Computer Aids in the Evaluation of Indexing Terminology*†

HELEN F. GINSBERG, RICHARD F. SCHMITZ, WENDELL K. HOLMAN, and MICHAEL D. HALL
American Dental Association, Chicago, Ill.

Received September 11, 1967

Computer programs have been written for the preparation of specialized term listings from the totality of the terms used to index the multidisciplinary current research project abstracts collected by the Dental Research Information Center. The logic of these programs and the use of the listings in the evaluation of the indexing terminology are presented.

During the survey of current dental research activities in the United States, short project descriptions in a variety of disciplines were collected from dental schools, dental research institutions, federal agencies, and other facilities at which dental research is conducted. These brief abstracts, submitted by the scientists responsible for the project, are the major source of information describing the projects available to the staff of the Dental Research Information Center (DRIC).

The project abstracts are indexed in-depth with a view toward two types of output from a single system. The desired outputs are: (1) the retrieval of all pertinent documents, with a minimum of extraneous material, in a search for specific information from the files and (2) the compilation and analysis of work being conducted in either narrow or broad areas of specialization.

Approximately 1200 terms have been utilized to index the major concepts of 3800 projects; an additional 4000 terms identify specific products, specific techniques, and infrequently occurring concepts mentioned in the abstracts. Because of the multidisciplinary character of the projects, it was deemed advisable to elicit the aid of subject specialists who would evaluate the indexing terminology of their special areas of competency. It would be preferable to submit to these subject specialists a list of terms related to their areas of interest, rather than the totality of indexing terms used. A computer program was written, therefore, to extract and list all terms for each abstract which was indexed also under the concepts identifying the specialties. These partial term listings for broad areas of research then could be reviewed by the corresponding subject specialists. Additionally, the listings would be used by the DRIC staff for internal evaluation of the terms.

Figure 1 outlines the interplay between people and the computer in the development of the thesaurus of index terms. After the abstracts have been indexed, the terms

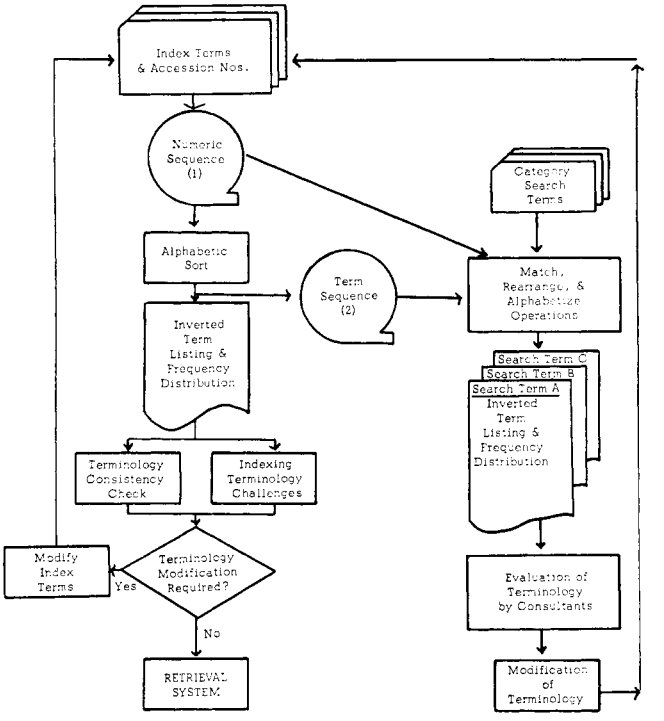

Figure 1. Thesaurus development

and accession numbers of the documents are keypunched, one term per card. The initial computer output is an inverted term listing of all the descriptors, with the document numbers in sequence and a count of the postings under each term. This listing is then used to check the consistency of indexing, both specifically and generically. It also serves as a dual dictionary for manual searches instituted as challenges for the retrievability of the abstracts from the indexed input. The frequency count is an additional tool in the evaluation of the potential utility of the terms for retrieval.

The cycle—card-to-tape, alphabetic sort, printing of the complete inverted term listing, and evaluation of the terminology followed by modification as required—can be

* Presented before the Division of Chemical Literature, 154th Meeting, ACS, Chicago, Ill., September 1967.

† This project was supported in part by USPHS Grant DE 02448 from the National Institute of Dental Research, National Institutes of Health, Bethesda, Md.

repeated until a reasonably stable dictionary has resulted. The next step in the over-all operation is the preparation of inverted term listings of the partial vocabularies specific to the fields of research to be evaluated by the consultants.

Although the DRIC has access to a Honeywell 200 computer on which the magnetic tapes can be read in both directions, a general program was written for the preparation of the specialized term listings which does not require the reverse read feature. Figure 2 is a flow chart of the logic of this program. The input for this second program consists of the numeric and term sequence tapes from which the initial inverted term listing is prepared, and alphabetically sequenced punched cards for the category search terms. There are no programming restrictions on the number of category search terms that can be processed in a single computer run.

In the "select" operation, the category search terms are compared with the alphabetized records on the Term Sequence Tape (Tape 2). The term and the accession numbers of all abstracts posted under that term are written on the Selected Term Tape (Tape 3) which is then sorted numerically by abstract number; the search term remains as part of the record during the sort and on the generated Sorted Selected Terms Tape (Tape 4).

The next sequence of operations obviates the need for the reverse read feature on the tape drives. As Tape 1 (the Numeric Sequence Tape shown in Figure 1) and Tape 4 are read and compared, a count is made of the number of times each matching abstract number is written on Tape 4. If an abstract number occurs only once on Tape 4, the entire record for that abstract (number and all index terms) which appears on Tape 1 is transferred without modification to Tape 5 (Terms with Counts). If an abstract number appears more than once on Tape 4, each corresponding index term from Tape 1 is written on Tape 5 the same number of times that the abstract number occurs on Tape 4, and each record is tagged. In other words, if Abstract 435 had been posted under three of the research specialties for which limited term listings are being prepared:

> Number 435 would appear three times on Tape 4.
>
> The value "3" would be in a counter set up to count the number of times an abstract number occurs on Tape 4.
>
> Each term used to index Abstract 435 (from Tape 1) would be written three times on Tape 5. The three records for each term would contain the identical term, but the identification numbers would now read: 435-01, 435-02, and 435-03.

A numeric sort of Tape 5 reorders the terms extracted from Tape 1 so that each set of index terms for multiple occurring documents is complete within the subset for which additional digits were assigned. In the example previously given, all the 435-01 terms will be written before the 435-02 terms, and so forth. The generated Tape 6 (Sorted Terms with Counts) is now used with Tape 4 for the final matching operation.

The records on both Tape 4 (Sorted Selected Terms) and Tape 6 are in sequence by document number. The next operation appends the category search or selected terms from Tape 4 onto the records for each appropriate index term on a match of abstract number. For example, all the terms identified by number 435-01 will have "Search Term A" affixed to their records; those for number 435-02 will have "Search Term B;" and 435-03, "Search Term C." These expanded records on the Final Terms Tape (Tape 7) are sorted alphabetically, by specific term within major or "select" term category.

Final output from this computer program is a series of coordinated inverted term listings (Figure 3). There will be as many separate listings as the number of category search terms used in the initial "select" operation; each will contain, in addition to the broad term, every term appearing in any abstract in which the broad term has been used. The printout lists also the document numbers in which both terms appear and a count of the frequency with which they both occur in the same abstract.

This program for the preparation of the partial term listings requires a minimum computer configuration of five tape drives and a 16K core, plus the card reader and printer. In addition to the tape read speed and the additional core that might be available, the following factors must be taken into account in estimating the length of time required for a single run for multiple listings:

> The number of category search terms to be processed.
>
> The number of documents in which the search terms appear.
>
> The number of terms associated with each abstract in which the search terms occur.

Obviously, it is difficult to calculate the duration of the computer run in advance.

In Figure 3, a sample page for the selected inverted term listing, the category search term is repeated in the

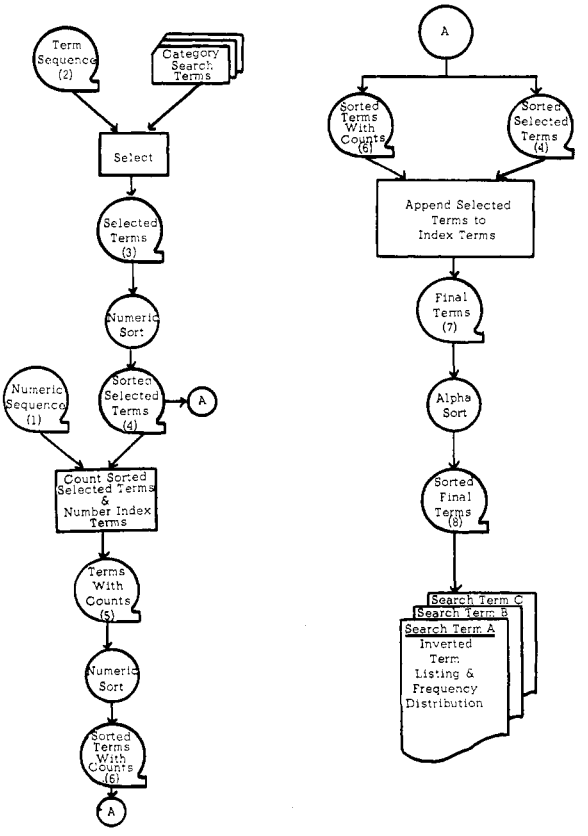

Figure 2. Term search routine

Category Term	Index Term	Document Number								Frequency
MICROBIOLOGY	COMPOSITION	0407	0463	0604	0874	0942	1156	1521		0007
MICROBIOLOGY	COMPUTERS	3330								0001
MICROBIOLOGY	CONCENTRATIONS	1866	2297	2535						0003
MICROBIOLOGY	CONNECTIVE TISSUE	0917	3090	3381						0003
MICROBIOLOGY	CONTAMINATION	0708	1849	2137	2294	2297	2455	2535	2594	
		3327	3341							0010
MICROBIOLOGY	CONTROL	2962								0001
MICROBIOLOGY	CORTICOSTEROIDS	0062	0700							0002
MICROBIOLOGY	CORTISONES	2715								0001
MICROBIOLOGY	CULTURES	0222	0274	0278	0385	0453	0473	0527	0531	
		0641	0825	0832	0847	0866	0877	0878	1120	
		1122	1138	1157	1158	1196	1230	1262	1367	
		1381	1457	1555	1556	1800	1810	1824	1849	
		1865	1866	1877	1888	1910	1968	1977	1999	
		2071	2294	2455	2516	2535	2715	2746	2749	
		2765	2812	2813	2814	2912	2940	2972	3070	
		3090	3097	3098	3132	3224	3339	3381		0063
MICROBIOLOGY	CYTOLOGY	0531	0602	0825	1262	1673	1824	1968	2071	
		2082	2084	2330	2813	3090	3224	3307	3330	0016
MICROBIOLOGY	DEBRIS	1849	2084	2448	3133	3325				0005
MICROBIOLOGY	DECALCIFICATION	0453	0517							0002
MICROBIOLOGY	DECARBOXYLATION	0832								0001
MICROBIOLOGY	DEFICIENCY	1865								0001

Figure 3. Sample page of selected inverted term listing

first column for each specific term. The second column contains the alphabetized index terms from the documents which were posted also under "Microbiology." The last column on the right presents the count of the frequency of occurrence of the two terms. The numbers between the terms and the frequency count are the pertinent abstract numbers. In essence this listing is the result of the coordination of a specific term with the entire vocabulary of the file.

It is evident from the sample page that the specific terms do not have to be related directly to the search term. Examples of these are "computer," "control," and "debris." In the case of the term "cultures" it is conceivable that each time the specific term is used it should be posted also under "Microbiology." A comparison of the frequency of occurrence in the specialized listing with the frequency count in the complete inverted term listing indicates that for the concepts included in our files this higher posting cannot be made automatically. There are almost twice as many postings under "cultures" in the complete listing as appear in the specialized one. As a result of this type of comparison of the postings under the uncoordinated terms in the complete dictionary against the number of postings in the specialized coordinated term listings, scope notes are generated for use in indexing. These scope notes serve also as an aid in processing a request for information from the files.

These specialized inverted term listings and copies of the scope notes will be sent to the subject specialist consultants. Upon completion of the review and modification of the terminology as required by the specialists' suggestions, a final corrected Term Sequence Tape (Tape 2) will be prepared. This tape will be used to produce a deck of punched cards which will contain the abstract number, term, and a code to be assigned to each concept. During the assignment of the codes, the same number will be used to identify terms which are to be considered synonyms in the final dictionary. The punched cards will be sorted numerically by code number and sent to the Jonker Service Bureau for the preparation of Trematrex cards. Our final search medium for the present will be the Trematrex optical coincidence system. When the volume of information in the system and the number of searches processed exceed the capacity of our present operation, we will explore the feasibility of converting the processing system to a completely computerized one.