAS, 1866-1923.

THE AGE OF SHAKESPEARE (1579-1631), BY THOMAS SECCOMBE AND J OHN W. ALLEN. WITH AN INTROD. BY PROFESSOR HALES.

FREEPORT, N.Y.,BOOKS FOR LIBRARIES PRESS<1971>

2 V. 23 CM. **LIBRARY OF SHAKESPEAREAN BIOGRAPHY AND CRITICI SM, SER. 3, PT. B.

SHAKESPEARE, WILLIAM, 1564-1616. **ENGLISH LITERATURE EARLY MODERN (TO 1700) HISTORY AND CRITICISM.

IS THIS TITLE USEFUL? YES NO CANNOT TELL COMMENT DO YOU WISH TO RECOMMEND FOR LIBRARY ACQUISITION? NO YES WHY?
LC 74-160993 P1602 EN 01 TW 000 WT 000 S R0323 FC LENG PR421 822.33 ISBN 0836958608

SELDOM EVALUATION QUESTIONNAIRE

1. What is your feeling about the SDI lists as a source for finding out about the existence of newly published works in your fields of interest? Would you say that the lists provided a source which was:

   (a) very useful (18) (b) useful (34) (c) moderately useful (12) (d) inconsequential (6)

2. Do you feel that the SDI lists brought to your attention works of interest which are not generally cited by other sources that you use to learn of new publications?

   (a) many works (10) (b) some works (39) (c) a few works (19) (d) none (2)

3. How would you characterize your feeling about the relative proportions of the items "of interest" (relevant items) and "those not of interest" (irrelevant items) included in the SDI lists?

   (a) the proportion of relevant items in the lists was satisfactory. 57
   (b) the proportion of irrelevant items in the lists was too high. 13

4. It is inevitable that some "not-of-interest" items are included in the SDI lists. Was the inclusion of irrelevant notices bothersome to you?

   (a) yes (6) (b) no (65)
   REASONS:

5. On the other hand, it is possible that for any given search run, some relevant items in the file are missed. The chance of relevant items being missed can generally be minimized by certain search adjustments, but with a resulting increase in irrelevant notices. Would you be willing to increase the number of irrelevant notices received in order to maximize the number of relevant ones?

   (a) yes (40) (b) no (29)
   REASONS:

6. The SDI lists notified you of an average of ___ items per list which you judged to be "of interest". On a purely quantitative basis, would you say that this number was satisfactory, or for some reason too small or too large?

   (a) satisfactory (48) (b) too small (16) (c) too large (1)

7. When the input to the MARC file is increased, your SDI output would also likely increase. Do you feel that you would like to be able to set some arbitrary upper limit on the quantity of items included in each SDI list even at the risk of missing a number of relevant items?

   (a) yes (17) (b) no (51) If yes, Maximum number ___
   REASONS:

8. The SDI lists alerted you to a number of items which you judged to be "of interest". Would you say that "of interest" items were new to you?

(a) in most cases 17 (b) frequently 33 (c) occasionally 17 (d) seldom 5

9. Do you feel that the proportion of items "of interest" which were also "new" to you was:

(a) satisfactory 54 (b) too low 17

10. Would you say that, in general, information given for the entries in the SDI lists is adequate to judge whether an item is or is not of interest to you?

(a) yes 58 (b) no 10

11. What elements of the entry did you most often find useful in making evaluations?

(a) author/editor 38 (b) title 55 (c) publisher 9 (d) series note 4 (e) subject headings 32 (f) classification numbers 8 (g) other (please specify) 1

12. What is the primary use to which you put the SDI information?

(a) recommendation for library acquisition 51 (b) personal purchase of item 12 (c) other (please specify) 15

13. If your recommendation originates the library order for a publication, it will be some time before the work is available; and even if already on order, most of the publications included in your lists were probably too new to be available from the library at the same time you received the list. Do you feel that this diminishes the value of the SDI service?

(a) significantly 2 (b) somewhat 20 (c) negligibly 45

For what reasons?

14. A potential value of SDI service, based on the large volume of newly published works catalogued by and for the Library of Congress, is to bring together in one list timely notices for those works in the file which correspond to your several fields of interest. Do you feel that the experimental SDI service demonstrated this capacity?

(a) amply 26 (b) adequately 36 (c) poorly 9

15. Is the format of the SDI notices satisfactory?

(a) yes 61 (b) no 9

If not, what format would you suggest?

16. Is the distribution schedule of once a week satisfactory?

(a) yes (b) no
71 0

17. On the average, how much time would you estimate it took to examine an SDI list? Roughly:

Minutes: (a) 5 (b) 5-10 (c) 10 (d) 10-15 (e) 15 (f) 15-20 (g) 20
23 16 9 11 5 1 5

18. A possible by-product of this SDI service is the building up of a cumulative MARC tape file which can be searched in various ways by computer. Would you make use of such a file?

(a) yes (b) no
40 18
If no, for what purposes?

19. Judging from your total experience with the SDI service, would you characterize its overall value to you as:

(a) very high (b) high (c) medium (d) low
8 24 30 9

20. The MARC file at present represents English monographs catalogued by the Library of Congress on a week by week basis. Sometime in 1972, the Library of Congress will begin to add some non-English monographs to the MARC file. Keeping in mind the forthcoming expanded MARC file on which future SDI service would be based, do you feel that its value to you would then be:

(a) increased (b) the same (c) less
REASONS: 30 33 7

21. Do you personally want this SDI service to be continued?

(a) yes (b) no (c) it doesn't matter
42 3 5

22. Do you feel that this SDI service should be offered to the entire faculty?

(a) yes (b) no
42 14
REASONS:

23. Do you feel that this SDI service should appropriately be made available by the university, i.e. that the university should organize and administer the service?

(a) yes (b) no (c) don't know
36 5 23

24. Do you feel that the university alone should pay for this faculty SDI service:

(a) yes (b) no (c) don't know
30 6 25

25. Optional: General comments, pros and cons, elucidation of above replies, attitudes, suggestions, etc. conerning the SDI service.

TITLE: Proceedings of the Western Canada Chapter of ASIS
(3 rd annual: October 3-5, 1971: Banff, Canada)

ABSTRACT: The proceedings contain papers given by the members of the chapter who come from both the University and Business environments. Some operational indexing, bibliographic, SDI and Retrospective Search Systems which include CAN/SDI, Compendex, TEXT-PAC, SIS II & III, KWOC and FAMULUS are discussed. Also included are papers on two projects conducted by the Computing Science department of the University of Alberta; the one project is an on-line thesaurus and the second an Information Retrieval Laboratory. Other papers are about the computerized circulation system at the University of Calgary's library, the Marc project at the University of Saskatchewan and the problems of design and coding questionnaires.