Selective Dissemination and Indexing of Scientific Information

Improved indexing methods will increase the precision of matching scientists with useful published documents.

John H. Schneider

Automated methods for selective dissemination of information (SDI) to individual scientists and engineers play an important role in dealing with the increasing avalanche of scientific information. This article presents some basic aspects of SDI systems and describes recent developments and problems. Two different approaches to indexing information for SDI systems are discussed, with emphasis on the desirability of using enumerative hierarchical classifications to improve the precision and quality of matching scientists with useful documents.

Role of SDI in the Intellectual Support of Scientists

The need for strong intellectual support of scientists and engineers is often overlooked by research administrators who do not realize how lack of information can seriously hamper research and development programs. Shaw suggests that information input may be just as essential for creative research as laboratory equipment or technicians (*1*). Scientists require several years of formal intellectual training to build up a store of basic knowledge before they are considered to have the minimum qualifications for professional work. Yet when they begin specific research tasks where current knowledge and constant updating are vital, they are expected to keep up with the flood of journal articles and reports on an informal, random, haphazard basis as time permits. Many professionals spend at least 25 percent of their time trying to keep abreast of developments in their field (*2*). In spite of this effort, a given individual is increasingly likely to miss useful information as the literature continues to expand far beyond his ability to examine all the sources that might contain valuable data. One study indicates that more than 30 percent of scientific manpower is wasted because needed information is unavailable to the right person at the right time (*2*).

Journals are becoming more specialized, reflecting the fragmentation of science into smaller areas, with an increasing number of articles published in an increasing number of journals. Estimates suggest that 120 million pages of scientific information were published in 100,000 journals in 1970 and that the number of pages doubles every 8.5 years (*2*). Even secondary literature services are reaching an unmanageable volume, with 1,300 abstracting services in science, technology, medicine, and agriculture (*3*). The resulting problems of literature dispersion, dilution, and scattering are partially solved by SDI systems, which identify, concentrate, and "purify" specified types of information.

Selective dissemination of information is a type of personalized current-awareness service which, under optimum conditions, involves screening a large number of documents, selecting information exactly tailored to meet the specific, unique research needs of each user (usually by an automated process), and supplying this information directly to each individual on a dependable, continuous basis. Systems for SDI do not eliminate the need for scanning or browsing through books and journals, nor do they diminish the value of scientific meetings or of verbal communication. But, by precise matching of users and documents, SDI systems *supplement* the researcher's efforts to keep abreast of current literature and help to ensure that documents directly related to current research are not overlooked. Data summarized in this article show that SDI systems can supply research scientists with a high percentage of *useful* information, most of which was not previously known by the scientist.

The cost of SDI systems is small in comparison with the total cost of research. For example, a feasible target figure of $200 spent each year to select and send summaries of articles to each scientist is less than 0.5 percent of the $45,700 average annual cost of supporting each industrial R & D scientist or engineer in 1968 (*4*).

Need for Precise Matching in SDI Systems

Scientists are often apathetic toward information services and follow a "principle of least action" in using them (*5*). For this reason, SDI systems should provide actual information such as summaries or abstracts as well as references and should require minimum effort from users of the system. More important, the best SDI systems should have a high degree of selectivity or precision (a high signal-to-noise ratio) when documents are matched against the information needs of an individual user. Users rapidly become disenchanted with systems that are unable to select a high proportion of relevant documents (*6*). Low precision results in a waste of time and money, first by the SDI system that selects and sends useless data, and then by the user who must spend considerable time looking through an excessive amount of "trash."

Figure 1 is a target model of precision in an SDI system, which emphasizes the advantage of using a rifle rather than a shotgun for disseminating scientific information. Ideally, a system should identify all information directly related to the subject in the center of the target. No matter how good the system, some information will always fall in the outer circles. If matching is poor, a high percentage of the information will be completely off the target. Different scientists, of course, require circles of different sizes. The best systems permit easy, accurate adjustment

The author is scientific and technical information officer of the National Cancer Institute, National Institutes of Health, Department of Health, Education, and Welfare, Bethesda, Maryland 20014.

of the level of generality or specificity to accommodate various appetites for information. If an SDI system is really selective, it must be able to discriminate among the various circles, no matter what their relative sizes may be. If it cannot relax (broaden) or tighten (narrow) the criteria for matching to meet the precise needs and wishes of each user, then an SDI system is just one additional, relatively expensive information source, little better than those already available.

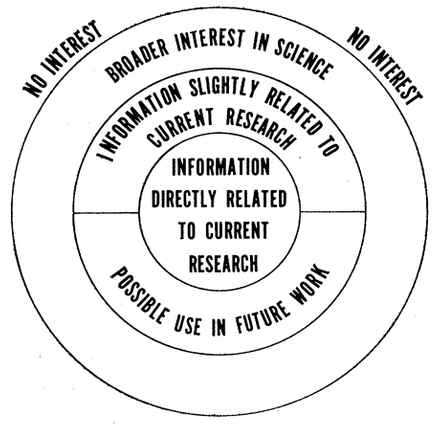


Fig. 1. Target model illustrating the need for precision in matching scientists with information.

Development of SDI Systems

Automated SDI systems were first proposed only 13 years ago (*7*) but grew rapidly with the development of high-speed computers and efficient operating systems. A listing compiled by Share Research Corporation in February 1969 showed 73 SDI systems already installed by industry, government, and academic organizations, and 18 systems in planning stages (*8*). Housman had gathered data on operational SDI systems in 64 different organizations as of June 1969, with pilot or planned systems in 29 others (*9*). More than 80 percent of these organizations are engaged in research and development in the physical sciences. Several other surveys of SDI systems have been completed (*10*) or are in progress. The literature describing specific systems is covered in various reviews (*11*, *12*). Although most SDI systems deal with published articles, reports, or abstracts, several cover patents and new chemical structures (*13*).

The rapid development of SDI systems is due in large part to the production of magnetic tape data bases containing titles, references, and sometimes index terms or abstracts of published articles and reports. Data describing approximately 90 percent of the 854,000 articles processed in 1970 by 17 organizations belonging to the National Federation of Science Abstracting and Indexing Services (NFSAIS) are on magnetic tape. Carroll has published data on 37 different tape services directly related to physical sciences, plus nine tapes that cover other subject areas (*14*), and Gechman (*15*) has described 40 technical tape data bases. Since retrospective searching of tapes is quite expensive (*16*), their main use is for input to SDI systems.

Although many organizations prefer to set up their own internal SDI systems, SDI services may also be purchased from at least 14 different suppliers (*15*). A group of those interested in providing and using magnetic tape data bases for SDI services and retrospective searching was formally organized under the acronym ASIDIC (Association of Scientific Information Dissemination Centers) in October 1969 after a year of informal meetings. By October 1970 it had 19 full and 32 associate member organizations (*17*). A similar association (EUSIDIC) has been formed in Europe (*18*). Both organizations publish periodic newsletters and hold regular meetings. At least 512 individuals belong to a Special Interest Group for SDI which was established for members of the American Society for Information Science late in 1968 and held its first meeting in October 1969. A conference on SDI systems with 55 participants, held at the University of Wisconsin in April 1969 (*19*), discussed developments, problems, and future directions of SDI operations.

The plethora of magnetic tape data bases and tape-processing operations has created a whole new generation of problems in the storage and dissemination of scientific information. Processing of the same article by two or more services results in duplicated effort and increased cost for both producers and users of tapes (*20*). Because each tape service covers only a limited range of topics, SDI profiles must often be matched against tapes from two or more sources at an annual cost ranging upward from about $100 per tape service used on a fee-for-service basis (*16*). The long-term viability of organizations furnishing such services is uncertain because of increasing competition for a limited number of potential customers and because of relatively high operating costs.

Tape processors must pay tape producers $1,700 to $10,000 a year for each tape data base, depending on the number of documents and type of data (ranging from simple title and references to abstracts with many index terms). Additional costs include rental of computers and the salaries of programmers, computer and keypunch operators, and information specialists who formulate SDI profiles for individual users. Because each producer of tapes has developed a uniquely different format, tape processors must often reformat existing tapes. To improve the efficiency of searching, the reformatted tapes are usually processed again to create special inverted files or to combine data from several tapes. Different ways of coding data from one year to the next can cause special problems for multi-year searches.

Apart from these problems in the practical world of operational SDI systems, a number of basic considerations appear to have been neglected. It is none too soon to be concerned with the quality and optimization of SDI systems and the organization and analysis of input so as to provide the best possible match between scientists and documents at the lowest possible cost. Small-scale SDI experiments with extensive evaluation and novel indexing procedures, such as one described in this article, may be useful in improving the design of second-generation SDI systems.

Nature of SDI Systems for Individual Scientists

Almost all information services and products other than SDI are prepared for groups rather than for individuals and have a *passive* element—a "try-to-find-it" orientation—requiring each user to search through much useless information for a few useful items. Fortunately, the trend is toward progressively smaller, more specialized groups, as reflected by the titles and scope of new primary journals and new abstract bulletins. Several services now divide their abstracts or references into many small divisions or topics, which are printed on separate lists, cards, or microfiche and sent to individuals who

select specific topics or "group profiles" (*21*).

In the usual sense of the concept, SDI refers to services that match documents against the needs of a single individual and his co-workers rather than a larger group of individuals. In this way, SDI systems are entirely *active*, automatically providing information to each user on a "here-it-is" basis, with a minimum amount of extraneous or useless data. In addition, by continuously screening published articles, SDI systems provide each user with a constant stream of up-to-date information. Most other types of information products and services must be used in a discontinuous, retrospective, "on-demand" mode whenever the user has time to query an on-line computerized system, to fill out forms that can be processed by a search specialist at an information center, or to examine published indexes or abstract bulletins.

The common components of most SDI systems are shown in Fig. 2. The information needs of each user are converted by an indexing process into a "user profile" or "interest profile." Indexed documents are matched against user profiles by a computer. Whenever a "hit" or "match" occurs between an interest profile and an article or report, the user is sent information about the matching document. The user returns evaluation slips, which are sometimes used to change the interest profile and the index terms.

Most matching techniques involve selection and coordination of index terms, just as pieces of a puzzle are joined together to form a picture. One type of coordination uses Boolean logic statements such as "(Term A or Term B or Term C) and (Term D or Term E) but not (Term F or Term G)." Another type of coordination uses weighting techniques: each useful search term is assigned a numeric value, and the values of all index terms assigned to a document must add up to more than a minimum threshold to achieve a match. Some SDI systems use feedback from users to automatically compute fractional adjustments in the weight of index terms in user profiles. Terms yielding a good match are given increased weight, whereas terms resulting in a poor match are slightly reduced. The ASCA (Automatic Subject Citation Alert) system (*22*) has the additional and unique capability of identifying a match whenever the indexed document cites (gives as a reference)

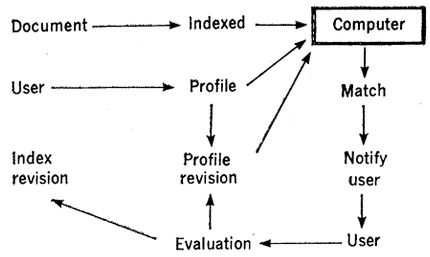


Fig. 2. Basic components of SDI systems.

previously published articles that are specified by each user.

A rarely used alternative to coordination of simple index terms or subject headings involves development of a detailed hierarchical subject classification which is so specific that an article indexed under any *one* of the categories assigned to a user has a high probability of being useful to that user. The following paragraphs describe this type of "single-hit matching" which has considerable potential for high precision and ease of operation in SDI systems.

INSTRUCTIONS - Please answer questions 1, 2, and 3 **for each summary received.** Return this form in the enclosed envelope.

1. Is this article OF INTEREST to you? *(Check one)*
- ☐ Of considerable interest
- ☐ Of moderate interest
- ☐ Of little or slight interest
- ☐ Of no significant interest

Comments

2. Is this article RELATED TO YOUR CURRENT RESEARCH? *(Check one)*
- ☐ Very close relation
- ☐ Direct but limited relation
- ☐ Slight or indirect relation
- ☐ No significant relation

Comments

3. How USEFUL is this article to you? *(Check one)*
- ☐ Very useful
- ☐ Definite but limited use
- ☐ Little or slight use
- ☐ No significant use

Comments

ANSWER QUESTIONS A, B, AND C ONLY IF ARTICLE WAS OF INTEREST TO YOU.

	Yes	No
A. Did you know about this research before it was published? *(Check yes or no)*		
Comments		
B. Were you already aware of this article? *(Check yes or no)*		
Comments		
C. Will you (or did you already) (1) retain this summary for files or reference? *(Check yes or no)*		
(2) order a reprint or make a copy of this article? *(Check yes or no)*		
(3) examine the article for details not given in the summary? *(Check yes or no)*		

YOUR COMMENTS will be of significant help in evaluation. Please make any such comments on the other side.

DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
PUBLIC HEALTH SERVICE
NATIONAL INSTITUTES OF HEALTH
SELECTIVE DISSEMINATION OF SCIENTIFIC INFORMATION
EVALUATION RECORD

Fig. 3. Example of a questionnaire used for in-depth evaluation of an SDI system.

Evaluation of an Experimental SDI System

The principal investigators of 137 research grants supported by the National Cancer Institute were invited to evaluate an experimental SDI system (*23*) based on linear, hierarchical, decimal classifications (*24*), and 104 scientists agreed to participate in the test (*25*). All had either an M.D., or a Ph.D., or in some cases, both. Although their research was cancer-related, it covered a wide range of biological disciplines and experimental approaches (*26*).

During the test, numbers representing categories in a very detailed subject classification were assigned to 1396 articles published in 12 major cancer research journals. Similar category numbers were also used to index the research interests of each scientist, as described in his most recent application for a research grant. When any *one* of the numbers assigned to a scientist matched any *one* of the numbers used to index the article, the scientist was sent a copy of the title page, which usually included the full reference, mailing address of the author, and the author-prepared summary of the article. Missing information was written or taped on the title page before copying (*27*). Matching was performed by a Fortran program written for an IBM 360/50 computer. During the 1-year evaluation period, participating scientists received 6458 summaries of articles. An evaluation slip (see Fig. 3), mailed with each summary, offered four possible levels for rating each article on the basis of interest, relation to current research, and usefulness. Nearly 82 percent of the evaluations (5278 slips) were returned (Fig. 4).

These evaluations showed that over 90 percent of the articles were of some interest; 68 percent were of either considerable or moderate interest. More important, 83 percent of the articles were somewhat related to current research, and 82 percent were of some use. Considering only the top two rating levels, 54 percent of all articles evaluated had a "very close" or "direct" relation to current research and were either "very useful" or of "definite" use (see Fig. 4). These ratings based on usefulness and relation to research are more meaningful than rather vague concepts of "relevance," "pertinence," or "value," which are often used for evaluating information systems.

Potential Value of SDI

Services and Published Data

The value of *published* research results is sometimes downgraded. It has been suggested that scientists and engineers know about work in their specialties long before publication because of "invisible colleges"—groups of scientists who exchange letters and preprints, attend the same professional meetings, visit and converse with each other frequently, and participate in other informal information exchange activities (*28*).

On the contrary, data in Fig. 5 suggest that published articles do supply much additional useful information: 76 percent of the articles rated as "very useful" describe research *not known* to the scientist before it was published; 88 percent of the articles of "definite but limited use" described unknown research. Even after delays due to indexing, matching, and mailing the summaries, scientists still did not know about 61 percent of the articles that they rated as "very useful" or about 72 percent of the articles rated of "definite but limited use." Thus, SDI systems have considerable potential for supplying individual scientists and engineers with summaries of useful reports that are directly related to their current research and contain previously unknown information. Retrospective searches at the National Library of Medicine have also identified many articles of "major value" that were unknown to requesters prior to searches. Lancaster proposes that a "novelty ratio" (*29*) be used to measure this lack of familiarity with useful literature.

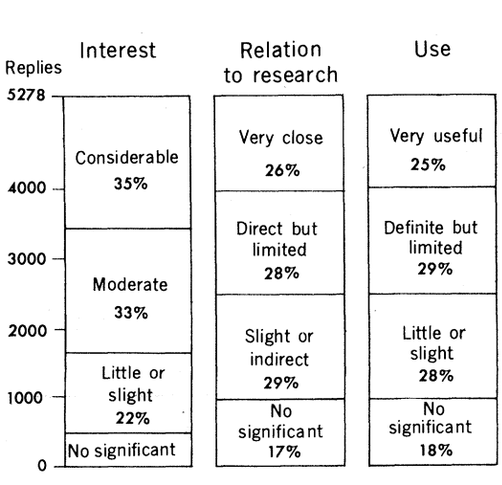
Hierarchical Classifications for Matching Users with Information

The precision of matching users with useful documents depends on the method employed for indexing information in documents and information needed by users. In fact, the value of an SDI system, or any information system, may well be directly related to the amount of intellectual effort expended on indexing.

One effective indexing method is based on subdividing fields of knowledge into hierarchical categories or classifications. Because the words "classification" and "hierarchical classification" are used in many ways, I will avoid semantic ambiguities and cumbersome wording by using the acronym HICLASS when referring to a special type of enumerative HIerarchical CLASSification that can be developed for a variety of scientific fields (*24, 30*) and used in automated systems (*23, 31*). The adjective "enumerative" is important because HICLASS classifications consist of very detailed listings of significant subdivisions or categories of information. Each category, described by one or more phrases or sentences in natural language, represents a complete concept or subject. Isolated words and terms are seldom used in a HICLASS system unless they are in a list of chemicals, organisms, or other items that require no further qualification or explanation.

The SDI experiment described above employed an automated HICLASS system. To illustrate the indexing method, I have chosen an article entitled "The Pathological Effect of Subcutaneous Injections of Asbestos Fibres in Mice: Migration of Fibres to Submesothelial Tissues and Induction of Mesotheliomata" (*32*). The indexing for this article is represented diagrammatically in Fig. 6. Each subject area is described by a series of concentric circles representing linear, general-to-specific logic. The meaning of the large set of circles in Fig. 6 is indicated by words within each circle. In the top set of circles on the right, logic from largest to smallest circle is: cancer (51.); agents that cause cancer (51.4); selected environmental and naturally occurring agents that cause cancer (51.43); minerals that cause cancer (51.4323); and a smaller circle (omitted in the diagram) for cancer induction by asbestos. Other sets of circles (represented by the unnumbered set in Fig. 6) are used for other important concepts in the article such as logic leading from general toxicology to environmental toxicology, to toxicity of minerals, to toxicity of asbestos.

It is not possible in this paper to describe in detail how subject categories are developed for a HICLASS classification or how they are used for



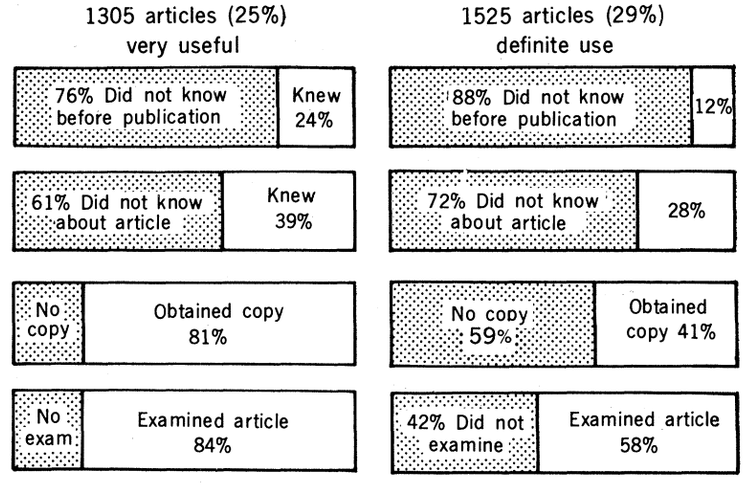


Fig. 4 (left). Summary of replies to questions 1 to 3 of the quest ionnaire shown in Fig. 3. This diagram and Fig. 5 are based on responses to 6458 summaries of articles selected for 104 scientists by an automated SDI system. Fig. 5 (right). Summary of replies to questions A to C of the questionnaire shown in Fig. 3. T he bars on the left apply to articles rated as "very useful"; the bars on the right apply to articles rated "definite but limited use" (see Fig. 4). The top line of bars indicates the percentage of articles for which results were known before they were published. The secon d line of bars indicates the percentage of articles that scientists were aware of before receiving a summary from an SDI service. The bottom two lines of bars indicate whether a copy was obtained or the article was examined for more detail.

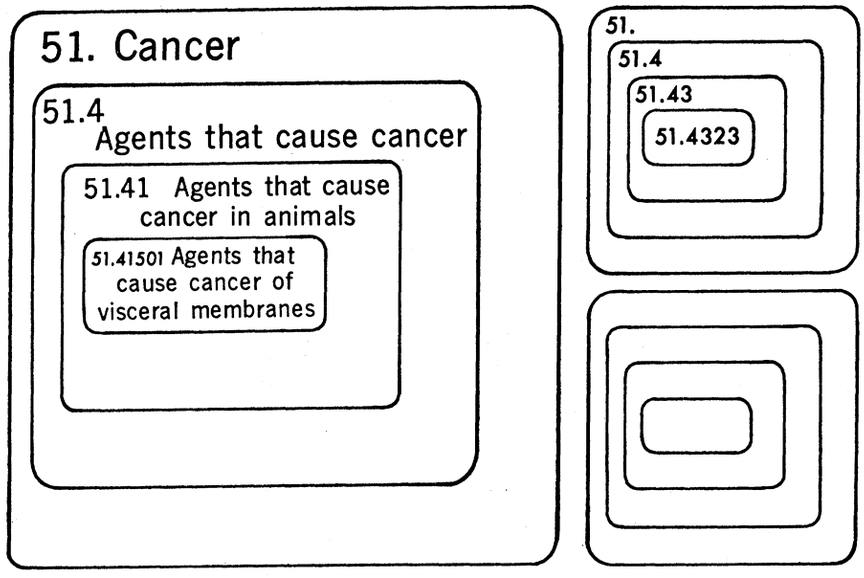


Fig. 6. Diagram of the linear, general-to-specific logic used to index and retrieve information in HICLASS systems based on enumerative hierarchical classifications. The multiple sets of circles represent different, integral concepts in an article on the causation of mesotheliomas by asbestos in mice.

indexing in an SDI system, but several important points should be noted.

1) Several numbers representing different concepts or categories are assigned to each scientist and document, as illustrated by multiple sets of circles for one article in Fig. 6. In the test of a HICLASS system described above, each scientist was initially assigned an average of 16.8 categories (requiring 26 minutes of indexing time per scientist). Because of profile changes, this average rose to 20 categories per scientist after about 14 months of operation. Each document was indexed by assigning an average of 3.8 category numbers in 5.3 minutes. Assignment of *multiple* category numbers to one document is possible because documents are physically separated from index terms in an automated system, in contrast to the classical procedure of using a single category number to "mark and park" a document on a library shelf.

2) In a HICLASS system, each category number represents an intact concept, complete in itself, which can be assigned, matched, or altered without reference or relation to any other category. Thus, the integrity of a scientific concept is preserved during all stages of indexing and retrieval operations. For example, Table 1 shows the profile of a scientist with a relatively large number of research interests. Single-hit matching of any *one* of these HICLASS categories (including all subdivisions of each category) with any *one* of the numbers assigned to a document will identify an item with a high probability of being useful to this scientist. If a returned evaluation slip indicates that one of these numbers caused a poor match, then the number can be deleted, subdivided to give more specificity, or otherwise altered without affecting any other number in the profile.

3) Classifications are simply tools for arranging subjects in some reasonable order that shows logical generic relationships. The categories must be flexible, with constant evolution to accommodate new concepts; they should not be regarded as sacred, immutable constructions requiring high-level benediction for every change. The best-known classifications (Dewey, Library of Congress, Universal Decimal Classification) were developed long ago for use in libraries and tend to be out of date because of resistance to modification, which is understandable when hundreds of different libraries use the same standard classification. However, many systems currently engaged in indexing published articles and reports use unique in-house indexing methods that are not used elsewhere. In these highly centralized information centers, a proposed change in a classification would require no more than verbal approval by one or two specialists. Under these circumstances, it requires less effort to develop new, up-to-date classifications that are free of restrictions on future expansion or change than to push revisions of existing classifications through sluggish administrative channels.

Table 1. Initial SDI profile for one scientist. Category numbers taken from enumerative hierarchical classifications (*24*) are assigned to scientists' interests on the left. The meaning of each number is given on the right.

Category No.	Brief explanation
11.137	Dye binding to nucleic acids
15.1224	Function of vitamin D
32.3	Molecular binding studies
44.161	Hormone-calcium interactions
44.4	Parathyroid (gland and hormone)
47.3512	Immunoglobulin G (IgG)
47.3514	Immunoglobulin M (IgM)
47.352	Methods for gamma globulins
47.353	Properties of gamma globulin
47.8532	Thymus antigens
48.422	Histocompatibility genetics
48.423	Histocompatibility antigens
51.45122	Antigens of cancer viruses
51.525524	Thymus leukemia in mice
51.743	Tumor antigens
53.2543	Lymphoid aspects of thymus
53.841	Mineralized tissues, general
53.8429	Osteolysis of bones

Use of Keywords to Index Scientific Information

Almost all operating SDI systems use various types of word lists, rather than classifications, for matching users with documents. Indexing with individual words and terms requires what might be called the Humpty-Dumpty approach: concepts are deliberately broken into individual words and terms during indexing, and an attempt is made to put the concepts together again during retrieval operations. This approach to indexing (for the same article discussed in relation to Fig. 6) is shown schematically in Fig. 7. Index terms derived from the original concepts are rejoined to form an area of overlap during retrieval.

As suggested by words outside each circle, the terms can be chosen from general subject areas or categories (sometimes referred to as facets or disciplines) such as body part, disease, type of animal, or substance. In other cases, the index terms (often called keywords, subject headings, descriptors, identifiers, or uniterms) come from alphabetical listings or other word lists (subject-authority lists, special dictionaries, and controlled vocabularies). More sophisticated word lists in which relationships (broader terms, narrower terms, related terms, preferred terms)

are indicated are called "thesauri" or "classed thesauri." In some cases, related words or subject headings are grouped together, arranged hierarchically, and assigned code names or numbers, resulting in structures called "trees," with "explosion" or "expansion" of major terms into subordinate terms. Although these various names for index terms and word lists imply differences, many of them are near-synonyms, referring to very similar sources and types of terms for indexing documents. In spite of minor differences in meaning, these words will be used interchangeably in this article.

No matter how the list of index terms is organized, it is used to find keywords or terms for indexing the component *parts* of a concept such as "mesothelioma induction by asbestos in mice." After breaking a concept into basic terms during indexing, the words must be coordinated or rejoined during the search and retrieval operations to form an intersection which, for the example given in Fig. 7, represents articles indexed under "cancer and mesothelium and asbestos and mice." This use of coordination is clearly different from the single-hit approach based on HICLASS categories with the general-to-specific logic represented by Fig. 6.

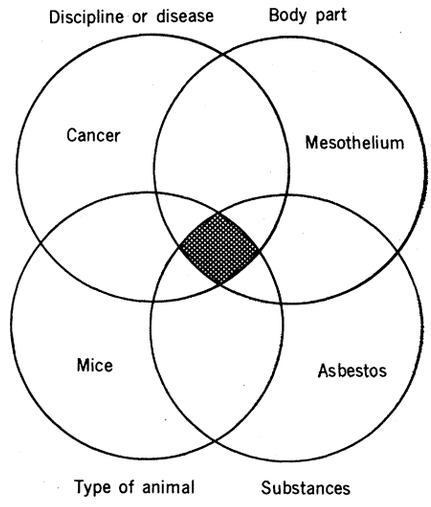


Fig. 7. Diagram of coordination logic required when information is broken into keywords or other isolated terms during indexing and is then recombined during retrieval. This is an alternative way of indexing the article referred to in Fig. 6.

Comparison of Keyword-Based and Classification-Based Indexing

Most operating information systems for processing scientific information are based on the use of keywords, descriptors, thesauri, subject headings, and similar indexing devices described in the previous section. Very few automated information systems are based on classifications (*23, 31, 33–35*). If the cost of *developing an indexing method* were the only consideration, then the overwhelming preference for keyword systems might be justified, since it is relatively easy to select keywords, arrange them alphabetically, and specify some minimum generic relations between them. Detailed classifications are more difficult to develop, since considerable understanding and analysis are required to divide subjects into subareas and to group related categories into adjacent or neighboring locations in a hierarchy.

Indexing operations based on classifications tend to require more intellectual input than keyword indexing, which is sometimes performed entirely by computer processing of machine-readable data. For example, some systems use computers for "free text" or "normal text" searching, with every word in the title, abstract, or full text treated as a potential index term (*36*). Other computer systems select the most significant keywords from the text by automatic indexing, automatic text analysis, or automatic classification procedures that involve complex syntactic and semantic analysis (*37*). The ability to use these computerized methods to index documents with minimum human participation is a major reason for the popularity of keywords as compared with classifications in information systems.

Classifications are open-ended and are easily expanded to include new ideas or increasing depth of detail. As a result, they tend to grow rapidly, often to an intimidating size. It has been difficult to update and use the bulky classifications, but conversion of classification schedules to magnetic tape format with automated updating (*38*) may overcome this problem. In contrast, thesauri and alphabetical lists of index terms can be revised with relative ease, since they usually consist of a limited number of words in a controlled, rigidly restricted vocabulary that changes only slightly from year to year. The updating problem is almost nonexistent for "free text" or "natural language" systems which use any significant word occurring in a title or abstract for "indexing" (sometimes after slight grammatical changes).

Despite these drawbacks, detailed enumerative classifications of the HICLASS type have at least four very important advantages over the use of keywords for indexing:

1) The HICLASS categories have a high degree of built-in "pre-coordination" *within* each category. Concepts may be indexed, processed, and retrieved as intact, integral units corresponding to single HICLASS categories in the classification. Other types of classifications, such as faceted classifications, also use pre-coordination, but this often involves coordination of two or more separate categories during indexing operations. In contrast, systems based on keywords have minimum pre-coordination and require a large amount of "post-coordination" or recombination of keywords during search and match operations.

2) Because classifications specify the generic-specific relations between concepts, they facilitate indexing at any desired level of generality or specificity. Expression of generic-specific relations is often quite complicated in keyword systems, even when the words are organized into "trees." As a very simple example, the intersection of circles in Fig. 7 should include articles on cancer induction by chrysotile, crocidolite, or amosite (specific forms of asbestos) in epithelial linings of the body cavity, the scrotal cavity, the pleura, omentum, peritoneum, and pericardium, although none of these terms appear in the diagram. Because of the need to include all possible combinations of alternative terms at various generic-specific levels and to exclude unwanted combinations, it is not unusual for search specialists to spend an hour or more formulating the search strategy required to retrieve answers even for relatively simple questions. In at least one large keyword-based system, searchers average less than 20 searches per week. On the other hand, at the Science Information Exchange, which uses a detailed hierarchical classification, only 5 to 20 minutes is required to analyze a request and complete a computer search form that shows numeric codes for all topics requested and indicates the optimum search strategy (*39*).

3) A high degree of precision in indexing or matching ideas is possible with HICLASS categories because each concept is expressed fully and clearly in normal English, whereas other systems involve awkward and sometimes incomplete synthesis of an idea by coordination of component keywords.

This advantage of classifications is particularly important for information dealing with the social sciences, education, humanities, politics, and other subject areas, where relatively common words with all their subtlety and ambiguity are often used in technical documents. Physical sciences are fortunate in having highly technical, unique terminology that guarantees at least a minimum degree of success no matter how naive the system or how inadequate the indexing vocabulary. Even for the physical sciences, however, concepts can be more precisely identified by numbers referring to well-defined categories than by coordination of several keywords.

4) In classifications, any given category number represents the same concept, no matter how many synonyms, abbreviations, singular and plural word forms, or other grammatical or linguistic variations are used to express the idea in the original text. The use of a common category number to describe a single concept, regardless of language or terminology, facilitates indexing and should promote international cooperation in the indexing and dissemination of scientific information.

Tests of Indexing Languages

The more effort devoted to organization and analysis of data during the input stage of an automated system, the easier it is to retrieve "clean" information rapidly with minimum noise or trash. The alternative is to dump raw, unorganized keywords into an automated system and then spend a large amount of time trying to retrieve organized, meaningful information. Although a direct relation between effort expended during input and the quality, extent, and ease of retrieval seems obvious and reasonable, little objective evidence can be cited to support such trade-off.

Three major studies have compared indexing languages requiring various degrees of analysis during indexing. After an extensive study of several procedures at the Comparative Systems Laboratory at Case Western Reserve University, Saracevic was unable to determine whether "the indexing languages tested are either a significant or not a significant factor which affects the performance of retrieval systems" (*40*). Salton has found that automatic analysis of text, with index terms chosen entirely by a computer and programmed logic, is at least as effective as more conventional manual (intellectual) indexing methods, and that a hierarchical arrangement of terms in a thesaurus does not appear to be sufficiently promising to advocate its use in a system for automatic retrieval (*37, 41*). Neither of these two studies included an enumerative classification in their comparisons.

The most widely quoted results are those of Cleverdon (*42*), who initially found that four different indexing methods—the Universal Decimal Classification, alphabetized subject headings, faceted classifications, and uniterms—all operate at about the same level of efficiency. Subsequent tests, however, did show consistent differences in the performance of 33 index languages (based on three general methods of assigning index terms) used to index 1400 research papers related to aerodynamics (*43*). Although Cleverdon deserves much credit for conducting these extensive, detailed, well-documented studies, both his methods and conclusions have been criticized (*44*). His finding that recall of relevant documents decreases significantly when controlled terms or simple concepts are used, rather than single (free, uncontrolled) natural language terms, conflicts with the opposite results of at least two more recent studies (*45, 46*).

Unfortunately, the comparative studies mentioned above, as well as some 20 other evaluations of indexing languages reviewed by Bourne in 1966 (*47*), have virtually ignored enumerative hierarchical classifications of the HICLASS type (with concepts already fully pre-coordinated *within* each category) which eliminate the need for either pre-coordination or post-coordination of categories. The classifications that were tested in the above studies are used with considerable coordination, frequently requiring both pre-coordination of category numbers during indexing and post-coordination of category numbers during retrieval operations. In view of this double disadvantage, it is not surprising that the classifications tested showed no definite advantage over keyword indexing methods.

In contrast, in a test at the Science Information Exchange (SIE), computerized use of an enumerative classification gave both more complete retrieval and higher levels of relevance than did a widely used, computerized, free text retrieval method (where every word in an abstract and citation is available as a search term). The SIE indexing procedure employs a type of HICLASS system. Since each category or "subject topic" is a complete concept in itself and is defined by its position in a hierarchy, in many cases coordination of categories is not necessary, and searching involves a substantial amount of single-hit matching. However, it must be emphasized that the SIE system (*34*) is a hybrid, which combines both enumerative classification and coordination in output operations (*48*).

Descriptions of 4600 research projects related to pesticides, reproduction, solid state physics, and physical oceanography were searched both by categories taken from the SIE classification and by free text searching. Examination of the answers to 39 questions asked by actual users of the SIE system showed that recall (the ratio of the number of relevant documents retrieved to the number of relevant documents in the data base searched) was 30 percent higher, and relevance (the ratio of the number relevant to the total number retrieved) was 15 to 20 percent higher with the SIE classification than with free text words (*46*). If this exciting finding can be corroborated and if enumerative classifications consistently perform substantially better than other widely used indexing methods, then the HICLASS type of classification can be used to upgrade the quality and performance of many information systems.

Concluding Comments and Summary

Selective dissemination of information to individuals provides a new and promising method for keeping abreast of current scientific information. Since SDI services are directed to the information needs of each individual, they are a significant step beyond group-oriented services and products, which require considerable expenditure of effort by each user as he sorts useful information from trash. However, SDI systems do require a high degree of precision in matching scientists against documents. They must operate more efficiently and economically than many current systems which occasionally provide a useful item of information to users. To meet these stringent requirements for quality, precision, efficiency, and economy, more research must be devoted to comparing and improving indexing methods, which are the basic component of all information storage and retrieval systems.

It is incredible that so much money

has been spent on the development and operation of scientific information systems before basic data on the comparative performance of various indexing methods have been gathered, analyzed, and confirmed by multiple investigators. The design of an effective information system would seem to require this type of basic knowledge, just as basic properties of alternative materials must be known before an engineer can design a building, bridge, or factory. Yet, except for the few studies mentioned in the previous section, research on indexing methods has been greatly neglected. Bourne's comment about studies of indexing languages is still an appropriate description of the situation: "In almost all the experimental reports, the investigator worked with an indexing language different than that of other experimenters. Consequently, no one has ever had his test results verified, or expanded, or made more precise by another experimenter" (*47*).

Most existing information systems are based on keyword indexing, with concepts broken into isolated terms during input operations and recombined to synthesize the original concept during search and retrieval. Such systems tend to involve imprecise indexing, with a high level of "noise" in retrieved documents, difficult search strategy involving extensive post-coordination, and lengthy, complex computer manipulations. This situation reflects the fact that many producers of indexed data originally focused the design of their systems on the production of a published product with entries printed under short, concise index headings. Production of magnetic tapes as a by-product of the publication process, and their use for retrospective searching or for SDI services, was a much later development, almost an afterthought. Yet use of these tapes is growing so rapidly that it may be time to redesign the tape-producing systems, with ease of tape use for SDI services and retrospective searching as the primary consideration, and with publication of abstract and index bulletins or title listings relegated to secondary importance (*49*).

The use of keywords to index documents creates a high degree of disorganization in information search and retrieval operations: Information is scattered under the many different terms that can be used to index different aspects of a concept. If the large-scale, comprehensive abstracting and indexing services were based on enumerative classifications with assignment of documents to logical hierarchical categories at the time of initial indexing, then many of the specialized information centers (*50*) and the 1300 abstracting and indexing services (*3*) would be unnecessary, and much of the reindexing and reprocessing of documents, the repackaging and reworking of abstracts and index data, and the resulting overlap and duplication characteristic of current information processing could be terminated.

Partly because of the disorganization resulting from keyword indexing, the cost of a 5-year retrospective search of information on just one data base on magnetic tapes is a major investment (*16*). The effort and cost required to find a few items of useful information scattered among 1,285,000 abstracts indexed on 116 full reels of magnetic tape (11 million characters per reel) which will be needed for the 5-year Eighth Collective Index to Chemical Abstracts (1967–1971) (*51*) staggers the imagination.

In contrast, when HICLASS systems based on enumerative hierarchical classifications are used, concepts that might be useful for later retrieval are identified and related items of information are grouped together during the indexing process. These enumerative classifications, with single-hit matching, make it possible to index and retrieve ideas as intact units and to perform simple sequential searches of the very small segment of a file that deals with a given topic (*31*). The experiments at both the Science Information Exchange and the National Cancer Institute, as described in this article, demonstrate that automated HICLASS systems are feasible and can operate at a very satisfactory level of performance.

Although considerable effort may be required for the development and constant updating of detailed enumerative classifications, HICLASS categories may facilitate organization of data at the time of input, improve the precision of matching documents with users, and greatly simplify search logic and computer manipulations. If so, then output savings and performance would more than justify input costs, and the development and use of enumerative classifications would be a better solution to information problems than the current keyword-and-coordination approach.

It is time to think beyond the ease of the single input step in information systems and to take a hard look at ways of easing retrieval problems for the multitude of information systems that process the indexed data (*52*). Indexing effort is expended only once, whereas search and retrieval effort is required by every user of a system. If information were better analyzed and organized during input operations, if more basic research were devoted to the effect of indexing methods on the performance of information systems, and if more emphasis were placed on the quality and usefulness of retrieved information, then the magnitude of problems related to the storage and retrieval of scientific information might be considerably reduced.

References and Notes

(Note that throughout this reference list six-digit numbers preceded by AD or PB identify documents available from the National Technical Information Service, Springfield, Va. 22151.)

1. R. R. Shaw, *Science* **137**, 409 (1962).
2. G. Schussel, *Data Manage.* **7**, 24 (1969).
3. Federation Internationale de Documentation (FID), *Abstracting Services*, vol. 1, *Science, Technology, Medicine, and Agriculture* (FID Publ. 455, The Hague, Netherlands, 1969). The 1300 abstract services listed in this publication include abstract sections in some primary journals but do not include many services that give only references and citations without abstracts. Volume 2 (FID Publ. 456) describes 200 abstract services in the social sciences and humanities. Contents of the volumes are described in *FID New Bull.* **19**, No. 12 (December 1969).
4. National Science Foundation, "Industrial R & D Spending, 1968" (NSF Publ. 70-12, Washington, D.C., May 1970).
5. D. R. Swanson, *Libr. Q.* **36**, 79 (1966).
6. F. W. Matthews, *Proc. Am. Soc. Inf. Sci.* **7**, 315 (1970).
7. H. P. Luhn is credited with the first description of automated SDI systems in two classic papers: "A Business Intelligence System" [*IBM J. Res. Dev.* **2**, 314 (1958)] and *Selective Dissemination of New Scientific Information with the Aid of Electronic Processing Equipment* (IBM Advanced Systems Development Division, Yorktown Heights, N.Y., 1959). *SDI Bibliography-1* (Publ. SRTP-1095, Share Research Corporation, Santa Barbara, Calif., February 1968) shows the growth of the field for the first 9 years in terms of number of references (in parentheses): 1958 (1); 1959 (1); 1960 (1); 1961 (10); 1962 (17); 1963 (19); 1964 (24); 1965 (41); 1966 (48); 1967 (46); and no year given (8). Total is 216 references.
8. Share Research Corporation, Santa Barbara, personal communication. Some 35 organizations thought to have SDI systems did not respond to the census questionnaire.
9. E. M. Housman, *Survey of Current Systems for Selective Dissemination of Information (SDI)* (Special Interest Group on SDI, American Society for Information Science, Washington, D.C., 1969) (AD-692-792); ——— and E. D. Kaskela, *IEEE Trans. Eng. Writ. Speech* **13**, 78 (1970); E. M. Housman, *Proc. Am. Soc. Inf. Sci.* **6**, 57 (1969).
10. W. A. Bivona and E. J. Goldblum, "Selective Dissemination of Information: Review of Selected Systems and a Design for Army Technical Libraries" (Information Dynamics Corp., Reading, Mass., 1966) (AD-636-916); M. Cooper, *J. Chem. Doc.* **8**, 207 (1968); A. G. Hoshovsky, "Selective Dissemination of Information (SDI): Analysis of Experimental and Operational SDI Services, 1967" (Doc. OAR 69-0016, Office of Aerospace Research, U.S. Air Force, Washington, D.C., 1969) (AD-691-012).
11. J. A. Holt, thesis, University of Chicago, Chicago, Ill. (1967); J. H. Connor, *Libr. Q.* **37**, 373 (1967). The *Annual Review of Information Science and Technology* (C. A. Cuadra, Ed., Encyclopaedia Britannica, Chi-

cago, Ill.) contains reviews of SDI literature each year; see P. L. Brown and S. O. Jones, in *ibid.* (1968), vol. 3, p. 276; H. B. Landau, in *ibid.* (1969), vol. 4, p. 242.

12. V. A. Wente and G. A. Young, in *Annual Review of Information Science and Technology*, C. A. Cuadra, Ed. (Encyclopaedia Britannica, Chicago, Ill., 1970), vol. 5, p. 259.
13. The ANSA (Automatic New Structure Alert) system, which uses Wiswesser line notations to identify compounds of interest to individual chemists, is in the final stages of testing at the Institute for Scientific Information, 325 Chestnut Street, Philadelphia, Pa. An SDI service for patents is supplied by Information Retrieval Limited, 38 Chancery Lane, London, W.C.2, England.
14. K. D. Carroll, *Survey of Scientific-Technical Tape Services* (AIP ID 70-3, American Institute of Physics, New York, and ASIS SIG/SDI-2, American Society for Information Science, Washington, D.C., September 1970).
15. M. C. Gechman, *Proc. Am. Soc. Inf. Sci.* **7**, 97 (1970).
16. Retrospective searches cost from $50 to $150 (depending on the size and frequency of the tape data base) for each year searched at the University of Georgia Computer Center, Athens, as of mid-1970. A 5-year search would cost $250 to $750. The same center charges $84 to $182 for running one SDI profile against one magnetic tape data base for a 1-year period. Charges to academic institutions are slightly lower than the standard commercial rates given above.
17. Full members in ASIDIC must be computer-based scientific dissemination centers which use at least two or more tape data bases from different suppliers either to furnish SDI services for more than 100 profiles or to make 1000 retrospective searches per year, on demand. Interested individuals or organizations can become associate members. The present secretary is Diane Follmer, 201-25, 3M Center, St. Paul, Minn. 55101.
18. Operators of computer-based chemical information services in Western Europe (using the acronym CHEOPS) met informally in the late 1960's under the sponsorship of the Organization for Economic Cooperation and Development. At a meeting in The Hague in April 1970, they broadened their scope beyond chemical information and began formal meetings of a group called the European Association of Scientific Information Dissemination Centers (EUSIDIC). The eight founding members represent three centers in England, two in Sweden, and one each in Denmark, France, and Holland. The present secretary of the group is A. K. Kent, UKCIS, University of Nottingham, England, NG7-2RD.
19. G. Schlachter, "Notes on SDI Conference at University of Wisconsin on April 16–18, 1969." Copies of this mimeographed report may be obtained from H. R. Knitter, Director, Data Processing Center, Graduate School of Business, University of Wisconsin, Madison 53706.
20. When the same article is processed and indexed by several tape producers, royalty payments for retrieval or use of information about that article could be a complex problem.
21. Examples of SDI-like announcement services to small groups of specialists are: the Selective Current Aerospace Notices (SCAN) system of NASA with about 200 topics [V. A. Wente and G. A. Young, *J. Chem. Doc.* **7**, 142 (1967); *Proc. Am. Soc. Inf. Sci.* **5**, 217 (1968)]; the CARD-A-LERT service of Engineering Index began on a manual basis in 1928 and is now automated with 167 divisions (brochure available from Engineering Index); the Selective Dissemination of Microfiche (SDM) system of the National Technical Information Service with about 300 subject areas (description available from National Technical Information Service, Springfield, Va. 22151); and the INSPEC Topics service of the Institution of Electrical Engineers in England which began with 21 major topics. See also Wente and Young (*12*, pp. 282–284) for a general discussion of this "shared-profile" type of SDI system.
22. E. Garfield and I. H. Sher, *J. Chem. Doc.* **7**, 147 (1967); E. Garfield and I. H. Sher, *Am. Behav. Sci.* **10**, 29 (1967); M. V. Malin, *Libr. Trends* **16**, 374 (1968); M. Weinstock, in *Encyclopedia of Library and Information Science*, A. Kent and H. Lancour, Eds. (Dekker, New York, 1971), vol. 5, pp. 16–40.
23. J. H. Schneider, *Proc. Am. Doc. Inst.* **4**, 301 (1967); *Proc. Am. Soc. Inf. Sci.* **5**, 243 (1968); in *Proceedings of The Fifth Annual National Colloquium on Information Retrieval*, J. W. Ramey, Ed. (Information Interscience, Philadelphia, 1968), pp. 137–143.
24. J. H. Schneider, *Biochemistry: A Decimal Classification* (American Univ. of Beirut, Lebanon, 1960 and 1961), vols. 1 and 2 (PB-170-793 and PB-170-794); "Proposed Categories for Classifying the Subject Content of Research Grants on a Public Health Service-Wide Basis" (National Institutes of Health, Bethesda, Md., 1964) (PB-170-736); "Hierarchical Decimal Classification of Information Related to Cancer Research" (National Institutes of Health, Bethesda, Md., 1968) (PB-177-209).
25. Only three of the 104 participants dropped from the project during the 1-year trial period.
26. Major disciplines of the 104 research projects were biochemistry (35); immunology (20); cytology and pathology (19); physiology (17); cytogenetics (7); and endocrinology (6).
27. Permission was easily obtained to copy and distribute title page and abstract when it was pointed out that this would stimulate interest in and use of their journal.
28. D. J. de Solla Price, *Science Since Babylon* (Yale Univ. Press, New Haven, 1961), p. 99; *Little Science, Big Science* (Columbia Univ. Press, New York, 1963), chap. 3.
29. F. W. Lancaster, *Evaluation of the Medlars Demand Search Service* (National Library of Medicine, Bethesda, Md., 1968), p. 126.
30. J. H. Schneider, *Organic Chemistry, the Chemical Industry, and Physics: A Decimal Classification* (American Univ. of Beirut, Lebanon, 1960) (PB-170-795).
31. J. H. Schneider, in *Proceedings of the Sixth Annual National Colloquium on Information Retrieval*, L. Schultz, Ed. (College of Physicians, Philadelphia, Pa., 1969), pp. 99–105.
32. F. J. C. Roe, R. L. Carter, M. A. Walters, J. S. Harington, *Int. J. Cancer* **2**, 628 (1967).
33. C. H. Davis and P. Hiatt, *J. Am. Soc. Inf. Sci.* **21**, 29 (1970); K. J. Bierman, *Proc. Am. Soc. Inf. Sci.* **7**, 87 (1970); H. Falk and H. E. Tompkins, *ibid.*, p. 283; R. R. Freeman and P. Atherton, "Final Report of the Research Project for the Evaluation of the UDC as the Indexing Language for a Mechanized Reference Retrieval System" (Rep. AIP/UDC-9, American Institute of Physics, New York, May 1968).
34. D. F. Hersey, W. R. Foster, M. Snyderman, F. J. Kreysa, *Methods Inf. Med.* **7**, 172 (1968).
35. I am compiling an annotated bibliography of references on classification-based, automated information systems. A preliminary version is available on request. The bibliography does not include automation of classical, routine library functions which often involve a standard library classification.
36. J. J. Magnino, "CIS—A Computerized Normal Text Current Awareness Technique" (Rep. ITIRC-006, IBM Technical Information Retrieval Services, Armonk, N.Y., 1965); "IBM's Unique but Operational International Industrial Textual Documentation System—ITIRC" (Rep. ITIRC-023, IBM Technical Information Retrieval Services, Armonk, N.Y., 1967).
37. G. Salton, *Science* **168**, 335 (1970); *Am. Doc.* **20**, 61 (1969).
38. An AUTOKLS system, which I have written in PL/1 for the IBM 360/65, can be used to replace, delete, move, and insert categories, descriptive text, notes, "see also" references, and alphabetic index entries for the type of classfications listed in (*24*). Changing a category triggers changes in the cross-references and index entries related to the changed category. Successful automation of the Universal Decimal Classification (UDC) is described elsewhere: M. Rigby, *Mechanization of the UDC* (Meteorological and Geoastrophysical Abstracts, American Meteorological Society, Washington, D.C., 1964) (PB-166-412); R. R. Freeman, *Modern Approaches to the Management of a Classification* (Rep. AIP/UDC-3, American Institute of Physics, New York, 1966) (PB-173-704).
39. Science Information Exchange, Washington, D.C., personal communication.
40. T. Saracevic, *An Inquiry into Testing of Information Retrieval Systems*, part 2, *Analysis of Results* (Center for Documentation and Communication Research, Case Western Reserve Univ., Cleveland, Ohio, 1968), p. 119 (PB-180-951).
41. G. Salton, *Information Storage and Retrieval* (Sci. Rep. ISR-12 to the National Science Foundation, Department of Computer Science, Cornell University, Ithaca, N. Y., 1967), p. I-5 (PB-176-536).
42. C. W. Cleverdon, *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems* (College of Aeronautics, Cranfield, Bedford, England, 1962); ———, F. W. Lancaster, J. Mills, *Spec. Libr.* **55**, 86 (1964); C. W. Cleverdon and J. Mills, *ASLIB Proc.* **15**, 106 (1963); F. W. Lancaster and J. Mills, *Am. Doc.* **15**, 4 (1964).
43. C. W. Cleverdon, J. Mills, M. Keen, *Factors Determining the Performance of Indexing Systems* (ASLIB-Cranfield, Bedford, England, 1966); C. W. Cleverdon, *ASLIB Proc.* **19**, 173 (1967).
44. D. R. Swanson, *Libr. Q.* **35**, 1 (1965); P. Richmond, *Am. Doc.* **14**, 307 (1963); *College Res. Libr.* **27**, 23 (1966); M. Taube, *Am. Doc.* **16**, 69 (1965).
45. C. W. Hargrave and E. Wall, *Proc. Am. Soc. Inf. Sci.* **7**, 291 (1970).
46. D. F. Hersey, W. R. Foster, E. W. Stalder, W. T. Carlson, *ibid.*, p. 265; *J. Doc.*, in press.
47. C. P. Bourne, in *Annual Review of Information Science and Technology*, C. A. Cuadra, Ed. (Wiley, New York, 1966), vol. 1, chap. 7. Space does not permit a discussion of the artificiality of many of the experimental protocols, the desirability of comparing operating rather than test systems, or the frequent use of "subject specialists" or "testers" rather than actual users to evaluate the performance of indexing procedures.
48. Category numbers from an enumerative classification can be coordinated by using either Boolean logic or by weighting techniques just as keywords or categories from faceted classifications are coordinated. However, if the classification is sufficiently detailed, this coordination is usually unnecessary. It would be useful to have a brief list of common experimental animals, standard experimental methods or approaches, major geographic locations, and a few broad age ranges that could be checked off during indexing, stored with the indexed data (as a short bit string on magnetic tape, for example), and screened on output to eliminate certain documents dealing with animals, methods, locations, or ages that are not of interest to a particular user.
49. Income from subscriptions to published abstract and index bulletins is a major factor preventing a significant shift away from printed abstract and index bulletins. However, Wente and Young (*12*, p. 267) note an increasing readiness to neglect printed documents and apply resources to SDI and other computerized services. They report that the journal *Road Abstracts* has been discontinued and replaced by an SDI and retrospective search service. W. S. Brown, J. R. Pierce, J. F. Traub [*Science* **158**, 1153 (1967)] and others have recommended that journals stop binding papers into issues and, instead, distribute papers, titles, and abstracts to individuals by a type of SDI system.
50. The *Directory of Federally Supported Information Analysis Centers* [Federal Council for Science and Technology, Committee on Scientific and Technical Information, Washington, D.C., 1968) (PB-189-300)] describes 113 centers. E. Mount [*Spec. Libr.* **59**, 107 (1968)] estimates that there are 200 information analysis centers.
51. Chemical Abstracts Service, *Report on the Fourteenth Chemical Abstracts Service Open Forum* (Columbus, Ohio, March 1971), p. 20.
52. Unfortunately, the producers and the users of indexed data bases are usually two different groups of people. The producer's main concern is ease and economy of input, rather than the effort and problems involved in retrieving useful data from his products. Still, the investment in an average $45,000 research project is so great that some source should be found to pay a subject specialist the few dollars required for thorough, in-depth indexing of every article resulting from that project.