By J. W. PERRY, ALLEN KENT AND M. M. BERRY

Mechanized Literature Searching—A Progress Report

The authors are members of the staff of Battelle Memorial Institute, Columbus, Ohio.

RECENT YEARS have witnessed an increasing amount of time and effort being devoted to the problems of scientific and technical documentation. Unfortunately, discussion of newly developed methods and equipment sometimes obscures the purposes to be served. Unrealistic evaluation of new documentation techniques may do more harm than good.

The material basis of our civilization is provided by science and technology. Efficient use of scientific and technical information is essential to a very wide range of activities, such as developing new products and designing new machines, planning and conducting research, and evaluating the results of tests and experiments. In planning production, either in farm or factory, in exploring for new mineral resources, in conducting trade and commerce, the effective use of scientific and technical information may well provide the margin between success and failure.

It should not be concluded, however, that scientific or technical information is so high in value, that rendering it readily available justifies unlimited cost. At best, the value of a given piece of information cannot exceed the cost of regenerating it through experiment, test or otherwise. The philosophy that should underlie the design of a documentation system in science and technology is the same as for designing a solvent-recovery system. The value of the recovered material—be it information or acetone—must, in no case, be less than the cost of recovery. Difficulties involved in estimating the value of retrieved information often render quantitative application of this philosophy difficult, but this in no way undermines the latter's validity.

The philosophy of cost justification, even if applied only in a qualitative way, provides considerable valuable guidance. We are reminded, first of all, that there are only disadvantages in retaining information for which absence of future use can certainly be predicted. Furthermore, this principle strongly suggests that time and effort invested in processing different types of information for future retrieval should be proportional to the benefits to be achieved. Such benefits may be measured as the frequency of use times the value provided by each use. In estimating value, thought should be given not only to the cost of eventually regenerating information by experiment or test, but also to the fact that the record of an experiment is far less informative than the experiment itself. Repeating an experiment can be expected to provide considerably more information than ıeading someone else's description of what he did and saw—or thought he did and saw, as the case may be.

A particularly important factor in determining the value of recorded information is the level of ability and competence of the user. A first-rate scientist may be stimulated by a report of a previous failure to try a slightly different approach or to repeat the previous experiment with just those precautions needed to achieve success. Mediocre talent, under the same circumstances, may be discouraged from any further research. Skill in evaluating and using the results of earlier work is an essential factor in determining the value of documentation. Such skill, it should be noted, is closely akin to the ability to observe and to evaluate the results of experiments in order to arrive at sound conclusions. Mediocre talent is at a disadvantage when confronted by new results regardless of whether they originate in laboratory or library.

Enough has been said, perhaps, to enable us to say that the goal of documentation in science and technology is to reduce the cost of research, development, and related activities by making scientific and technical information available as needed. This definition of purpose presupposes nothing concerning methods to be used. The choice of methods will be determined, of course, by the parameters of the situation to be met and the needs to be served.

Consider, as an almost trivial, though instructive, example, a small collection of a dozen or so working papers of moderate length. Classifying or indexing such a small collection can provide little or no saving of time or effort when using them. No organization is, for this simple situation, almost sure to be the best organization.

The situation changes, of course, with increasing number of papers. Here, it may be helpful to distinguish between two cases. In one of these, let us assume that the various items may be arranged in some one order in accordance with some one feature, such as the serial number on the patents of a given country or the senior author name on reprints. Setting up a monodimensional array on such a basis can produce satisfaction to the user, if other aspects, especially those relating to subject matter, will never need to be available as leads to documents of interest. Even when subject matter is of importance, it may be advisable to maintain an author index or a numerical index for patents if frequency of use and resulting saving in time provide justification for the expense involved.

In the more general case, of course, considerable advantages are achieved by providing a subject approach to the contents of scientific and technical documents. Before considering in detail some of the methods, techniques, and devices that have been developed or proposed for use in this connection, it is well to note that certain steps are involved regardless of what methods, techniques, or devices may be used. These basic steps might be summarized as follows:

(1) *Analysis of subject matter.*

During this step, decisions must be made as to which aspects of subject matter are likely to be important in retrieving the document or correlating its information with that of other documents. Failure to detect an important aspect may result in the document being overlooked when its contents should be considered. Over-evaluation of a document may result in its being selected as pertinent when it is of so little interest as to be merely troublesome. For these reasons, the expense incurred in having an expert conduct this analytical step usually can be justified as a good investment.

(2) *Processing important aspects of subject matter.*

The goal of such processing is to establish means for facilitating the identification of pertinent documents to meet the users' requirements. As is well known, such processing may assume a variety of forms.

Words and phrases may be selected to constitute index entries which are then alphabetized to provide an array and thus facilitate a person's finding a given entry, as a lead to a desired paper.

An alternate possibility, that has been extensively discussed during the past year, is to establish a separate card for each aspect and enter on each aspect card the numbers of the document to which the given aspects pertain. A book devoted to this system, sometimes referred to by the name "Uniterm," appeared a few months ago. Earlier embodiments of this approach were the Batten-Cordonnier and National Bureau of Standards (Office of Basic Instrumentation) systems, in which holes were punched on aspect cards to permit easier manipulation of the reference cards. (See R. S. Casey and J. W. Perry, eds., *Punched Cards—Their Application to Science and Industry*. New York, Reinhold Publishing Company, 1951. Chapter by Dr. Batten.)

A third alternative is to relate the various aspects selected as being important in an array of headings, such as the Universal Decimal Classification System, or the Manual of Classification of the U. S. Patent Office. One or several headings of more or less generic scope may be used to draw together papers having certain aspects in common.

A fourth procedure is to use the various aspects as a basis for making entries on a medium to be searched by mechanical or electronic means. For examples, holes may be punched in cards, or magnetic spots may be recorded on tape, or transparent spots recorded on photographic film. Various systems, usually referred to as codes, are used for attributing meaning to such holes or spots in a well-defined orderly manner.

An alternate possibility, to which we have been and still are devoting considerable time and attention, is to organize the words and phrases that correspond to important aspects so that a telegraphic-style abstract is generated. Encoding its component words and phrases renders it "readable" by recently designed scanning equipment. Various problems involved and advantages that can be achieved will be discussed subsequently.

(3) Utilizing processed aspects to identify pertinent documents.

The importance of this step seems to be underestimated with surprising frequency. It is obvious, of course, that indexes, classification systems, and other forms of documentation cannot be regarded as successful unless they enable persons needing information to identify documents of pertinent interest at reasonable cost of time and energy.

(4) Acquiring needed information.

Identifying those documents most likely to contain information of value to the user must be followed by his actually contacting and absorbing the needed information. At this stage, effective abstracts can provide valuable assistance in aiding the user either to reject papers of marginal interest or to obtain a quick bird's-eye view of papers of sufficient interest to warrant detailed study.

Problems involved in supplying copies of needed documents or difficulties caused by language barriers will not be discussed in this paper. However, new approaches to some of the problems involved in supplying copies of needed documents are being developed in France (Film-O-Rex) and in the United States (Eastman Kodak's Minicard), etc.

It is not the purpose of this paper to attempt to define with precision the circumstances under which various methods, techniques, and equipment will prove to be most advantageous. The determining parameters, as is perhaps evident from preceding discussion, are as follows:

(1) Size of file, i.e., number of reports. reprints, and other papers.

(2) Complexity of subject matter involved, range of aspects of importance, number of aspects involved on an average per document, important relationships between aspects.

(3) Frequency of use of file.

(4) Type of search required; in particular, range and number of aspects involved in defining an average search.

(5) Need for speed in responding to search requirements.

In terms of these parameters, each of the methods previously outlined has its range of usefulness. The machine searching system now to be discussed has been designed to meet the most severe conditions with respect to size of file, complexity of subject matter, frequency of use, complexity of search, need for speed and, last but not least, minimum cost.

To provide maximum flexibility, the new indexing-abstracting methods have been designed so that any important aspect or feature of the subject matter of documents or any combination of such aspects may be used to define the scope of a search to be conducted by machine. Searching methods have been devised so that the machine scans encoded abstracts of special design. Aspects of subject matter mentioned in these abstracts may include spatio-temporal entities (things or substances, organisms, persons, devices, apparatus, or machines), processes and actions, conditioning circumstances, geographical regions, locations and dates, concepts of generic or theoretical nature, attributes including those involving numerical values, such as melting points. The abstracts are so prepared as to render explicit relationships between these different aspects of subject matter.

The analysis of subject matter already previously discussed is the first step toward preparing these abstracts. This analysis is best accomplished by a subject-matter expert who decides what the important aspects of subject matter are and then designates these aspects by appropriately selected words and phrases. Underlining or similar marking may be used to direct attention to the selected words and phrases or they may be provided as marginal notations. No restrictions are imposed on the words or phrases that may be so selected.

The next step is to organize the words and phrases into telegraphic style abstracts and finally to encode the abstracts to render explicit important points of meaning which then serve as reference points in defining and conducting the scope of a search.

Before undertaking to outline the principles involved in preparing the abstracts and in establishing codes for words and phrases, it should be pointed out that the development of these methods proceeded hand in hand with the development of the new scanning equipment for searching encoded abstracts and selecting those of pertinent interest.

Early in this development, it was recognized that the scanning machine should have the ability to perform selections defined by requiring the simultaneous presence of all of several search criteria. Such a requirement which is said to involve the logical product of the criteria may be represented by $A \cdot B \cdot C \cdot D \cdot E$ (for five criteria). Another possibility is a search in which any one of several criteria suffices for a positive response. This requirement, sometimes called the logical sum, may be represented by $A+B+C+D+E$ (again for five criteria). When it is required that some criterion be absent, the logical difference is involved as symbolically expressed by $A-B$ (in this case, the search requirement is for A to be present and B to be absent).

It was realized early in this development of new methods and equipment that the use of selected words to designate the important aspects of the subject matter of documents has proved highly effective in constructing subject indexes to be scanned by human experts. Their ability to interpret the meaning of technical terminology, and other words, is also essential to successful use of conventional classification systems. The use of phrasing to indicate relationships between substances, processes, attributes, and concepts in general plays an important role in abstracting, indexing, and classifying along traditional lines. The problem is to provide means for rendering explicit for machine searching those implied aspects of the meaning of words and phrases which can be understood easily and quickly by the human expert but which cannot be interpreted by electronic machines of practical design at the present level of electronic technology.

The most convenient approach to this problem is to consider the problem presented by the meaning of words. We observe, first of all, that certain technical terms are so derived etymologically that their Greek and Latin roots render certain basic aspects of meaning explicit in the words' spelling. Examples of such words are "thermometer" (heat, measure), and "cytolysis" (cell, dissolution). If all the words of our language were derived in the same way, basic aspects of meaning, denoted by the Greek roots, would be rendered explicit by certain letter combinations in the spelling, such as "therm" for heat, "meter" or "metric" for measure, "lysis" for dissolution, etc. The limited degree to which the spelling of technical terms explicitly exhibits, in a logical way, the basic elements of their meaning has made it advisable to generate codes to render such meaning explicit and thus make it possible for machines to detect certain combinations of symbols used to designate elements of meaning. In the code we have developed, three-letter combinations are usually so used. Thus, MAC is used to designate "machine, apparatus, equipment, device," MES "measure," NAL "analysis," REH "heat," REL "light," etc. The three-letter combinations can be set up so as to be mnemonic in character and otherwise convenient in use.

A typical group of words selected from our code dictionary are provided for illustrative purposes.

abaca	TEX FIB
abortion	BIL DED GEG
abortifacient	BIL DED GEG DOG
abrade	BAR
abrasion	BAR
abrasive	BAR
absorb	BAS
absorbent cotton	BAS TEX
absorber	BAS MAC
absorption band	BAS RAL CAP
absorption tower	BAS MAC
acaricide	DED PES

In developing such codes for scientific terminology, our efforts have been directed, as already noted, to rendering explicit in the codes, those aspects of the meaning of terms most useful as reference points for defining searching operations to be performed by automatic equipment.

It is perhaps obvious that two problems have been encountered.

One of these is to arrive at a set of basic aspects of meaning as exemplified by MAC for "machine, apparatus, equipment, device" or MES for "measure." Such aspects of meaning—or semantic factors, as they are sometimes called—provide a framework of generic reference points for searching operations.

The second, no less important problem, has been to apply the semantic factors in a consistent fashion when setting up the codes for individual terms used in science and technology. At the present time, we are carrying through a revision of the original tentative codes for 7000 terms frequently used in traditional indexing and classifying systems.

It is not possible, within the limits of this paper, to set forth in detail the various considerations that have come to our attention in developing codes based on semantic factors. It may be illuminating, however, to point out the nature of the logical relationships that exist between specific terms and the semantic factors from which the codes of the terms are constructed. These relationships are of four types: (1) Functional; (2) Attributive; (3) Whole-part; (4) Class inclusion.

Functional attributes are illustrated by those semantic factors which designate operations such as MES for "measurement" or NAL for "analysis." The following sample codes illustrate the use of semantic factors indicating functional aspects, which are pointed out by underlining.

Examples of Factors Indicating Functions

Grindstone	BAR; MAC (abrade; machine)
Saber	CUT; WEP (cut; weapon)
Insecticide	DED; PES (kill; pest)
Navigator	GUD; PEP (guide; person)
Filter	SEP; MAC (separate; machine)

Attributive semantic factors indicate a relationship of a descriptive or adjectival character. This type of relationship may be illustrated by the following examples, in which codes for the attributive factors have been underlined.

Examples of Factors Indicating Attributive Relationship

Forestry	WOD; PAN; SIC (wood; plant; science)
Hour	TIM; LEM (time; unit)
Meteorology	WET; SIC (weather; science)
Forceps	SUG; MAC (surgery; machine)
Bombsight	MIL; POT; MAC (military; optics; machine)

The role of the whole-part relationship is particularly well illustrated by codes for geographic units, for which departure from the use of three-letter combinations permits more compact codes to be established.

Geographical Codes Illustrating Whole-Part Relationships

United States	US
New England	USNE
Massachusetts	USNEMA
Vermont	USNEVT
New Hampshire	USNENH
Connecticut	USNECT
Maine	USNEME
Rhode Island	USNERI

In coding scientific and technical terminology the whole-part relationship comes into use relatively infrequently. The reason for this is the fact that this type of relationship is more frequently an important aspect of meaning at the indexing-abstracting level rather than a semantic aspect of the meaning of individual terms.

Examples of Factors Indicating Whole-Part Relationships

Roof	SUT; COV (structure; cover)
Tau-saghyz	PAN; RUB (plant; rubber)
Fleet	MIL; SAS; SIP (military; organization; ship)

From the philosophical point of view any set of individuals having one or several functions or attributes in common constitute a class. Similarly, the various parts of a given whole also form a class. Hence, from this point of view, class inclusion—our fourth logical relationship between specific terms and semantic factors—might be regarded as all inclusive with respect to the other three. The semantic factors of the type exemplified by MAC "machine, apparatus, equipment, device" or TEX "textile" may be said to be based on generic terms denoting classes. Such class designation does not involve direct explicit reference to the attributes involved, for example, in textiles or machines and similar devices. From this point of view, we may find logical justification for speaking of semantic factors based on class inclusion. Certainly, this concept proves helpful in establishing codes for terms.

Examples of Factors Indicating Class Inclusion

Cytology	CEL; SIC (cell; science)
Curie	RAX; LEM (radiation; element)
Beer's Law	RUL; BAS; RAL (law; absorption; light)
Chaplain	REL; PEP (religion; person)
Gasoline	PET; FUL (petroleum; fuel)

In discussing the establishment of codes for terms, rendering their meaning explicit is the problem with which we have been concerned. Designating a thermometer as a device for measuring temperature (MAC; MES; TEM) does not, however, indicate in what respect a thermometer may be involved in a given situation reported in some document. Thermometers might, for example, be referred to as being manufactured or the document may be concerned with research on thermometers. The relationships between the various things, processes, and circumstances are also important when developing a system aimed at rendering explicit as many important aspects of subject matter of a document as may be advantageous for searching and correlating.

In analyzing these relationships, guidance is provided by the same logical relationships as underlie the development of codes employing semantic factors. When indexing a document, we may note, for example, that a certain chemical compound is used as a component of a mixture (whole-part relationship) or that a given compound or mixture may be used to treat a disease (functional relationship) or that a given medicinal preparation has a certain physical consistency (attributive relationship). The class inclusion relationship will be based, when conducting indexing, on some whole-part, functional or attributive relationship, which is much more explicit in character than is the case in the establishment of terminology.

Linda Hall Library

(Continued from page 141)

Rare Book Room is located at the southeast corner of the first floor, adjacent to the librarian's office. A small elevator and two book-lifts to all levels are located on the lower stack levels. Book capacity of the new building is estimated at about one-half million volumes, and the total cost at less than one and one-half million dollars.

There is every indication to believe that the Linda Hall Library, with a large and rapidly growing collection in science and technology, meets a real need in a region which has shown a remarkable rate of expansion in industrial and technical facilities since the end of the war. With a new building, providing space for at least twenty years, and an endowment which should prove adequate for the steady growth of the collection, Linda Hall should become an increasingly valuable research asset.