# A Mechanized Information and Document Retrieval System*

HAROLD W. BATCHELOR and CLIFFORD J. MALONEY**
U. S. Army Biological Laboratories, Fort Ketrick, Frederick, Maryland

Received November 7, 1963

This mechanized information and document retrieval system has proved useful for retrieving information from the several hundred documents that have been indexed in depth thus far. The primary objective was to design a system that would be adequate to retrieve highly specific topics from many scientific disciplines that are involved in the Biological Laboratories research and development program. Other objectives were to keep the system as simple as possible, to minimize ambiguities in the glossary of terms, and to have the system operable by a Solid-State 90 electronic computer and its ancillary equipment that are available at this installation.

No preconceived notions dictated the selection of a particular retrieval system. Neither was there any intention to restrict the system to a conventional design. Rather, the specific intention was to select the best features of existing systems, to avoid their undesirable features, and to devise any innovations that might be needed. The result is a nonconventional system that we have designated "*Co*mposite *Me*chanized *In*formation and *Do*cument *R*etrieval *S*ystem" (COMEINDORS). Interest shown by others in the system as a result of several preliminary published notes[1-7] and the presentation of the paper at the March, 1962 Washington meeting of the American Chemical Society appear to warrant publication of further details.

The chief components of COMEINDORS are:

a. A "pure decimal" notation, familiar throughout the world, that is adapted to both machine and manual operation.

b. An ordered classification of subject headings of terms and their corresponding numerical codes or schedules.

c. An alphabetized glossary or thesaurus of subject headings or terms that constitute definitions by context by use of multiple-word entries that are ordered in a manner comparable to their counterparts in the classification.

d. A series of lattice-type classification forms that simplify and expedite the development and control of the ordered classification of terms.

e. Printed forms that assist in the indexing of documents, reduce clerical assistance, and minimize errors in the assignment of codes.

f. A flow sheet of operations that delineates the types and number of personnel needed and their specific duties.

Major features of COMEINDORS include:

a. The issue of the system's 250-page manual as a loose-leaf document that provides means for almost indefinite internal "growth" and revision of the ordered classification and of the alphabetized glossary without disturbing the system as a whole.

b. The avoidance of essentially all scope notes that are a necessary component of some systems.

c. The avoidance of "concealed classifications" inherent in some nonclassified systems.

A discussion of some of the individual components of the system follows.

## THE NOTATION

The Dewey and Universal Decimal systems owe much of their world-wide acceptance to the employment of a nearly "pure decimal" notation. A system intended for Department of Defense use, with its global responsibilities involving allies speaking the widest variety of languages written in nearly every conceivable script, can therefore do no better than to emulate the example of the Dewey and Universal Decimal systems. Use of other symbols to shorten code length is no advantage to the majority of computers, which in general operate most efficiently on digital codes—or even binary codes. Moreover, decimal codes involve little penalty in the case of fixed-word-length computers.

## THE CLASSIFICATION

In one or another manner, the classification resembles the categories of Roget, the ASTIA thesaurus, the codes of Dewey, the arrays and chains of Ranganathan,[8] and the faceted classification of Vickery.[9] Dewey's "zero" is extended to serve as a code marker for categories that are, or may be, divided into more specific terms. Only a few examples extracted from various sections of the classification will be discussed. Because the original plans for COMEINDORS included its possible expansion to all fields of knowledge, the whole of knowledge was first divided into an array of eight coordinate classes that resemble Roget's categories. This first division will be mentioned again under the heading "Lattice Forms." A second example of division into arrays is provided by the eight main classes of the 700 section, as shown in Table I. The long codes for these classes are 710, 720, 730, etc. The long codes that correspond to the classes in an array are therefore designated by increasing the value of the integer in the same digit space.

* This paper was presented under the title "A Classified Deep Index" on the program of the Division of Chemical Literature, 141st National Meeting of the American Chemical Society, Washington, D. C., March 24, 1962.

** Now at the Division of Biologics Standards, National Institutes of Health.

Table I
Extract from Two-Digit Schedules

| | |
|---|---|
| 700 | Professions |
| 710 | Religion |
| 720 | Law |
| 730 | Business |
| 740 | Agriculture |
| 750 | Medicine |
| 760 | Government |
| 770 | Military Science |
| 780 | Engineering |

Table III
An Example of an Array of Terms

| | | |
|---|---|---|
| 7742.6212.58 | Aerosols | 539-01* |
| 7742.6212.581 | Theory | 539-50 |
| 7742.6212.582 | Properties | 540-01 |
| 7742.6212.583 | Behavior | 560-01 |
| 7742.6212.584 | Dispersion | 561-01 |
| 7742.6212.585 | Evaluation and Sampling | 583-01 |
| 7742.6212.586 | Evaluation Systems | 592-01 |
| 7742.6212.587 | Evaluation Results | 602-01 |
| 7742.6212.588 | Counter Measures | 619-01 |

A second, inherent feature of the structured classification is the division of any category in an array into a hierarchical chain, as shown in Table II. The long codes that correspond to successive links in such a chain are formed by inserting the proper integer after the long code of the higher category. In effect, a term or category in an array may also be a link in a chain. For example, the term "Aerosols" that is shown in Table II as generic to only a single term "Properties," in the full classification is in fact divided into an array of eight categories shown in Table III, the term "Properties" being the second term in this array.

Such a combination of arrays and chains in the classification provide "definition by context," an expression suggested by Paul Klingbiel. The association of terms, one with another in this way, imparts such precise meanings to the terms that, for the most part, scope notes are unnecessary. This recalls the rather trite saying, "A picture is worth a thousand words." Thus, reading up the classification in Table II, this chain of terms indicates by context that the term "Aerosols" is limited specifically to BW aerosols, which in turn are part of the larger field of Military Science. Similarly, reading down the classification, this particular chain terminates with an array of terms that are measures of particle size, MMD, NMD, and VMD, (Mass Median Diameter, Number Median Diameter, and Volume Median Diameter, respectively). These examples are but short extracts from the Classification Schedules. As would be expected, the category "BW Aerosols" comprises prominent portions of the COMEINDORS classification and glossary. Other categories of aerosols may appear in other sections of the classification, such as medical or therapeutic aerosols, industrial smokes and smog, agricultural dusts and sprays.

Table II
Extracts from Military Science Schedule

| Long code | | Short code |
|---|---|---|
| 770 | ... Military Science | 096-01 |
| 774 | .... Weapons Systems | 133-50 |
| 7742 | ..... CBR Weapons | 138-01 |
| 7742. 6. | .... Technical Project Programs | 242-01 |
| 7742. 62. | .... In-Force Tasks | 262-01 |
| 7742. 621. | .... Field Agencies | 282-01 |
| 7742. 6212. | .... BW | 492-01 |
| 7742. 6212. 5. | ... Agents | 524-45 |
| 7742. 6212. 58. | ... Aerosols | 539-01 |
| 7742. 6212. 582. | ... Properties | 540-01 |
| 7742. 6212. 5821. | ... Physical Properties | 540-13 |
| 7742. 6212. 5821. 1. | .. Mensural | 540-25 |
| 7742. 6212. 5821. 11. | .. Particle Size | 540-37 |
| 7742. 6212. 5821. 111. | .. MMD | 540-50 |
| 7742. 6212. 5821. 112. | .. NMD | 540-61 |
| 7742. 6212. 5821. 113. | .. VMD | 540-73 |

Some classificationists might prefer to put all terms related to aerosols in one section, which in COMEINDORS is done by employing the 100 schedules. In deep indexing, however, the term aerosols may be too broad and varied in its meaning to occupy a single position. The classification is therefore designed to be sufficiently flexible to accept additional entries that may be needed and to permit existing areas to be reorganized without disturbing the over-all system. Nevertheless, the success of a structured classification depends in large measure on the care with which categories in the arrays and chains are selected. In this connection, Vickery's concepts[9] and Ranganathan's canons[8] for arrays and chains have been particularly helpful in developing the COMEINDORS classification. The array for BW aerosols, shown in Table III, is a good example. The eight categories in this array are believed to encompass all foreseeable knowledge in the field of BW aerosols, to be mutually exclusive, and of helpful and consistent sequence. A ninth category can be added if needed or the whole array may be reorganized without disturbing the rest of the classification.

## THE GLOSSARY

The glossary provides an alphabetized entry to the system by any term that may occur to the searcher or indexer. Its role is strictly parallel to the index of Roget or the thesaurus of ASTIA. The ordered multiple word entries that correspond to the ascending generic tree of the classification accurately delineate the meaning of the terms. A five-digit short code or index shown in Tables II, III, V, and VI is common to the glossary and classification. It provides ready access to either from the other.

## THE LATTICE-TYPE CLASSIFICATION FORMS

The development of the COMEINDORS structured classification was an arduous and at times an exasperating task as long as conventional nested tables such as Tables I, II, III, and V were used. Although these are concise, they have a marked deficiency. Because they are essentially unidimensional, comparison of terms in such tables is limited to essentially adjoining terms. Moreover, because the items in such tables are written on adjoining lines on a sheet of paper, one may not appreciate the existence of useful spaces between adjoining items, nor may one readily appreciate the magnitude of these spaces that the long and short codes imply. Lattice tables were therefore designed to avoid these deficiencies. Forms for this purpose were printed on 11- × 11-in. paper in the form of a 10 × 10 equilateral grid with 0.75-in. graduations. The category to be divided was then written at the upper left corner of the grid. The terms into which this category is divided were then written across the top line of the grid at the junctures with the vertical lines. The junctures were also given their respective long codes. Each item along the top line forms a heading for a column that may be divided into an array of eight categories. By this means a single term inserted at the upper left corner of the table can be divided into a total of eight categories, and each of these in turn divided into eight subcategories, or a total of sixty-four categories in a single table. Such a table, reproduced in miniature as Fig. 1, permits four-way comparisons with adjoining terms. The division can then be continued by writing any of these categories in the upper left corner of another form, and this form in turn can be subdivided into sixty-four subcategories. When this second lattice is properly superimposed over the first lattice, as shown in Fig. 1, a multiple dimensional concept of the classification develops, the interrelations and inconsistencies among terms become evident, and the exist-

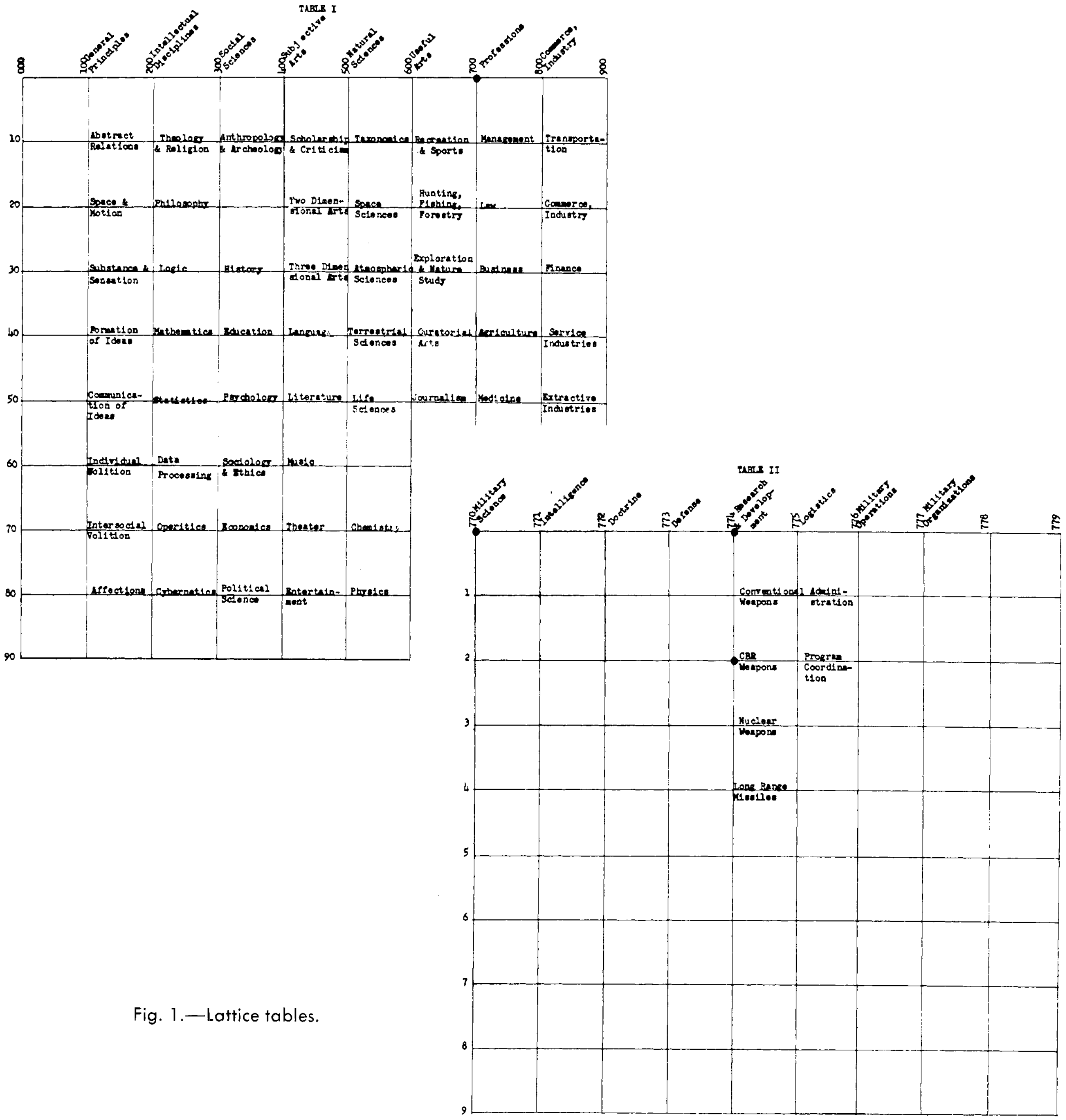

TABLE I

| | 000 | 100 General Principles | 200 Intellectual Disciplines | 300 Social Sciences | 400 Subjective Arts | 500 Natural Sciences | 600 Useful Arts | 700 Professions | 800 Commerce, Industry | 900 |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | | Abstract Relations | Theology & Religion | Anthropology & Archeology | Scholarship & Criticism | Taxonomics | Recreation & Sports | Management | Transportation | |
| 20 | | Space & Motion | Philosophy | | Two Dimensional Arts | Space Sciences | Hunting, Fishing, Forestry | Law | Commerce, Industry | |
| 30 | | Substance & Sensation | Logic | History | Three Dimensional Arts | Atmospheric Sciences | Exploration & Nature Study | Business | Finance | |
| 40 | | Formation of Ideas | Mathematics | Education | Languag | Terrestrial Sciences | Curatorial Arts | Agriculture | Service Industries | |
| 50 | | Communication of Ideas | Statistics | Psychology | Literature | Life Sciences | Journalism | Medicine | Extractive Industries | |
| 60 | | Individual Volition | Data Processing | Sociology & Ethics | Music | | | | | |
| 70 | | Intersocial Volition | Operitics | Economics | Theater | Chemistry | | | | |
| 80 | | Affections | Cybernetics | Political Science | Entertainment | Physics | | | | |
| 90 | | | | | | | | | | |

TABLE II

| | 770 Military Science | 771 Intelligence | 772 Doctrine | 773 Defense | 774 Research & Development | 775 Logistics | 776 Military Operations | 777 Military Organizations | 778 | 779 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | Conventional Weapons | Administration | | | | |
| 2 | | | | | CBR Weapons | Program Coordination | | | | |
| 3 | | | | | Nuclear Weapons | | | | | |
| 4 | | | | | Long Range Missiles | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | |
| 8 | | | | | | | | | | |
| 9 | | | | | | | | | | |

Fig. 1.—Lattice tables.

ence of many unused spaces is readily apparent. The lattice tables are thus a useful tool in the construction of the classification but do not form a final component of either the classification or glossary.

## THE INDEXING AND CODING FORMS

Printed indexing and coding forms have been provided to assist the document indexers and card-punch operators in their operations. The index form contains terms that recur frequently in a particular group of documents that is being indexed. A pair of brackets follows each term. A list of links and roles with their respective codes is in a box at the upper right corner of the form. Near their lower margins, the index and coding forms contain the tracing information and other terms needed to identify the document. In addition, the code form contains the short codes that correspond in context and in exact position to their respective terms on the index form. In practice, the forms are assembled in pairs in pads, first an index form and below it a code form. The indexer inserts a third, blank sheet of paper below such a pair. Carbon paper is then interleaved. The tracing information that is needed to identify the document is written in or checked at the bottom of the indexing form, and the lower carbon paper is then removed. The indexer then inserts between the brackets of any pertinent term on the index form any pair of codes from the table of "roles" that corresponds most nearly to the importance or role of the term in the investigation. Useful terms found in the document that do not appear on the index form are written on the third, blank sheet along with the pertinent pair of role codes. An information specialist or classificationist checks the adequacy and correctness of indexing, assigns codes to the terms written on the third sheet, retains the indexing form for his records, and transmits the code form and the third sheet to the card-punch operator. A table of links and roles is shown in Table IV.

Table IV
Table of Links and Roles

| Subject | Code[a] |
|---|---|
| Major importance in the document | 01 |
| Minor importance in the document | 03 |
| Importance not applicable | 05 |
| Cause | 02 |
| Effect | 04 |
| Cause or effects not applicable | 06 |

[a] These codes will be used as the sixth and seventh digits in the short code.

## COMPARISONS OF INDEXING IN THE COMEINDORS AND ASTIA SYSTEMS

Information specialists recognize that, as large multiple disciplinary information systems grow progressively larger, it may be necessary to restrict the depth to which the systems are indexed or to provide a series of microthesauri or satellite systems in order to ensure efficient operation. The Defense Documentation Center (DDC), formerly the Armed Services Technical Information Agency (ASTIA), recognizes that this situation presently exists even with its recently issued second edition of its thesaurus. The need for greater depth of indexing than is provided by the ASTIA thesaurus is shown by a comparison of searches for the terms "Particle Size," "Aerosols," and "Impactors" in the ASTIA thesaurus and in COMEINDORS. The ASTIA thesaurus gives nothing specific on any of these terms. The ASTIA Code Manual merely indicates that 1,019 documents in its collection have been indexed under "Particles" but no specific term "Particle Size," nor breakdowns for such a term, are listed even though this is a major subject in a number of scientific and industrial fields. Similar results are found in the use of the terms "Aerosols" and "Impactors." Many similar examples might be found in a number of disciplines. The existence of such differences should not be regarded as a discredit to any particular system. Indeed, they should be accepted as the inevitable result of dynamic knowledge in highly specialized fields of endeavor.

The ease with which specialized retrieval systems such as COMEINDORS can cope with such situations and provide concise meaning to its terms is shown in Tables V and VI. Table V, an extract from the COMEINDORS Glossary, indicates that its first entry, "Particle Size," is related progressively to the larger topics, "Physical Properties," "Fill Properties," and "Evaluation." Noting its short code, 629–09, which has been marked by an asterisk for this particular illustration, one looks for the same short code in the Classification Schedule, a portion of which is extracted in Table VI. As one reads up the Classification from this entry he again finds the context in which this Particle Size entry is used, and may continue to trace its context as far up the Schedules as he wishes.

Table V
Glossary Extract

| | Short code |
|---|---|
| Particle Size, Physical Properties, Fill Properties, Evalua | 629–09* |
| Particle Size, Temperature-Humidity Interactions, Meteorol | 604–15 |
| Particle Counts, Temperature-Humidity Interactions, Meteor | 604–08 |
| Particle Size, Humidity, Meteorological Factors, Evaluatio | 603–15 |
| Particle Counts, Humidity, Meteorological Factors, Evaluat | 603–18 |
| Particle Size, Temperature, Meteorological Factors, Evalua | 602–25 |
| Particle Counts, Temperature, Meteorological Factors, Eval | 602–19 |

Table VI
Classification Schedule Extract

| Long code | Short code |
|---|---|
| 7742. 6212. 67. . . . Evaluation of Product | 628–01 |
| 7742. 6212. 671. . . . Fill Properties | 628–12 |
| 7742. 6212. 6712. . . . Physical Properties | 629–01 |
| 7742. 6212. 6712. 1. . . Particle Size | 629–09* |
| 7742. 6212. 6712. 11. . . MMD | 629–17 |
| 7742. 6212. 6712. 12. . . NMD | 629–25 |
| 7742. 6212. 6712. 13. . . VMD | 629–33 |
| 7742. 6212. 6712. 2. . . Specific Gravity | 629–41 |
| 7742. 6212. 6712. 3. . . Surface Tension | 629–50 |

Similarly, reading down the Schedules he is able to see at a glance the specific components into which this entry is divided, and additional terms that are coordinate with it. No confusion should arise as to the specific meaning of this particular entry. Other meanings of the same or related terms, designated with equal preciseness by the same method, may be traced by means of the short code. The requestor need not be concerned with the long codes that are used by the computer for searching.

An obvious solution to the problems posed by the large, comparatively shallow indexed information system, such as the ASTIA thesaurus, is to develop detailed or micro-thesauri.

## OPERATION

As it is presently constituted, COMEINDORS is a semi-mechanized system with the information from the indexed documents stored in several thousand computer cards and in several print-outs. No fixed searching procedure is followed. Requests are still sufficiently few in number to permit each request to be processed individually according to its particular needs. The computer has been used during the development of the system and may be used for searches, including high-speed print-outs.

The operation of the system can be illustrated by a special procedure devised to answer a request. The request was for references to all tests on a given subject, but references were to be subdivided into fourteen subcategories to permit retrieval of documents in all possible combinations of the fourteen subcategories. The conventional hierarchical classification system might have been unable to cope with such a request. A special program might have been written for the computer, but this would have resulted in some delay. The request was therefore handled in the following manner. Because the system's punch cards happened to be sorted with one decklet containing all the cards of the main category wanted, no preliminary sorting was needed. The concerned decklet, which happened to contain 157 references, was passed through a tabulator and a print-out was obtained with all test numbers listed in numerical order in the left column. The fourteen subcategories with their respective short codes were then typed across the top of the print-out. Lines were drawn to identify the rows and columns. A previously prepared print-out that contained all test numbers in the collection, segregated according to subcategories and short codes and test numbers, was then consulted for the desired fourteen subcategories. Whenever a test number was found in a subcategory that corresponded to one of the test numbers in the left column of the request print-out, a check mark was placed in the corresponding box. The resulting table, prepared in less than 2 hr., permitted the requester to obtain all documents with any desired combinations of subcategories. Now that this procedure has been developed, future complex requests of this nature can be processed even more promptly.

## REFERENCES

(1) C. J. Maloney, "Machine Indexing of Literature, Superposition, Coding, and Information Theory," paper presented at the American Statistical Society, 114th Annual Meeting, Montreal, Canada, 1954.

(2) C. J. Maloney, "Abstract Theory of Retrieval Coding," Proceedings of the International Conference on Scientific Information, Washington, D. C. (Natl. Acad. Sci.-Natl. Res. Council), November, 1958, pp. 1365–1382.

(3) C. J. Maloney, *Am. Doc.*, **9**, 1 (1958).

(4) C. J. Maloney, *Current Res. Dev. Sci. Doc.*, **7**, 44 (1960).

(5) C. J. Maloney and H. W. Batchelor, *ibid.*, **8**, 49 (1961).

(6) C. J. Maloney and H. W. Batchelor, *ibid.*, **9**, 78 (1961).

(7) C. J. Maloney, *Am. Doc.*, **13**, 276 (1962).

(8) S. R. Ranganathan, "Prolegomena To Library Classification," The Library Association, London, 1957.

(9) B. C. Vickery, "Faceted Classification, A Guide to Construction and Use of Special Schemes," Aslib, London, 1960.