DOCUMENT RESUME

ED 072 794 LI 004 147

AUTHOR Tell, Bjorn V.; And Others

TITLE The Use of ERIC Tapes in Scandinavia, Searching With Thesaurus Terms in Natural Language.

INSTITUTION Council of Europe, Strasbourg (France). Council for Cultural Cooperation.; Royal Inst. of Technology, Stockholm (Sweden).

REPORT NO ECS-DOC-72-15

PUB DATE 11 Nov 72

NOTE 24p.; (0 References); Ad Hoc Committee for Educational Documentation and Information. EUDISED Project

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS *Data Bases; Foreign Countries; *Information Dissemination; *Information Retrieval; *Information Services; Information Utilization; Relevance (Information Retrieval); *Search Strategies

IDENTIFIERS *Educational Resources Information Center; ERIC; ERIC Tapes; Scandinavia; SDI; Selective Dissemination of Information

ABSTRACT

Since February 1971 the Royal Institute of Technology, Stockholm, has been running the ERIC data base mainly for SDI purposes. The implementation of the data base into the generalized search system, ABACUS, is described. One hundred and fifty-eight users received SDI service at present, 99 from governmental and educational institutions, 23 from industry, and 36 from abroad. Retrospective searches have also been made. Two methods of matching users to documents have been employed - the controlled vocabulary of the Thesaurus of ERIC Descriptors, and the free text words of titles. Some users' assessments of the relevance of the output have been gathered, and examples are given of query formulation into profiles and the resulting printout of references. A number of the profiles for the ERIC data base have also been run on data bases such as ISI, INSPEC and COMPENDEX. The practice of writing profiles which contain term types which are appropriate to only one of several data bases, against which they are searched, is discussed. A frequency list of words in the titles has been compiled in order to use a weighting procedure for sorting the printout. A training program for acquainting the user with this new service is needed. (Author/NH)

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

STRASBOURG, 11th November 1972

DECS/DOC (72) 15

COUNCIL FOR CULTURAL CO-OPERATION

AD HOC COMMITTEE FOR EDUCATIONAL DOCUMENTATION AND INFORMATION

EUDISED PROJECT

THE USE OF ERIC TAPES IN SCANDINAVIA, SEARCHING WITH THESAURUS TERMS IN NATURAL LANGUAGE

by

Björn V. Tell, Kerstin Wessgren and Winnie Hemborg

Royal Institute of Technology
Stockholm

SUMMARY

Since February 1971 the Royal Institute of Technology, Stockholm, has been running the ERIC data base mainly for SDI purpose. The implementation of the data base into the generalized search system, ABACUS, is described. 158 users receive SDI service at present, 99 from governmental and educational institutions, 23 from industry, and 36 from abroad (Finland 26, Norway 8, Switzerland 1 and the United Kingdom 1). Retrospective searches have also been made.

Two methods of matching users to documents have been employed for the ERIC data base - the controlled vocabulary of the Thesaurus of ERIC Descriptors, and the free text words of titles. Some users' assessments of the relevance of the output have been gathered, and examples are given of query formulation into profiles and the resulting printout of references.

A number of the profiles for the ERIC data base have also been run on data bases such as ISI, INSPEC and COMPENDEX. The practice of writing profiles which contain term types which are appropriate to only one of several data bases, against which they are searched, is discussed. A frequency list of words in the titles has been compiled in order to use a weighting procedure for sorting the printout in an helpful order. A training programme for acquainting the user with this new service has been needed. However the present results show that a great number of users have found it of interest to use the SDI service in their work. On the other hand many of the queries have also had to be searched on other data bases in order to assure a reasonable coverage.

1. INTRODUCTION

The ERIC data base is run by the Royal Institute of Technology Library. Usually, the library functions are those of acquisition, cataloguing, storage and circulation. How did it happen then that the Institute considered it within its scope to include machine-readable data bases such as ERIC, and provide an information service based on them? Why should the library offer an information service which was not otherwise available? Could it justify the costs of acquiring and maintaining mechanized data bases and the computer operations? This paper will try to answer these questions.

The Swedish government has taken an active interest in developing a policy for economic growth. In 1967 it launched a programme for the promotion of technological development and industrial growth, and a plan for the development of scientific and technical information was included. The government was especially interested in studying the viability of mechanized information services in the field of science and technology, and the utility they could offer to users in research and industry. The Institute library was chosen as the agency responsible for the establishment of a mechanized service for users in science, industry and education.

The requirements for the computer operation of a service had been thoroughly studied during Tell's years as department manager of the Swedish nuclear establishment, AB Atomenergi. Then in 1967 the Institute library received a first grant of 80,000 Sw.Cr. ($ 16,000) to initiate a computerized service in the field of mechanical engineering. Over the years the scope has extended and the grant has increased, and it has now stabilised around 1 Million Sw.Cr. outside the ordinary budget of the library. Half of that sum goes to the salaries for documentalists who have been added to the library staff. Thus, the fundamental requirements for staff and funds have been fulfilled by the new policy.

2. THE ORGANISATION OF A NEW COMPUTERIZED SERVICE

In order to start a computerized service the best choice, at least at the time, seemed to be a current awareness service - SDI - Selective Dissemination of Information. SDI is a system developed by the late Hans Peter Luhn at IBM in 1959 for alerting participants about new publications such as journal articles, reports, conference papers etc. The acronym SDI has the special connotation that the process makes use of a computer. This is possible when the references to the literature are stored on magnetic tape.

TABLE I.

The Reformatting of ERIC Report Resume Master Data Set Fields into the ABACUS Format.

ERIC		ABACUS		
Field name	Field identification no. in hexadecimal	Searchable	Printout	Deletion
Sequence	0000			X
Add Date	0001			X
Change Date	0002			X
Accession Number	0010		X	
Clearinghouse Accession Number	0011		X	
*Other Accession No.	0012			X
*Program Area	0014			X
*Publication Date	0017		X	
Title	001A	X	X	
Personal Author	001B	X	X	
*Institution Code	001C			X
*Sponsoring Agency Code	0020			X
Descriptor	0023	X		
Identifier	0024			X
*EDRS Price	0025			X
*Descriptive Note	0026			X
Issue	002B		X	
Abstract	002C			X
*Report Number	002D		X	
*Contract Number	002E			X
*Grant Number	002F			X
*Bureau Number	0030			X
*Availability	0031			X
Journal Citation	0032	X	X	
*Institution Name	0080	X		
Sponsoring Agency Name	0084			X

* Not Used in CIJE

3. IMPLEMENTATION OF THE ERIC DATA BASE INTO ABACUS

The basic approach employed has been to use a general processing format into which a record of a particular output such as the ERIC files can be converted by a reformatting program so that its records can be searched. Thus, the search routine will be the same as for records of other systems similarly converted by individual reformatting programs.

The success of this pragmatic approach to the compatability problem of various tape formats greatly depends upon the hospitality of the search record format. The ABACUS was designed in 1966, before the MARC pilot program and the interchange format reflected in International Standard ISO/DIS 2709 (Coward 2) which is foreseen as the standard for UNISIST and EUDISED. However, the ABACUS record has many characteristics in common with MARC and ISO. A directory to the whole record maps out the record length, the data elements present, and the number of characters in each element. The directory is a fixed field header followed by variable data fields. The fixed fields give the address to, and the length of the variable fields. The items of interest in the external data base are selected, and fields in the ABACUS format are allocated by the reformatting program. Depending on the amount of information on the external tape, the identification process differs from one format to another.

The most extensive format in ERIC is the Report Resumé Master Data Set, many of which fields are not applicable in the shorter format of Journal Article Master Data Set. Not all fields are of interest to the Scandinavian users. Thus, at present, some fields are deleted when reformatting into the ABACUS. Table I shows the ERIC Report Resumé Master Data Set Fields and their treatment in the ABACUS record. Even if documentation is provided by a data base producer, the reformatting specification is written after inspection of tape dumps.

In general, the reformatting of the ERIC tape formats was a rather straightforward job of 30 hours programming, even if they deviated from the ISO interchange format into which, it is hoped, they will eventually change. ERIC files in their present form are grouped in variable length blocks, the maximum length of which is included in the label. The first two bytes of each block specify the length of the block. Similarly, the two bytes of each record specify the length of the record. Within each record, the first two bytes of each field specify its length and the third and fourth are the field identification number. Essentially, the allocation of fields in the ABACUS program depends on the field identification numbers within the two ERIC record types for reports and articles. As can be seen from Table I, the 26 fields in the ERIC format yield 5 fields in the ABACUS set of searchable fields. The search terms can operate within these, since they are specified with regard to the type of field in which they are to be searched.

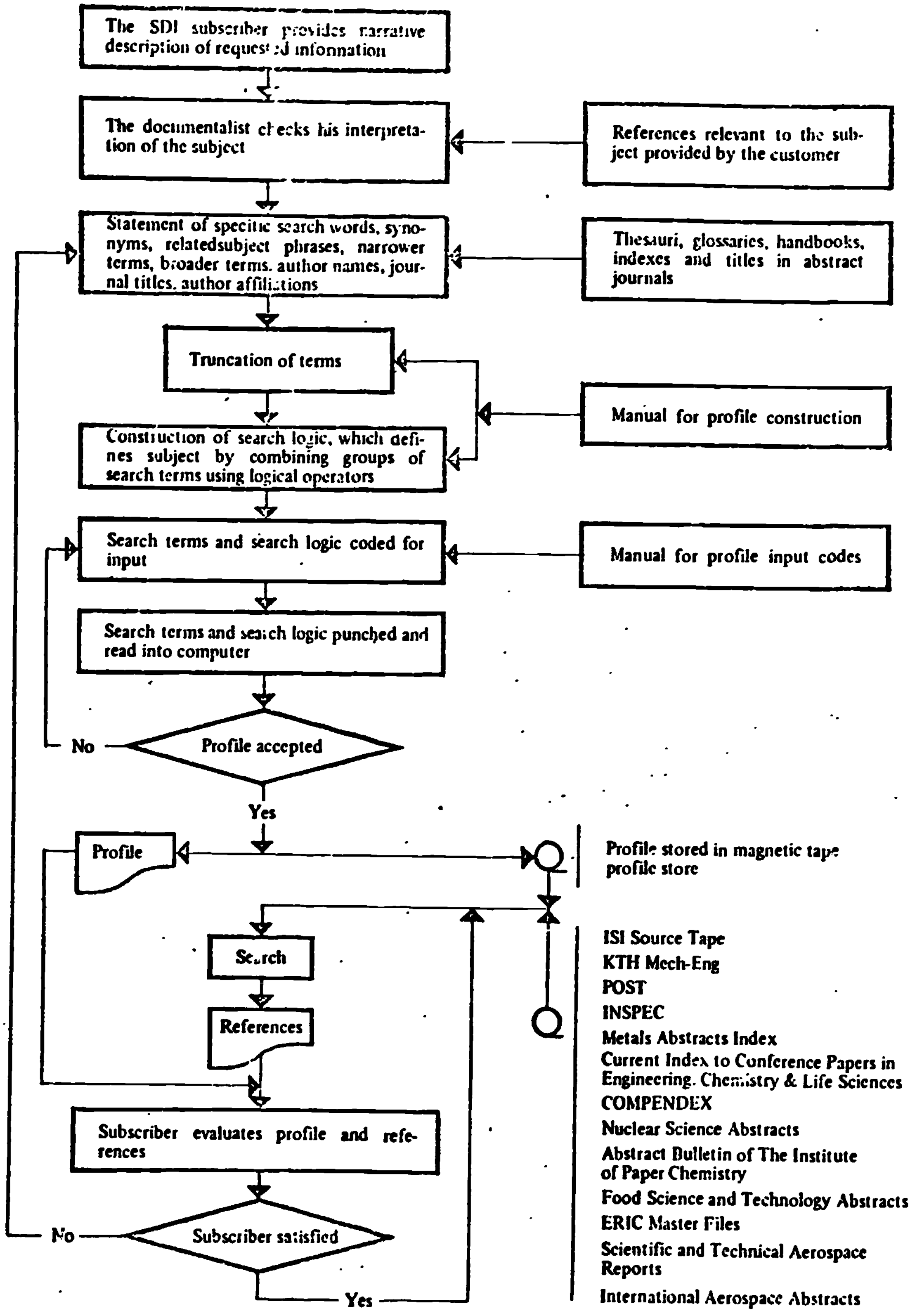
4. PROFILE CHARACTERISTICS

The construction and revising of query profiles is another essential task in an SDI system which demands an effort from the user and the subject specialist. When a user wants to submit a question to the SDI system he is requested to formulate his field of interest in natural language which means in a normal narrative way, describing his interest in some detail. It has proved very useful for the user also to supply some references to papers which he considers relevant to his query. He could also provide a list of significant terms, and if possible, make a draft of the actual search profile. The staff have prepared a Profile Design Manual which explains the principles of a computer-operated information retrieval system and describes all details of the profile construction.

Within the research and development programme of the Swedish National Board of Education the introduction of the ERIC tapes has also resulted in a report by Bernhard Bierschenk. Under the title "To Search for Literature by Means of Computers" (in Swedish), the author elaborates upon the information and communication problems in education. The construction of profiles is given a broad treatment, and much of it has been copied from the Profile Design Manual. Tools of this nature help the user to visualize the usage he can make of the computerized service and aid him in developing his individual profile.

Fig. 1

GENERALIZED FLOW CHART FOR PROFILE CONSTRUCTION


The SDI subscriber provides narrative description of requested information
The documentalist checks his interpretation of the subject
References relevant to the subject provided by the customer
Statement of specific search words, synonyms, relatedsubject phrases, narrower terms, broader terms. author names, journal titles. author affiliations
Thesauri, glossaries, handbooks, indexes and titles in abstract journals
Truncation of terms
Manual for profile construction
Construction of search logic, which defines subject by combining groups of search terms using logical operators
Search terms and search logic coded for input
Manual for profile input codes
Search terms and search logic punched and read into computer
No
Profile accepted
Yes
Profile
Profile stored in magnetic tape profile store
Search
References
Subscriber evaluates profile and references
No
Subscriber satisfied
Yes
ISI Source Tape
KTH Mech-Eng
POST
INSPEC
Metals Abstracts Index
Current Index to Conference Papers in Engineering. Chemistry & Life Sciences
COMPENDEX
Nuclear Science Abstracts
Abstract Bulletin of The Institute of Paper Chemistry
Food Science and Technology Abstracts
ERIC Master Files
Scientific and Technical Aerospace Reports
International Aerospace Abstracts

Z.Gluchowicz

The interaction between the staff and the user is essential for a successful search. On the basis of the user's statements the subject specialist specifies the question by making a list of significant terms, either from the ERIC Thesaurus, or terms which might occur as potential words in the titles of documents. Among the staff there are subject specialists in education, psychology, business administration, etc. Furthermore, the list might also include authors, affiliations, and journal titles. As the system permits search both on keywords and on natural language used in titles, the subject specialist uses thesauri, handbooks, dictionaries and all other means he might find helpful and relevant for the formulation of the profile. He has to make a special point of checking the printed volumes of Current Index to Journals in Education, and Research in Education, and other appropriate sources to find the occurrence of terms when used alone or in combination with other terms. A generalized flow chart, Figure 1, has been constructed by Zofia Gluchowicz (3).

While the keywords must be written exactly as they appear in the Thesaurus and on the tape, the free text terms in potential titles can be truncated both at the beginning and at the end. Truncation facilitates retrieval of items containing word fragments which are common to different forms of a word, and words within words can be searched for. As will be seen from the examples below, suffix (right-hand) truncation occurs very often, while prefix (left-hand) truncation is more unusual. Combined suffix and prefix truncation is, on the other hand, more common. For example, the truncated term /CASSETT/, where the slashes stand for truncations, will retrieve STERIOCASSETTES, VIDEOCASSETTE, CASSETTE-RECORDER, CASSETTE/CARTRIDGE, etc.

As can be seen from Figure 2, the terms are numbered sequentially in the profile printout to facilitate updating. The terms are also grouped together, and the groups are indicated by capital letters A, B, C, etc. Terms, or groups of terms, are linked together in a logical manner by using "and", "or", and "not" logic. The number of terms in one profile might be up to the system-allowed 150 in ABACUS. In the new VIRA program there are no such restrictions. On the other hand, as charging policy is to count 30 terms as one profile, the average number of terms per profile varies around 24.

The printout of the profile even includes a description in natural language of the query, the search logic, and the list of terms classified according to type of terms such as words, keywords, author names etc. The profile printout and every updating of it is sent to the user. For verification a copy of the profile as well as a copy of the search results are kept in the files of the service, transferred every 9 months into microfilm cassettes.

The user's responses to early selections based on the first profile approximation to his field of interest are used for improving the profile. Thus the maintenance of the profile is carried out by adding new terms and deleting old ones which do not give satisfactory results, or by opening and tightening the logic. False co-ordinations between search terms from different term groups can also be detected and should be avoided.

While constructing the initial profile we try to choose the logical strategy considering the user's wishes, and accordingly decide on the degree of restrictivity for the initial computer run. Often we use a less restrictive logic, i.e., not too many "and" or "not" restrictions, in the initial profile, even if it will result in an output of many irrelevant references, i.e., noise, and then after a few searches adjust the profile on the basis of the user's evaluation of the output.

5. PROCESSING METHODS AND COSTS

An inevitable characteristic of large retrieval systems is that a strategy for searching a small or medium size data base might differ significantly from a search strategy for a large base. During the years our search methods have passed through the mere masking-off technique, yielding search times proportional to the number of references and terms in the profiles, into a more elaborate technique making use of hashcoding and tree structure searches, thus arriving at an almost logarithmic increase in time when the number of terms in the profiles grows. The newest program, having the acronym VIRA and written by Rolf Larsson, is run in parallel with ABACUS (Zennaki 4). The present profile program, PROSA, includes 2,500 statements in COBOL, and the VIRA search program counts 2,000 statements in IBM assembler language.

Figure 2.

Profile 70E
Subject: Audiovisual aids for the mentally retarded
Data bases: ERIC, ISI, INSPEC
Logic: A & B

Term No.	Term Group	Search terms	Weight	Term Type
010	A	TAPE RECORD/	2	KEYWORD
020	A	VIDEO TAPE RECORD/	2	KEYWORD
030	A	EDUCATIONAL TELEVISION/	2	KEYWORD
040	A	INSTRUCTIONAL TELEVISION/	2	KEYWORD
050	A	AUDIOVISUAL/	10	KEYWORD
060	A	CASSETT/	2	WORD
070	A	CARTRIDGE/	2	WORD
080	A	EVR	2	WORD
090	A	VTR	2	WORD
100	A	VCR	2	WORD
110	A	ETV	2	WORD
120	A	ITV	2	WORD
130	A	CTV	2	WORD
140	A	SELECTAVISION/	2	WORD
150	A	TELEVISION	2	WORD
160	A	TV	2	WORD
170	A	/VIDEO/	2	WORD
180	A	CARTRIVISION/	2	WORD
190	A	8MM/	2	WORD
200	A	AUDIOVISUAL/	10	WORD
210	A	AV	10	WORD
220	A	A-V	10	WORD
230	A	VIDICORD/	10	WORD
240	A	VISUAL AID/	10	WORD
250	A	MEDIA/	2	WORD
260	A	PICTURE/	2	WORD
270	A	LONG-DISTANC/	10	WORD
280	A	AUDIO-VISUAL	10	WORD
290	B	EDUCATIONALLY DISADVANTAG/	2	KEYWORD
300	B	LOW ABILIT/	2	KEYWORD
310	B	SLOW LEARNER/	2	KEYWORD
320	B	MENTALLY HANDICAP/	10	KEYWORD
330	B	EDUCABLE MENTALLY HANDICA/	10	KEYWORD
340	B	RETARDED/	10	KEYWORD
350	B	RETARDATION/	10	KEYWORD
360	B	MENTAL RETARDATION/	10	KEYWORD
370	B	EXCEPTIONAL/	2	KEYWORD
380	B	SPECIAL/	2	KEYWORD
390	B	RETARD/	10	WORD
400	B	LOW/	2	WORD
410	B	SLOW/	2	WORD
420	B	FAILUR/	2	WORD
430	B	DISADVANTAG/	2	WORD
440	B	HANDICAP/	2	WORD
450	B	BELOW/	10	WORD
460	B	EXCEPTION/	2	WORD
470	B	DROPOUT/	2	WORD

The ERIC files have been run with the ABACUS program for the SDI service, while for retrospective searches the combination of ABACUS and VIRA has proved more economical. During 18 months, from February 1971 to the time of writing, ten SDI runs and one retrospective search, divided into three batches, have been performed. The search statistics for these runs are shown in Tables II and III. It is obvious that the VIRA program is more efficient when e.g., Run 2 in Table II is compared with Batch B in Table III. In economic terms, the search costs of VIRA are less than one-quarter of those of ABACUS taking account of the present pricing structure per hour for the two computers used, IBM 360/30 and 360/75. Table III gives the computer time for the retrospective search. A sequential search of a large data base is often regarded as excessively time consuming without compression methods. However, the VIRA program permits such searches to be made economically. The search time for about 60,000 ERIC records using 42 profiles containing 1,137 search terms was less than 20 minutes CPU time and 40 minutes input/output time. If conversion and printout time is added, the costs of the data processing amount to 3,100 Sw. Cr. ($ 620), or per profile 74 Sw. Cr. Since the price charged to the user was equivalent to the price level set by the European Space Research Organization, ESRO, for searching their files, 300 Frs, the balance also covered part of the cost of constructing the profiles.

In order to carry out a rough check of the performance of the profiles on a "management by exception" basis, two statistical tools have been developed. The critical values of the printout to a user are (1) an over-abundance of references, and (2) no printout. In order to reveal these extremes, every search results in search statistics indicating the number of references for each profile. The form is designed like the scale of the speedometer of many cars, the longer the row of "stars" the more the reason to put ones foot on the brake. Figure 3 displays part of the search statistics for Run 7 of ERIC. The columns give the number of references to the first digit, the second, etc. Thus, the first profile has resulted in 6 + 40 = 46 references; the second in 8 + 60 + 300 = 368 references. On the other hand, profile no. 26R has given no output. Furthermore, at the bottom of the form an indication is given of which profiles have received no hits, and those which have received more than 40 hits.

These search statistics give an indication of where the exceptional cases are located among the profiles. The next step is to analyse what causes the no-hits or the great number of hits. In order to find out about the latter case, a listing is also given for every profile stating which terms or term combinations have caused the printout, together with the frequencies of these terms. See Figure 4. In this case the first step would be to analyse the combination MEASUREMENT TECHNIQUES and MEASUREMENT INSTRUMENTS which occurs 13 times, perhaps in order to change the logic or to place these words in separate groups, if they have given rise to many irrelevant references. The second column in Figure 4 indicates the weights we are experimenting with which will be discussed later on.

6. SEARCHING KEYWORDS AND WORDS IN TITLES

The ABACUS program is designed in such a way that it can process natural language by searching titles or abstracts. In the case of another data base, Science Citation Index Source Tapes, the ISI tapes, which among 2,000 journals includes around 80 core journals in the field of education and psychology, there are no keywords or subject indicators other than the titles. Thus free text search is the only way to open the files. Free text search can be regarded as using a set of skeleton keys to open up any machine readable file.

The ERIC files make use of keywords chosen from the Thesaurus of ERIC Descriptors. Searching these keywords becomes an additional means for the subject specialist or the user to augment the search performance of the ERIC files compared with the ISI tapes. When a data base contains keywords we have recommended that they be used in combination with words in natural language.

In a multi-data-base environment the same profile in natural language can easily be used on various data bases, while the use of keywords is restricted to each specific data base which has to be taken into account when formulating the profile. Many of our profiles on the ERIC tapes are also processed on the ISI, INSPEC and COMPENDEX tapes, since our main principle is to answer the query in its broadest sense disregarding from which data base the responding references will stem.

TABLE II.

ERIC Search Statistics for SDI Feb. 1971 - July 1972 with ABACUS

Run No.	Data Base	No. Eric Records	No. Profiles	No. Answers	Search Time 360/30 CPU Minutes	Search Time per Profile Minutes
1	RIE	5402	38	2356	94	2.5
2	CIJE	19427	38	3360	332	8.7
3	RIE	2473	65	1514	185	2.8
4	CIJE	4036	70	2793	218	3.1
5	RIE	3982	72	4699	322	4.5
6	CIJE	4263	82	3062	214	2.6
7	RIE	5874	87	6095	341	3.9
8	CIJE	4204	92	3099	257	2.8
9	RIE	2867	132	4497	285	2.2
10	CIJE	4144	134	5269	310	2.4

TABLE III.

ERIC Search Statistics for one Retrospective Search in Batches with VIRA

Batch No.	No. Eric records	No. Profiles	No. Answers	Conversion Time 360/30 CPU Minutes	Search Time 360/75 Minutes I/O	CPU	Print Time Minutes
A	12,285	42	10,097	72	14	7	49
B	19,919	42	6,696	60	8	4	24
C	27,575	42	9,677	103	17	8	33
	59,779	42	26,470	235	39	19	106

FIGURE 3. SEARCH STATISTICS FROM RUN 7

```
0131.******    .****      .          .
02E1.********* .******    .***       .
02U1.***       .*         .          .
07F1.          .**        .          .
08F1.***       .          .          .
10A1.**        .********  .****      .
11E1.          .********  .*         .
12F1.*         .****      .          .
13F1.**        .****      .**        .
14F1.*****     .*         .*         .
15F1.********* .          .          .
16R1.**        .*         .          .
19C1.*******   .******    .          .
26R1.          .          .          .
28E1.****      .********  .          .
31C1 ***       .          .          .
3141.******    .**        .          .
32S1.********  .**        .          .
36D1.*         .*         .**        .
36F1.********  .***       .          .
38E1.          .****      .**        .
38F1.**********.*         .          .
39F1.*****     .***       .          .
40G1.******    .**        .          .
41F1.***       .********  .          .
44G1.          .          .          .
45B1.****      .          .          .
45D1.**        .*****     .          .
45F1.;******** .*         .***       .
45G1.*******   .*****     .          .
51A1.*******   .          .**        .
51F1.***       .          .**        .
52F1.*         .*         .          .
54F1.*****     .********  .          .
5411.*******   .********* .          .
56E1.**        .****      .          .
56F1.****      .**        .          .
5631.******    .***       .*         .
57E1.          .***       .          .
58E1.**        .***       .          .
58R1.*********.*******    .          .
59E1.*****     .*         .          .
```

```
FÖLJANDE PROFILER ERHÖLL INGA TRÄFFAR
26R1   44G1   6241   70F1   80A1   8851

FÖLJANDE PROFILER ERHÖLL MER ÄN 40 TRÄFFAR
0131   02E1   10A1   11E1   12F1   13F1   14F1   19C1   28E1   36D1   38E1
41F1   45D1   45F1   45G1   51A1   51F1   54F1   5411   56E1   5631   58R1
```

FIGURE 4. FREQUENCIES OF COINCIDENCES OF PROFILE 64G IN RUN 7

```
 1   VIKT=30,00   * BEHAVIOR* CLASSROOM OBSERVATION TECHNIQUE*
 2   VIKT=30,00   * BEHAVIOR* CLASSROOM* CLASSROOM OBSERVATION T
 1   VIKT=30,00   * BEHAVIOR* PERFORM* CLASSROOM* TEACH* CLASSRO
 1   VIKT=30,00   * BEHAVIOR* PUPIL* CLASSROOM OBSERVATION TECHN
 3   VIKT=30,00   * BEHAVIOR* TEACH* CLASSROOM OBSERVATION TECHN
12   VIKT=30,00   * CLASSROOM OBSERVATION TECHNIQUE*
 5   VIKT=30,00   * CLASSROOM* CLASSROOM OBSERVATION TECHNIQUE*
 1   VIKT=30,00   * CLASSROOM* TEACH* CLASSROOM OBSERVATION TECH
 2   VIKT=30,00   * EDUCATION* CLASSROOM OBSERVATION TECHNIQUE*
 2   VIKT=30,00   * OBSERVATION* CLASSROOM OBSERVATION TECHNIQUE
 2   VIKT=30,00   * PUPIL* TEACH* CLASSROOM OBSERVATION TECHNIQU
 1   VIKT=30,00   * TEACH* BEHAVIOR* CLASSROOM OBSERVATION TECHN
 5   VIKT=30,00   * TEACH* CLASSROOM OBSERVATION TECHNIQUE*
 3   VIKT=30,00   * TEACH* METHOD* CLASSROOM OBSERVATION TECHNIQ
 1   VIKT=30,00   * TECHNIQUE* CLASSROOM OBSERVATION TECHNIQUE*
 1   VIKT=32,00   * OBSERVATION* TEACH* METHOD*
 1   VIKT=35,00   * OBSERVATION* EDUCATION* EVALUATION TECHNIQUE
 1   VIKT=41,00   * OBSERVATION* BEHAVIOR* MEASUREMENT TECHNIQUE
 2   VIKT=50,00   * ACHIEVEMENT* MEASUREMENT TECHNIQUES * MEASUR
 1   VIKT=50,00   * EDUCATION* TEACH* TECHNIQUE* MEASUREMENT TEC
 1   VIKT=50,00   * INSTRUMENT* MEASUREMENT TECHNIQUES * MEASURE
13   VIKT=50,00   * MEASUREMENT TECHNIQUES * MEASUREMENT INSTRUM
 1   VIKT=50,00   * TEACH* MEASUREMENT TECHNIQUES * MEASUREMENT
 1   VIKT=50,00   * TECHNIQUE* MEASUREMENT TECHNIQUES * MEASUREM
 1   VIKT=52,00   * ACHIEVEMENT* MEASUREMENT TECHNIQUES * EVALUA
 1   VIKT=52,00   * CLASSROOM* MEASUREMENT INSTRUMENTS * EVALUAT
 1   VIKT=52,00   * INSTRUMENT* MEASUREMENT INSTRUMENTS * EVALUA
 1   VIKT=52,00   * INSTRUMENT* MEASUREMENT TECHNIQUES * EVALUAT
 3   VIKT=52,00   * MEASUREMENT INSTRUMENTS * EVALUATION TECHNIQ
 4   VIKT=52,00   * MEASUREMENT TECHNIQUES * EVALUATION TECHNIQU
 1   VIKT=52,00   * STUDENT* MEASUREMENT TECHNIQUES * EVALUATION
 1   VIKT=52,00   * TEACH* TECHNIQUE* MEASUREMENT INSTRUMENTS *
 2   VIKT=57,00   * EVALUATION TECHNIQUES * CLASSROOM OBSERVATIO
 1   VIKT=57,00   * TEACH* EVALUATION TECHNIQUES * CLASSROOM OBS
 1   VIKT=57,00   * TEACH* STUDENT* BEHAVIOR* EVALUATION TECHNIQ
 1   VIKT=60,00   * OBSERVATION* TEACH* BEHAVIOR* CLASSROOM*
 1   VIKT=70,00   * OBSERVATION* TEACH* STUDENT* CLASSROOM OBSER
 1   VIKT=80,00   * OBSERVATION* MEASUREMENT TECHNIQUES * MEASUR
 1   VIKT=82,00   * CLASSROOM* MEASUREMENT INSTRUMENTS * EVALUAT
 1   VIKT=86,00   * OBSERVATION* CLASSROOM* TEACH* METHOD* CLASS
```

Especially for questions of inter-disciplinary nature it is obvious that they should be processed on several data bases in order to assure good coverage. It is true, however, that the reformulation of a query into a profile for the SDI system takes place in a kind of dialogue with the computer, focusing on one data base at a time and considering both the terminology used in free text and the metalanguage of keywords or other subject indicators. In order to arrive at a standardization of the query formulation, allowing for different degrees of complexity of natural text and metalanguages, one method would be to develop a translation system between the various scientific disciplines reflected in the data bases by the generation of vocabularies and concordances for words in natural language and the various thesauri used.

We have started work in this area by the compilation of word frequency lists for various data bases. Thus, two years of ERIC CIJE 1969-70 tapes containing [illegible]7 575 references have been processed in order to compile an alphabetic and a frequency ordere[illegible] used in the titles +. Out of the 110,000 word occurrences, 10,642 different words were reco[illegible] The non-informative words like:

A OF T.IE AND IN FOR TO ON AN
AS WITH AT BY FROM OR SOME IS

account for 24 per cent of all word occurrences. The following 20 words account for another 10 per cent (frequencies are given in parentheses):

EDUCATION (AL)	(1900)	CHILDREN	(614)	RESEARCH	(370)
SCHOOL (S)	(1287)	TEACHING	(493)	SOCIAL	(356)
PROGRAM (S)	(742)	LEARNING	(394)	LANGUAGE	(328)
TEACHER (S)	(708)	TRAINING	(392)	CURRICULUM	(304)
STUDY(IES)	(707)	COLLEGE	(380)	EVALUATION	(304)
REPORT	(671)	DEVELOPMENT	(380)	FINAL	(303)
STUDENT (S)	(587)	READING	(373)		

It should be noted that the first significant word in the list is EDUCATION (AL) which has a frequency placing it between the two prepositions FOR and TO (2053 and 1295). The information value of these 20 words in the ERIC Thesaurus, in which all occur except for the last word FINAL, could be questioned. The word EDUCATION (AL) is found in 7 per cent of the document titles.

That the use of the language (the scientific "jargon") differs between various disciplines was illustrated when compiling frequency lists for other disciplines. Thus, for instance, the first significant word in the INIS system - nuclear energy - was REACTOR, and the first in CAC - organic chemistry - ACID. The non-informative words mentioned above occur in almost the same order in these data bases as they do in ERIC.

Of the 116 search profiles in Run 9 on ERIC RIE, 65 were also searched on the ISI tapes, and 34 of these latter also on INSPEC tapes. 8 profiles contained only ERIC keywords, 67 both keywords and free text words, and 39 were searched exclusively with free text words. On the average the keyword profiles contained 20 keywords. The mixed profiles had 11 keywords and 18 words, or together 29 search terms. The free text profiles contained 39 words. In total 2,529 words and 930 keywords were used in this run.

+ See Appendix for a brief note on the merged frequency list of words in the bibliographic references to [illegible] 917 FRIC reports and 27,573 journal articles.

TABLE IV.

ERIC Search Profiles Organized into Broad Categories

Subject Field	No. of Search Profiles
1. Administration	8
2. Communication; Methods and Characterstics	12
3. Counselling	1
4. Curriculum	2
5. Education and Instruction; General Education Concepts Specific Types of Education Instructional Techniques, -methods, -equipment	41
6. Evaluation; Evaluation Techniques Tests and Measurement	11
7. Health and Safety; Recreation	21
8. Language and Speech	4
9. Library Science	5
10. Psychology; Learning and Cognition Development Behavior Attitudes Adjustment	44
11. Sociology; Environment Socialization Social Relations	9
Total	158

The 42 profiles which participated in the retrospective search contained 498 keywords and 639 words. 4 profiles included only keywords, 36 had keywords and words mixed, and 2 were formulated in natural language words. The average number of keywords in the keyword profiles were slightly higher than in SDI, namely 26 keywords. The mixed profiles had 28 search terms, and the natural language profiles 32 terms on the average.

How efficient the keywords of the ERIC Thesaurus are can be judged, in a way, from the examples given below. In many cases the words out of titles perform equally well or even better. This confronts EUDISED indirectly with what Jean Viet(5) calls the fundamental question of whether it is really necessary to have a thesaurus at the input end.

The following remarks based upon our experience might illuminate this question. The combined search strategy we use, mixing keywords and words in free text, reveals that the present indexing habit in ERIC of using keywords identical to words in the titles is futile. If some of the keywords instead took the place of broad subject categories like those we have used for subdividing the user population in the following Chapter, it would add a new dimension to the search. This is, for instance, the case with another data base, the INSPEC.

A study should also be made about the proportion of titles that are not useful as content indicators and, thus, not suitable for free text searching. If only a small proportion of titles are meaningless, manual indexing using thesaurus keywords should be questioned.

On the other hand, if something needs to be done, especially if we believe that keyword indexing is necessary for the quality of printed indexes or for future on-line retrieval systems of the RECON type, title augmentation or automated keyword assignment seem to be attractive alternatives to expensive manual indexing. Such a strategy might lead authors to improve the information content of their titles. This has happened in areas where the KWIC indexing technique is used.

Because of the cost of indexing we could never afford it for our own data base in mechanical engineering, wood, paper and pulp industry, covering 250 journals (60,000 references/yr) in three languages. Only title augmentation is permitted in the case of short titles (less than 60 characters). We know that we can give satisfaction to the users by free text searching only, because at present we receive orders for several hundred photocopies a month as a result of the output. It might be that the vocabulary is more stabilized in these fields than in education, a case to investigate.

7. EVALUATION AND FEED-BACK

At present 158 users receive SDI service on ERIC, 99 from governmental and educational institutions, 23 from industry, and 36 from abroad (Finland 26, Norway 8, Switzerland 1 and the United Kingdom 1). A breakdown of the profiles into subject categories is found in Table IV.

After five years of operation on tapes in general, and 18 months on ERIC, we feel that we are still just scratching the surface of computerized information retrieval. We think, for instance, that the printout we now deliver as answers to the queries should go through further refinement before reaching the user.

When we consider the construction of a profile as reflecting a specific query, it is difficult to provide a measure of its effectiveness, especially as our practice is to retrieve references from multiple files. Questions about recall and precision lose interest. The essential measure which we can assess is the user's satisfaction, which can be expressed on a scale from highly relevant to irrelevant, or by counting the number of documents he orders. Time and costs of the computer are other factors which can be measured, and for ERIC we break even, more or less, between computer costs plus the costs for the tapes and the subscription fees for the profiles, leaving other costs, e.g. the construction of the profiles to be defined as common library costs.

The ability of the system to adapt to the expectations of a particular researcher in education is illustrated by the following examples of some rather simple profiles.

Example 1

The feedback problem is of interest to many in the educational field. The following profile has been constructed in order to cover the specific interest of a user attached to the university training centre at Lund in Sweden:

Profile 63G
Subject: Feedback
Data bases: ERIC, ISI
Logic: A & B or C or D

Term Group	Search Term	Term Type
A	FEEDBACK	WORD
B	IMMEDIATE	"
B	DELAY	"
B	PARTIAL	"
B	TEACHING MACHINES	"
B	ACHIEVEMENT	"
B	PERFORMANCE	"
B	PROGRAMMED	"
B	PROGRAMED	"
C	KNOWLEDGE OF RESULTS	•
D	FEEDBACK	KEYWORD

It should be noted that spelling variants have to be written as single words, Cf. PROGRAMMED and PROGRAMED. The output from four searches on the ERIC tapes has been evaluated by the user, if we use his requests for photocopies as a feedback to the service. The first run resulted in 247 references, 105 of which were picked up only by the keyword FEEDBACK, 97 only by the free text words, and 45 by both methods. Of more interest is, perhaps, that the user ordered 7 of the documents retrieved by natural language only, 1 retrieved by the keyword, and 8 which could have come out by either method.

The results of the four runs were as follows:

	Documents retrieved		Documents ordered	
Retrieval method	No.	Per cent	No.	Per cent
Keyword only	153	42%	10	24%
Words only	127	35%	15	37%
Keywords/Words	83	23%	16	39%
	363	100%	41	100%

It is a duplication of effort to use thesaurus terms which already exist in titles, a phenomenon that we encounter here. The keyword FEEDBACK occurs also as title word in 23 per cent of the titles, which shows the futility of repeating indexing terms which are identical with title words. It can be noted that one of the references ordered from the set retrieved by the keyword only could have been retrieved by the free text search if a lefthand truncation had been used before the word FEEDBACK. The title word was POSTFEEDBACK.

In this case over 11 per cent of the received references were of such interest that the user ordered photocopies. On the average of all data bases the requests for photocopies as result of the SDI service are lower, around 7 per cent.

Example 2

In present-day society, with a trend to continuing education, a reappraisal of training methods and teaching procedures is important. This raises the question of measurement techniques. The following profile has been constructed to meet the need of a researcher at the training centre at Uppsala University:

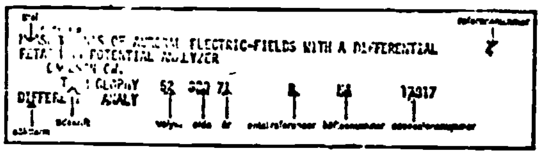
Profile 64G
Subject: Measurement techniques
Data bases: ERIC, ISI
Logic: A & (B or C or D) or D & E or G & (F or H) or I

Term Group	Search Term	Term Type
A	INSTRUCT/	WORD
A	TEACH/	"
B	PROGRAMMED/	"
B	PROGRAMED/	"
B	AID/	"
B	MEDIA/	"
B	INSTRUMENT/	"
B	EQUIPMENT/	"
C	ELECTRONIC/	"
D	TECHNOLOGY/	"
E	EDUCATION/	"
F	PUPIL/	"
F	CLASSROOM/	"
G	OBSERVATION/	"
H	TECHNIQUE	"
I	PROGRAMMED INSTRUCTION/	KEYWORD
I	PROGRAMED MATERIALS/	"
I	DIAGNOSTIC TESTS/	"
I	ELECTRONICS/	"
I	EVALUATION TECHNIQUES/	"
I	CLASSROOM OBSERVATION TECHNIQUES	"
I	MEASUREMENT TECHNIQUES/	"
I	MEASUREMENT INSTRUMENTS/	"

The slashes stand for truncation so any flexion form after the slash will be accepted in the free text field, and any word after the slash in the keyword field if more words are used to define the concept the keywords stand for. This profile lists 23 terms, 15 of which are free text words and 8 keywords. This is one of the profiles in the retrospective search which in total covered 59,779 references. It resulted in 558 references of which the user selected 55 by requesting photocopies. Of these 26 were picked up by keywords alone, 19 by free text, and 10 by both methods. Most efficient were the following search terms:

Selektiv Delgivning av Info.mation
Kungl Tekniska Högskolans Bibliotek
Dokumentationstjänst

DECS/DOC (72) 15



sökprofil nr
70E1

LITTERATURLISTA

Formulär 3

datum
10/03/72

beställare

kontaktperson

företag/institution

adress

postadress

UNDERSTANDING THE LAW: A GUIDE FOR TEACHING THE MENTALLY RETARDED.

BR-6-2883
AUG 69
VIKT=200,00 * AUDIOVISUAL* MENTALLY HANDICAPPED* RETARD* ED044835 1

USE OF THE PEABODY PICTURE VOCABULARY TEST WITH THE EDUCATIONALLY HANDICAPPED
FITZGERALD, BERNARD J. AND OTHER
JOURNAL OF SCHOOL PSYCHOLOGY; 8; 4; 296-299
W '70
VIKT=100,00 * PICTURE* HANDICAP* EJ032660 2

A PREVOCATIONAL AND SOCIAL ADJUSTMENT PROGRAM FOR EDUCABLE RETARDED ADOLESCENTS: A PILOT PROJECT. MILWAUKEE MEDIA FOR REHABILITATION RESEARCH REPORTS. NUMBER 10.
BEEDY, VERNON AND OTHER
JAN 71
VIKT=80,00 * MEDIA* RETARDATION* MENTAL RETARDATION* MENT ED046041 3

NARRATIVE EVALUATION REPORT ON THE INSTITUTE FOR IMPLEMENTATION OF MEDIA PROGRAMS IN DISADVANTAGED AREAS.

70
VIKT=24,00 * AUDIOVISUAL* MEDIA* DISADVANTAGED* ED047754 4

SIXTH IASLIC SEMINAR PAPERS. PART I: REFERENCE SERVICE-IN-ACTION. PART II: PROCESSING & SERVICING OF SPECIAL MATERIALS IN LIBRARIES.

70
VIKT=20,00 * AUDIOVISUAL* SPECIAL* ED047750 5

THE ROLE OF MEDIA IN THE EDUCATION OF EMOTIONALLY HANDICAPPED CHILDREN.

70
VIKT=20,00 * MEDIA* HANDICAP* ED046158 6

NATIONAL CENTER ON EDUCATIONAL MEDIA AND MATERIALS FOR THE HANDICAPPED: POLICIES AND PROCEDURES.

AUG 70
VIKT=20,00 * MEDIA* HANDICAP* ED044857 7

Retrieval method	Search term	Documents retrieved	Documents ordered
KEYWORD	EVALUATION TECHNIQUES	247	7
"	MEASUREMENT TECHNIQUES	145	5
"	MEASUREMENT INSTRUMENTS	88	1
"	CLASSROOM OBSERVATION TECHNIQUES	76	14
WORD	TEACH/ (in various combinations)	74	13
"	EDUCATION/ - " -	29	1
"	INSTRUCT/ - " -	23	9
"	TECHNOLOGY/ - " -	21	5
		703	55

We have not tried to eliminate overlapping occurrences of keywords and words here. From the user's point of view his evaluation by documents ordered shows equal preference for the references selected by keywords and by free text.

Example 3

The next example treats the problem of audiovisual aids for the mentally retarded. See Figure 2, Profile 70 E in Chapter 4 above. This profile has special interest since the user has sent back the graded evaluation form for the output of both ERIC and ISI. The evaluation was as follows:

Tape service	Number of runs	Very interesting	Interesting	Irrelevant	Copy order
ERIC	5	17	28	31	16
ISI	12	1	5	25	1
	Totals	18	33	56	17

ERIC is obviously the central data base for queries of this kind. However, the journals covered by ISI cannot be completely neglected. In spite of the high noise level of ISI, some relevant material has come out. "Formulär 3" shows what the output looks like. The user has ordered photocopies of item no. 1 and 4 which he has circled. The noise level for ERIC is over 40 per cent in this special case where ERIC is the central data base.

Example 4

Library services belong primarily to the educational field, so we could assume that ERIC would be an appropriate data base. The query was about automation in libraries by data processing. The profile (Figure 5) which contains 36 free text words, has been run on five data bases. The evaluation of 482 references from the different data bases follows:

Data base	No. runs	Very interesting	Interesting	Irrelevant	Noise Level
ERIC	3	30	36	196	75%
CICP	7	3	11	11	44%
COMPENDEX	5	16	7	7	23%
INSPEC	5	7	13	37	55%
ISI	25	9	26	74	62%
	Total	64	93	325	

Figure 5.

Profile 563
Subject: Automation in libraries by data processing
Data bases: CICP, COMPENDEX, ERIC, INSPEC, ISI
Logic: G & (F & (A or B) or A & B or C:4 or D & E)

Term No.	Term Group	Search terms	Weight	Term Type
05	A	ARCHIVE/	2	WORD
06	A	BOOK/	2	WORD
07	A	DOCUMENT/	2	WORD
08	A	JOURNAL/	2	WORD
09	A	LIBRAR/	2	WORD
10	B	E D P	2	WORD
11	B	EDP	2	WORD
12	B	ADMINISTR/	2	WORD
13	B	AUTOMAT/	2	WORD
14	B	COMPUT/	2	WORD
15	B	CONTROL/	2	WORD
16	B	DEVELOP/	2	WORD
17	B	GOVERNMENT POLICY/	2	WORD
18	B	ORGAN/	2	WORD
19	B	POLICY/	2	WORD
20	B	RESEARCH/	2	WORD
21	B	RETRIEV/	2	WORD
22	B	ROUTIN/	2	WORD
23	B	SEARCH/	2	WORD
24	B	STOR/	2	WORD
25	B	SYSTEM/	2	WORD
26	B	TECHNIQUE/	2	WORD
27	B	TREND/	2	WORD
28	B	/PROCESS/	2	WORD
29	C	LIBRAR/	2	WORD
30	C	UNIVERSIT/	2	WORD
31	D	COURSE/	2	WORD
32	D	CURRICUL/	2	WORD/
33	D	EDUCAT/	2	WORD
34	D	SCHOOL/	2	WORD
35	D	TRAINING/	2	WORD
36	D	UNIVERSIT/	2	WORD
37	E	DOCUMENTATION/	2	WORD
38	E	INFORMATION/	2	WORD
39	E	LIBRAR/	2	WORD
40	F	LEND/	2	WORD
41	F	LOAN/	2	WORD

Even if ERIC covers the major part of the very interesting references, the output from the other data bases cannot be neglected. The noise level is notably high in ERIC.

The delay time for the same reference appearing in the various services has been studied. We know that ISI is much faster than COMPENDEX or INSPEC, and also than ERIC. Thus for 51 identical references for this profile, the ERIC data base was 8 months later as a median value than ISI and 5 months later than INSPEC. However, delay time often has not even the slightest effect on the user. What happends instead is that when he receives an early reference, he judges it as of low interest or irrelevant, while the same reference appearing 3-6 months later is evaluated as interesting, and he orders a copy. In several cases it seems that the continuous SDI service has a sort of subconscious learning effect on the user.

8. METHODS TO ESTABLISH A HELPFUL OUTPUT ORDER

This paper is not intended as a primer on information retrieval for those interested in education, but the reader might already have noticed in Figures 2 and 4, and in the "Formulär 3" that there are indications of a weighting procedure (VIKT - Weight). We should, therefore, like to mention that we are experimenting with various weighting methods in order to establish a helpful arrangement of the output so that references early on the list will have higher probability of interest to the individual user than the later ones. The method shown in Figure 2 is based upon the assumption that the words used in the profile and the words occurring in a reference are related in such a way that the more the words co-occur, the higher the probability that the reference is relevant to the query.

This gives us one way of arranging the output. Thus we note the number of co-occurrences and let the search logic operate arithmetically to arrive at the values upon which we base the order. As can be noted from the profile 70E in Figure 2, the weight, in general, is 2 assigned to all terms. However, the user has regarded some terms of greater importance and assigned the weight 10 to them. The three words which pick up the first reference in the printout "Formulär 3" have all the weight of 10, two of which are in the same term group, thus, 10+10. The logical "and" is translated into multiplication, so the complete expression will be; 10 x 2 (10) = 200, as the weight shows. In this case it seems to have worked to the user's satisfaction, since he has ordered a copy by circling the reference.

Usually, we do not encourage the user to ascribe subjective weights, as we want to find out more about objectively assigned weights. This brings us back to the list of word frequencies dealt with in Chapter 6 above. We might, in particular, order the references on the basis of the frequencies of the words in the data base, which is our next step in preparation. The underlying reasoning is as follows.

When forming the logical expression in a keyword based system arranged as an inverted file, it is common to base the logical expression upon the number of documents indexed by each keyword. This number indicates the frequency with which this keyword has been used for indexing. Thus, on-line searches on a display terminal usually end by forming the logical expression that gives the minimum output. This means that high frequency terms are looked upon as having less value than those with low frequencies.

In a free text search system in the batch processing mode, a search can also be based upon term frequencies using natural language if we build a frequency table from a large sample of references of each data base, say around 30,000 references. The values for rank ordering could then be established as the sum of the values of the co-occurring terms, if these are expressed in 1/n, where n is the frequency of the term given by the frequency table (Tell 6). Such frequency tables are under construction for ERIC and the other data bases.

The weighting procedure is only the first step. We are going to study parsing and computational linguistic methods in order to find out the contribution such methods can offer to output ordering. We hope to arrive at shorter lists by introducing a cut-off when the weights are too low, thus saving computer and user time.

9. PERSONNEL AND TRAINING

Being responsible for exploring the utility of computerized information services to scientific research, higher education and industry, we have felt that one task has been to carry out research and development of the kind which has been disclosed above. The other tasks are production, management, clerical support, and supporting library service. The overall staff picture for running the SDI service is 10 full-time equivalents. The number of subject specialist is 5, clerical equivalents 4, and programmers 1. When the ERIC data base was introduced there was a pressing need for another subject specialist.

An agreement was reached between our library and the National Library for Psychology and Education to make available a subject specialist on a 4/5 full-time basis; and at the same time they pay for the tapes.

This specialist, the co-author Hemborg, has taken over the profile negotiation, coding, and retrieval analysis which had previously been the task of the other co-author Wessgren. In the transitional stage that we are in at present, operating with two systems, ABACUS and VIRA, the profile updating is laborious which has made it difficult, for example, to devote time to the construction of group profiles of interest in the educational area. Both SDI and retrospective searches are tailor-made for the individual and require personal attention of the subject specialist, and become relatively time-consuming, while group profiles are cheaper to update, there being no necessity to adapt to individual requirements.

Also the library back-up service has been put under pressure since the introduction of the ERIC files. Even if requests for copies of the references put out by ERIC RIE are passed on to the National Library for Psychology and Education where the microfiche collection is located, most references to journal articles and technical reports outside the ERIC clearing-house collection are handled by our library from its collections or by inter-library loans. In many cases photocopies are ordered from the National Lending Library at Boston Spa, United Kingdom. This follow-up service is found to be important in order to keep the interest of the users.

The SDI service has undertaken to train two subject specialists from the National Library. The training period has been stretched out over several months. In May 1971 when the ERIC tapes were well run in, a 10-day seminar on modern information retrieval methods in the field of education and instructional technology was also arranged. 24 persons whose activity had close connection with educational research and training participated; they came from Sweden and the other Scandinavian countries.

As a result of this seminar several profiles were received. Recently, a number of profiles (26) have been received from Finland where the University of Jyväskylä has obtained a grant for experimenting with computerized systems like ERIC. In this case the library in Jyväskylä tries to construct and maintain the profiles, and undertakes the individual mailing of the search results they receive from us.

Our experience from trying to market the data bases in other fields to scientists and people in industry has been that the most effective method is one-day seminars where the afternoon session is devoted to group work when every participant under the guidance of one of our staff constructs a profile in his field of interest. We then promise to run it on a trial basis free of charge for a few months. Such a procedure of "taking the service to the user", is, we think, valid also in the educational field in order to attract potential users.

9. CONCLUSIONS

The introduction of the ERIC tape service in our SDI service has been described in this report, where from February 1971 we have built up a user population around 150 profiles, a promising result for a service new to the social science field. However, in order to assure good coverage a great number of the profiles are also run on other services like ISI, COMPENDEX and INSPEC.

The efficiency of using keywords from the ERIC Thesaurus has been discussed. However, in order to utilize the other data bases the profiles have to include natural language words. In the next few months the Bulletin Signalétique will be added to the files; in which case the profiles will then have to be translated into French in order to be searched on title words. In many cases the free text words perform equally well as keywords on the ERIC files. As we use a combined search strategy, we regard indexing with keywords identical to words in the title as futile. It would be more useful if some keywords took the place of broad subject categories.

The user population which now subscribes to the ERIC service might not be representative of the potential user population when the service has been run for some years, and the validity of the examples given might not be high. However, user reactions in the form of requests for photocopies are high enough to be able to infer from the other services we run that here is a good indicator of the user acceptability of ERIC.

The "noise level" - the proportion of irrelevant references - of ERIC seems so far to be higher than we are generally used to for a subject oriented file, but this may be because of the novelty of the service to which we have not yet become accustomed. The users, however, are tolerant and accept it. We believe that more efforts must be made in order to improve the quality of the output by sifting out some irrelevant references. A weighting procedure is only the first step which later on will, for example, involve parsing or computational linguistic methods.

One essential feature required to keep the interest of the user is a fast backup service of documents or photocopies to the output from the SDI service.

In summing up we would like to note:

(1) that the ERIC tapes are a useful addition but often not enough to answer the queries in the educational field, whereas in a multi-data-base environment the user is assured of a good coverage,

(2) the present system is timely, and its economic aspects are not such as to prevent us from doing occasional retrospective searches even on small profile batches,

(3) The quality of the output can with the present line of thinking eventually be brought under control, and

(4) the present backup service with full documentation is necessary to give the user a reasonable degree of convenience.

ACKNOWLEDGMENT

The material for the statistics has been gathered by Mrs. Wessgren and Hemborg, leaving most of the text of the report to the fist author, Tell, who is the principal investigator. Mrs. Wessgren has chosen the examples presented. We are indebted to Dr. Harvey Marron, Department of Health, Education and Welfare, Washing D.C., who has been kind enough to read the manuscript and offer some valuable remarks.

REFERENCES

(1) Tell, B.V., Larsson, R. & Lindh, R., Information retrieval with the ABACUS programme - an experiment in compatibility - IAG Journal, 3(4), 1970, 323-41.

(2) Coward, R.E., Preparation of a range of standards for educational documentation. - EUDISED Technical Studies, Council of Europe, 1971, 105-31.

REFERENCES (continued)

(3) Gluchowicz, Z., Selective Dissemination of Information - a transdisciplinary information retrieval system at the Royal Institute of Technology, Stockholm - IAG Journal, 4(2), 1971, 131-43.

(4) Ze[illegible] M., Exposé VIRA. Paris CNRS-CDSH Equipe Informatique,1972, 22p.

(5) Viet, Jean, Problems in compiling the multilingual EUDISED thesaurus - EUDISED Technical Studies, Council of Europe, 1971, 94-6.

(6) Tell, B.V., Free text retrieval and an ordering for the printout - ENEA Tutorial Seminar on "Indexing vs. Free Text Retrieval" Studsvik 1972. 8 p. (TRITA-LIB-1037).

APPENDIX

FR^QUENCY LIST OF THE TERMS USED IN THE TITLES OF THE ERIC DATABASE

By
Björn V. Tell

The need for reformulation of the query when passing from one discipline to another is important in an information retrieval system. This means that the system must be capable of performing juxtaposition of disciplines and establish association links over the discipline barriers in reaction to the queries. However, account has to be taken of the use of the language (scientific "jargon") in each discipline, if a linkage is to be developed between existing disciplines, so that multi-disciplinary problems can be attacked.

One way to solve this is to develop a communication and translation system between the various scientific disciplines comprising the machine generation of vocabularies, concordances and thesauri, as an interim stage which might eventually result in a common language for all sciences.

As a first step a frequency list of the words used in the ERIC data-base has been constructed. The bibliographic references to 19,917 ERIC reports and 27,573 journal articles have been merged, and in the 47,490 titles we have found 454,466 word occurrencies, of which 19,856 were found to be unique words, of which 551 were numbers. 10,453 of the words occurred with a frequency greater than one. The list is presented alphabetically and by descending order of frequency.

The construction of the list has been done on an IBM 360/75 using the VIRA retrieval program supplemented by a special program for sorting, editing and printout.

Secretariat Note

The word frequency list has been submitted to Mr. Jean Viet, Paris, for his use in compiling the Multilingual EUDISED Thesaurus.

It is interesting to observe that although there are four times as many word occurrences involved in this analysis as there are in the CIJE-only analysis referred to in Chapter 6, the list of "top twenty" significant terms accounting for 10 per cent of all word occurrences is hardly affected. The terms CURRICULUM, EVALUATION and SOCIAL are replaced by PROJECT(S), UNIVERSITY(IES) - both of which are included in the ERIC Thesaurus - and NEW. The frequency of occurrence of EDUCATION(AL) reaches 16 per cent of the document titles, which means that it is used in some 30 per cent of bibliographic references to ERIC reports.