Development and Implementation of a Thesaurus for the Visual Sciences*

MARY M. EICHHORN and ROBERT D. REINECKE
Vision Information Center, Francis A. Countway Library of Medicine and Howe Laboratory of Ophthalmology, Harvard Medical School, Boston, Mass.

Received December 30, 1968

A thesaurus has been developed to serve as the integrating unit for the computerized information storage and retrieval system of the Vision Information Center. The center maintains records of information in the visual sciences which are available to the user in the forms of computer-assisted instruction, literature retrieval, and patient records. The numerical coding system used in the thesaurus permits seven levels of specificity; this specificity is required for depth of indexing, as well as to limit the retrieval to those bibliographic citations which are relevant to a highly specific search request. The flexible design of the thesaurus facilitates frequent revision and addition of new terminology.

The Vision Information Center (VIC), a project combining the efforts and facilities of the Francis A. Countway Library of Medicine and the Howe Laboratory of Ophthalmology, is one of the network of specialized centers sponsored by the National Institute of Neurological Diseases and Stroke. The four centers are conceived of as an information network for neurological and sensory diseases. Each center strives to facilitate the dissemination of biomedical information pertaining to its subject area. The Vision Information Center has developed the capability to organize and maintain records of new knowledge relevant to ophthalmology and the visual sciences. This knowledge will ultimately be available to the user in three forms: computer-assisted instruction (CAI), which will provide information on a specific topic at an elementary level; retrieval of bibliographic citations, which will give the user access to the published literature in his specific area of interest; and specific patient data, which will contain diagnostic and therapeutic information on specific case reports. The computer-assisted instruction and literature retrieval are already available at a computer terminal; however, it will be several years before the patient data are incorporated into the system. The over-all structure of VIC is shown in Figure 1. This paper is primarily a report on the development and implementation of the thesaurus, which is the keystone to the entire operation.

The development of a thesaurus containing the highly specific terminology used by ophthalmologists and scientists working in the diverse areas relating to vision was one of the first tasks undertaken by VIC. The thesaurus integrates the various aspects of the VIC data base into a unified system. The user at a computer terminal, with the aid of a thesaurus, can request information in the form of instruction, bibliography, case histories, or any combination of these appropriate to his needs. In addition to its function as a link between the various modes of the data base, the thesaurus also introduces the user to the system. Perusal of the thesaurus printout is, in effect, browsing through the VIC library. The thesaurus has been generated largely by the input of indexing terms, and is used by the indexers to assign descriptors from the thesaurus to documents which are incorporated into the VIC system. When a necessary term is not available

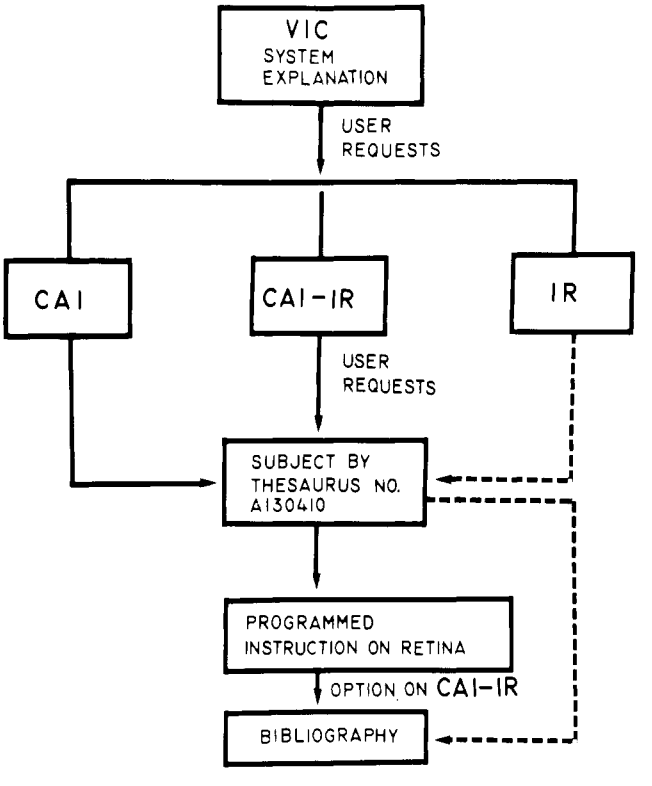

Figure 1. Design of the VIC computer-user interaction; following the system explanation, the user has the option of selecting computer-assisted instruction (CAI), literature retrieval (IR), or a combination of both (CAI-IR)

* Presented before the Division of Chemical Literature, 155th Meeting, ACS, San Francisco, Calif., April 1968.

A.	Anatomical Terms
B.	Organisms
C.	Diseases
D.	Chemicals and Drugs
E.	Analytical, Diagnostic, and Therapeutic Technics
F.	Psychiatry and Psychology
G.	Biological Sciences
H.	Physical Sciences
I.	Social Sciences
J.	Technology, Commerce and Industry
L.	Communication Media
M.	Named Groups of Persons
N.	Health Care

Figure 2. Outline of the VIC thesaurus

in the thesaurus, the term is used as an index term and added to the thesaurus as a new term.

The literature retrieval system of VIC was designed to be compatible with the MEDLARS system developed by the National Library of Medicine. Therefore, the overall structure and divisions of subject categories of the VIC thesaurus are almost identical to those of the Medical Subject Headings of the National Library of Medicine (MESH).[1] The outline of the VIC thesaurus structure is shown in Figure 2. A description of the general organization and goals of VIC has been presented.[2]

THESAURUS STRUCTURE AND DEVELOPMENT

Computer programs for the implementation of the VIC thesaurus, as well as for other programs of VIC, were written for the IBM 360/50 computer by members of the Harvard University Computing Center under the direction of A. Mildred Koss and Theresa C. Lee. Details of the programming work will be published elsewhere.

To facilitate machine searching of the data base, it was important to have a numerical, hierarchical coding system for the thesaurus terminology. An example of this type of thesaurus in the biomedical sciences is that in use at the Lovelace Foundation.[3] Since neither the Lovelace thesaurus nor the alphanumeric codes employed by MEDLARS satisfied the VIC requirements, a new hierarchical coding system which allows seven levels of specificity has been developed. The first, or broadest, level is designated by a letter; for example, the letter D indicates Chemicals and Drugs. The remaining six levels are specified by a numerical classification consisting of six pairs of two digits. Thus, each of these six categorical levels has a capacity of up to 99 terms. Where both digits of a pair are zeroes, the indicated level of specificity has not been designated by that term. An example of the hierarchical structure and coding system is shown in Figure 3.

To achieve the depth of indexing and degree of specificity required by the specialists who are the primary users of VIC, it was necessary that the VIC thesaurus include many more detailed or specific descriptors than are found in MEDLARS. This was especially true in the areas of the anatomy, physiology, and diseases of the eye, and ophthalmic surgery. Therefore, the VIC thesaurus has deviated from MEDLARS in setting up separate categories for these subjects, which MEDLARS includes under more general groupings. For example, in the VIC thesaurus, the category of A13, which in MEDLARS is used for terminology of animals, is designated for anatomy of the eye, and C13 for ophthalmology, or diseases of the eye. The category for Diseases of Nutrition and Metabolism, which in MEDLARS is C13, was transferred to C17 in the VIC system. The "Coding System for Disorders of the Eye,"[4] developed for the International Council of Ophthalmology, has been adapted as the basis of the C13 category. According to this coding system, the first pair of digits following the C13 indicator designates the anatomical site, the second pair of digits indicates pathology, and the third pair of digits indicates etiology. When the latter pair of digits are both zeroes, an unknown or unspecified etiology is signified. When it has been necessary to further subdivide a disease category listed in the coding system, the zeroes to the right of the etiology code have been replaced by appropriate numbers. A section of the eye disease category is illustrated in Figure 4.

The availability of seven levels of specificity has allowed VIC detailed classification in categories other than ophthalmology which are of importance in handling the

```
D23 10 CC 00 00 00    ENZYMES
D23 10 C1 00 00 00        OXIDOREDUCTASES
D23 10 C1 00 00 00        (DEHYDROGENASES)
D23 10 C1 01 00 00            ACTING ON CH-OH GROUPS
D23 10 C1 01 01 00                WITH NAD OR NADP AS ACCEPTOR
D23 10 C1 01 01 22                    UDP GLUCOSE NAD OXIDOREDUCTASE
D23 10 C1 01 01 22                    (URIDINE DIPHOSPHOGLUCOSE DEHYDROGENASE)
D23 10 C1 01 01 27                    L-LACTATE NAD OXIDOREDUCTASE
D23 10 C1 01 01 27                    (LACTATE DEHYDROGENASE)
D23 10 C1 01 01 27                    (LACTIC ACID DEHYDROGENASE)
D23 10 C1 01 01 37                    L-MALATE NAD OXIDOREDUCTASE
D23 10 C1 01 01 37                    (MALATE DEHYDROGENASE)
D23 10 C1 01 01 49                    GLUCOSE 6 PHOSPHATE DEHYDROGENASE
D23 10 C1 01 01 71                    SORBITOL NAD OXIDOREDUCTASE
D23 10 C1 01 01 71                    (SORBITOL DEHYDROGENASE)
D23 10 C1 01 01 71                    (POLYOL DEHYDROGENASE)
```

Figure 3. The hierarchical structure of the VIC thesaurus; this section illustrates the coding system as applied to enzymes

```
C13 75 CC 00 00 00    CORNEAL DISEASES
C13 75 1C 00 00 00        KERATITIS
C13 75 1C 00 00 00        (CORNEAL INFLAMMATIONS)
C13 75 1C 15 00 00            VACCINIAL KERATITIS
C13 75 11 00 00 00        KERATITIS PUNCTATE SUPERFICIAL
C13 75 11 00 00 00        (PUNCTATE KERATITIS)
C13 75 11 00 00 00        (KERATITIS PUNCTATA)
C13 75 11 00 01 00            EPITHELIAL KERATITIS
C13 75 14 13 00 00        HERPES SIMPLEX KERATITIS
C13 75 14 13 00 00        (CORNEAL DENDRITIC ULCER)
C13 75 14 13 00 00        (DENDRITIC KERATITIS)
C13 75 14 13 01 00            PRIMARY OCULAR HERPES SIMPLEX
C13 75 14 13 02 00            CHRONIC OCULAR HERPES SIMPLEX INFECTION
C13 75 15 13 00 00        DISCIFORM KERATITIS, HERPES SIMPLEX
C13 75 15 13 00 00        (KERATITIS DISCIFORMIS)
C13 75 16 00 00 00        INTERSTITIAL KERATITIS
C13 75 16 23 00 00            KERATITIS INTERSTITIAL, SYPHILITIC
C13 75 18 00 01 00            NECROGRANULOMATOUS KERATITIS
C13 75 19 13 00 00        METAHERPETIC KERATITIS
C13 75 19 13 00 00        (KERATITIS METAHERPETICA)
C13 75 19 13 01 00            CORNEAL METAHERPETIC ULCER
```

Figure 4. A section of the eye disease category from the VIC thesaurus

vision literature. These include the B03 category, where "Bergey's Manual"[5] has been followed for the classification of bacteria. At present, all drugs are coded under the D04 category; the drug classification is a physiological one and is essentially based on the outline given by the chapter headings in "The Pharmacological Basis of Therapeutics" by Goodman and Gilman.[6] It is anticipated that the drug index currently in preparation by the joint effort of the American Chemical Society and the National Library of Medicine will be a valuable aid for the development and maintainance of the classification of drugs in the VIC thesaurus. Proteins are classified under D23 in the VIC thesaurus; the subgroup D23 10 is reserved for enzymes. The nomenclature and code assignations for enzymes of the International Union of Biochemistry[7] have been followed and incorporated into the VIC system. The four pairs of digits following the D23 10 indicator consist of the enzyme code numbers, as recommended by the International Union of Biochemistry.

Synonymous terms are included in the thesaurus where several expressions for a concept are in common current usage. The present program allows for the inclusion of up to five synonyms. A synonym is so indicated by enclosure in parentheses, and is identified with its preferred term by the code number. Whenever a term in the VIC thesaurus occurs in MEDLARS, the MEDLARS term is used as the preferred term. If the VIC term does not exist in the MEDLARS system, the term preferred in the VIC thesaurus is selected in accordance with the terminology preferred by the scientists who are specialists in each particular field.

The VIC thesaurus has been designed to allow frequent updating by the addition of new terminology, and revision when necessary. During the first few months of operation, the thesaurus was updated monthly. At present, updated editions are printed quarterly by computer. This updating and revision can be accomplished without changing the basic structure of the thesaurus. This flexibility is an essential feature of the Vision Information Center, which is seen as an evolving system designed to serve users whose scientific interests and levels of specialization cover the wide range of biological and medical sciences relevant to vision. To encourage uniformity rather than diversity in the efforts of the various groups working with the scientific literature, nomenclature and coding systems already in existence have been incorporated into the VIC thesaurus where it was possible and appropriate to do so.

DATA INPUT

Except for broad terms used as category headings—i.e., eye diseases, bacteria, drugs, etc.—the terminology in the thesaurus was derived almost exclusively from the indexing terms assigned to the journal articles which constitute the VIC bibliographic data base. Each indexing term is assigned a code number, which indicates the location of the term in the hierarchical structure. The code number and term are keypunched in their specific fields on a single card. The cards are then entered into the computer storage file. The entry program prevents the entry of more than one preferred term under any given code number. Likewise, the program provides for appropriate bibliographic file maintainance when a term is deleted from the thesaurus or inserted under a new code number.

After 18 months of operation, the VIC thesaurus contains approximately 5000 terms. Currently, 90 to 95% of the indexing terms assigned to articles by the VIC indexing staff are found in the thesaurus. Within the next year this figure should approach 99%, as new terms are added to the thesaurus at a gradually decreasing rate. The current rate of indexing input is approximately 3000 articles per year.

Once the indexing of a journal article is completed, the code number of each indexing term and the VIC acquisition number for the article are punched in specific fields on a card. The acquisition number serves as the link in the computer file between the terms and the bibliographic data which are punched in fields specified by the VIC format on separate cards, which are also punched with the acquisition number.

COMPUTER OUTPUT

Programs have been written to generate computer printouts of the thesaurus in two forms, the hierarchical structure shown in Figure 3, and the permuted alphabetic sort illustrated in Figure 5. In the latter printout, every thesaurus term appears in alphabetic order under each significant word in the term. The permuted alphabetic sort is most useful for the indexer or searcher who must look up in the thesaurus the term he wishes to use as an index or search term. Once the proper term is found in the alphabetic sort, reference to the hierarchical listing permits the user to see the term he has selected in context with its related broader and narrower terms, and to determine the appropriate level of specificity for his indexing or searching purposes.

An additional program provides a printout of an index to the data base. The index is a listing, in hierarchical order, of the code numbers of all thesaurus terms. Beside each term's code number are printed the acquisition numbers of all articles indexed under that term. Reference to this index, therefore, gives the searcher information about the number of citations which may be retrieved under any given term.

SEARCHING THE VIC DATA BASE

The search capability of the VIC literature retrieval system was designed to allow the user to have direct access to the computer files. The inexperienced user working for the first time at the computer console can request instructions from the computer both on the use of the console and on the operation of the VIC system, including the information necessary to use the thesaurus for conducting a search. The experienced user has the option of bypassing the preliminary instructions.

The hierarchical coding system permits facile machine search capability; in a simple operation, one goes from a broader term to a narrower, more specific term, a subcategory of the broader term, or in the opposite direction from the specific term to a more general one. For example, if a search request produced too few citations indexed under the specific term "retinal capillaries," one would reformulate the search using the next broader term, "retinal blood vessels." Use of the appropriate thesaurus

```
                (LACTIC ACID DEHYDROGENASE)                                D23 10 01 01 01 27
                    (MALATE DEHYDROGENASE)                                 D23 10 01 01 01 37
                    (POLYOL DEHYDROGENASE)                                 D23 10 01 01 01 71
                  (SORBITOL DEHYDROGENASE)                                 D23 10 01 01 01 71
                  SUCCINATE DEHYDROGENASE                                  D23 10 01 03 99 01
(URIDINE DIPHOSPHOGLUCOSE DEHYDROGENASE)                                   D23 10 01 01 01 22
                           (DEHYDROGENASES)                                D23 10 01 00 00 00
                           (DEHYRATASES)                                   D23 10 04 02 01 00
                STRABISMUS DELAY IN TREATMENT                              E02 01 06 01 01 00
                           DELAYED HYPERSENSITIVITY                        C14 06 03 01 00 00
                           DELAYED OFF RESPONSES                           G01 08 09 12 01 00
                           DELLEN                                          C13 75 26 00 04 00
                           DELTA CRYSTALLIN                                D23 03 01 01 04 00
                           DEMECARIUM BROMIDE                              D04 01 02 01 05 03
                           DEMENTIA                                        C10 02 12 00 00 00
                          (DEMEROL)                                        D04 02 02 02 09 00
                          (DENDRITIC KERATITIS)                            C13 75 14 13 00 00
                   (CORNEAL DENDRITIC ULCER)                               C13 75 14 13 00 00
                           DENERVATION                                     E07 13 12 00 00 00
                SYMPATHETIC DENERVATION, CHRONIC                           C10 01 02 00 00 00
                       FILM DENSITOMETRY                                   E05 06 08 00 00 00
                           DENSITY GRADIENT ULTRACENTRIFUGATION            E05 02 21 01 01 00
                          (DENTAL ANOMALIES)                               C04 05 26 01 00 00
                           DENTAL CARIES                                   C04 05 06 02 00 00
                           DENTAL MESODERMAL DYSPLASIA                     C04 05 26 01 03 00
                          (DEOXYRIBONUCLEIC ACID)                          D24 02 02 00 00 00
                           DEPOLARIZING NEUROMUSCULAR BLOCKING AGENTS      D04 01 02 04 02 00
                 CYSTINE DEPOSIT                                           C16 03 04 01 08 01
```

Figure 5. A section of the permuted alphabetic printout of the VIC thesaurus

code numbers rather than words at both the indexing and search formulation steps assures consistency of terminology from the input to the output processes.

When a single term does not suffice for search purposes, as many terms as are indicated by the user's query can be input into the computer. The terms are combined by the computer's search strategy in such a manner that only articles which have been indexed under all the given terms are retrieved. For example, a reader who wants references on the metabolism of glutathione in the crystalline lens would input the thesaurus code numbers for the three terms, metabolism, glutathione, and crystalline lens. This mechanism serves to avoid the retrieval of many irrelevant citations. An example of the conversational interaction between the computer and the searcher is shown in Figure 6.

```
* Do you wish to make another search?  (yes or no)

  yes
* Input thesaurus number, please.

  a130405   (crystalline lens)
* There are    43 references filed under that number, including possible duplications.

  Do you wish to continue with the number or change it to another with a different
  set of references?  (Type "con" or "chg".)

  con
* Type "list" for references, or input additional number to make search more specific.

  d220301   (glutathione)
* There are     2 references filed under that number, including possible duplications.

  Do you wish to continue with this number or change it to another with a different
  set of references?  (Type "con" or "chg".)

  con
* The combination of these numbers yields     1  articles, which eliminated duplications.

  Type "list" for references, or input additional number to make search more specific.

  list

*000842 WOOD D.C
 RESPONSE OF RABBITS TO CORTICOSTEROIDS. II. INFLUENCE
 OF TOPICAL THERAPY ON LENS, AQUEOUS HUMOR, SERUM
 AMER J OPHTHAL 63;849,1967

*Do you wish to make another search?  (yes or no)

  no
*Please type sign off.

  sign off
*LINE IS SIGNED OFF
```

Figure 6. Search query of the VIC IR data base

The computer's directions (or questions) are preceded by an asterisk. The user's replies follow. In response to the request for a bibliography on glutathione in the crystalline lens, the computer lists the one citation entered into the data base under both terms, "glutathione" and "crystalline lens."

Computer-assisted instruction courses on basic ophthalmology and glaucoma have been programmed and are available at the computer terminal. At present, the student must enter the instructional mode at the beginning of the course he has chosen and proceed sequentially through it. However, thesaurus code numbers have been assigned to frame sequences of the programmed instruction which deal with specific subjects. When necessary programs are implemented, it will be possible for the student to request instruction on a specific subject of interest, for example, on the technique of ophthalmoscopy. The frames of the basic ophthalmology course which explain ophthalmoscopy will then be presented. In this manner, a user can readily receive both instruction and a bibliography on the subject he has selected from the thesaurus.

Additional search capabilities are being planned which will be incorporated into the VIC system. These include search by author, language, and date of publication as well as by journal title. The latter capability provides the potential of generating the annual index for any given journal. It is anticipated that use of the present VIC data base by members of the scientific community will provide feedback and critical suggestions which will guide the Vision Information Center in its future development plans.

LITERATURE CITED

(1) U. S. Department of Health, Education, and Welfare, Public Health Service, National Library of Medicine, "Medical Subject Headings," Vol. 8, U.S. Government Printing Office, Washington, D. C., 1967.
(2) Reinecke, Robert D., "Vision Information Center—Direct User Access via Computer-Assisted Instruction," *Proc. Am. Soc. Inform. Sci.*, Vol. 5, "Information Transfer," Greenwood Publishing Corp., New York, 1968.
(3) Roth, Emanuel M., Charles W. Sargent, and Marjorie M. Benson, "Aerospace and Environmental Medicine Information System Thesaurus," Department of Aerospace Medicine and Bioastronautics, Lovelace Foundation for Medical Education and Research, Albuquerque, N. M., 1966.
(4) Schappert-Kimmijser, J., A. Colenbrander, and S. Franken, "Coding System for Disorders of the Eye," S. Karger AG, Basel, Switzerland, 1968.
(5) Breed, Robert S., E. G. D. Murray, and Nathan R. Smith, Eds., "Bergey's Manual of Determinative Bacteriology," 7th ed., Williams and Wilkens Co., Baltimore, Md., 1957.
(6) Goodman, Louis S., and Alfred Gilman, "The Pharmacological Basis of Therapeutics," 3rd ed., MacMillan, New York, 1965.
(7) Florkin, Marcel, and Elmer H. Stotz, Eds., "Comprehensive Biochemistry," Vol. 13, 2nd ed., Elsevier, New York, 1965.

Supported by DHEW Contract PH 43-66-911 from the National Institute of Neurological Diseases and Stroke.

Compatibility in Chemical Information Systems

DAVID P. JACOBUS
Walter Reed Army Institute of Research, Washington, D.C. 20012

KENNETH H. ZABRISKIE
BIOSIS, Philadelphia, Pa.

MAXWELL GORDON*
Smith Kline and French Laboratories, Philadelphia, Pa. 19101

Received January 31, 1969

As increasing amounts of data are processed for computer storage and retrieval, it becomes important to amortize the cost of this operation more rapidly, by a wider exchange of data. Obviously, exchange of machine-processed data introduces problems of compatibility and convertibility. It is urged that planning of information operations take these problems into account at an early stage, since seemingly minor differences between data systems can have disastrous effects on data exchange efforts. The compatibility and convertibility of chemical data, as special cases of the foregoing, have their own special problems. A survey of these problems is presented, together with some suggested solutions to compatibility and convertibility in this area.

The problem of compatibility among information systems becomes increasingly important as systems become more and more mechanized. It is our feeling that, inasmuch as the cost of input of scientific information is great, maximum use of this information will be possible only if it can be exchanged among a large community of users to reduce the unit cost and optimize the use of our intellectual resources. It has been suggested that the cost of information processing and transmission will one day exceed the cost of input. However, compatibility considerations have relevance to information processing and transmission, as well as to input costs.

For the past several years, the Committee on Chemical Information of the National Academy of Sciences/

* To whom questions should be directed.