ERRATA:

Several errors have crept into formulas in two papers in the series entitled "Machine Literature Searching." In paper VI, *American Documentation*, Vol. 5, No. 4, Page 242, right hand column, the third formula should read:

$$(A.B.C)\ (<C.B.F>) - H$$

In paper VIII, *American Documentation* Vol. 6, No. 2, Page 101, right column the third equation should read:

$$\left[1 - (x/n)_0\right]^K + K(x/n)_0\left[1-(x/n)_0\right]^{K-1} = 1-q$$

and the fourth equation should read:

$$1 - K(x/n)_0 + \frac{K(K-1)}{2}(x/n)_0^2 - - -$$
$$+ K(x/n)_0[1 - (K - 1)\ (x/n)_0$$
$$+ \frac{(K-1)\ (K-2)}{2}(x/n)_0^2 + - - -] = 1 - q$$

In addition to correcting misprints, the last equation has been rewritten to show more clearly the expansion of the second binomial expression.

The authors wish to express regret for any inconvenience these errors may have caused.

MACHINE LITERATURE SEARCHING X. MACHINE LANGUAGE; FACTORS UNDERLYING ITS DESIGN AND DEVELOPMENT

JAMES W. PERRY, ALLEN KENT, and MADELINE M. BERRY

Introduction

A documentation system that employs machine literature searching methods must link together a number of different operations and functions. The more obvious of these are:

1. The analysis of information requirements to express them in terms of operations to be performed by automatic equipment;
2. The analysis and encoding of the subject content of graphic records to render them amenable to machine searching; and
3. The routine operations performed by the machines in accomplishing the desired identification and selection of documents of probable pertinent interest.

This linking together of diverse operations and functions is most conveniently accomplished by a system of symbolism whose design and development is influenced and controlled by a variety of factors to be reviewed in this paper. This system of symbolism has come to be called "machine language." (1) Perhaps the origin of this term could be traced to certain points of similarity and contact with human language. Although misinterpretation of such similarity or overestimation of its importance may have been occasioned in the past by use of the term "machine language," it will be used in this paper as a matter of convenience. At the same time, we shall attempt to place the relationship of machine language to human language in proper perspective.

To emphasize the importance of our subject, we recall to mind that extensive experience has taught the need for care and precision in the use of terminology when constructing conventional subject indexes and classification systems based on hierarchical arrangement of classes and subclasses. (2,3) Awareness of the problems encountered in constructing and using conventional aids to searching may lead to an excessively pessimistic evaluation of the problems that are encountered, if we attempt to depart from tradition and introduce new methods. Since, it is argued, so much human skill is required to generate and to use alphabetized indexes, it will be very difficult to conduct the analysis of the subject content of documents in such a way that brainless machines, by merely performing routine operations, can accomplish useful selection of documents of pertinent

interest to a given problem, question, or situation. (4) Similarly, awareness of problems encountered in establishing and using conventional patent classification systems that are based on pigeon-holing may lead to expressions of doubt as to the applicability of machine searching methods for patents in general, even though applicability in certain specialized fields, such as chemical compositions of matter, may be recognized as offering practical advantages. (5) It is hoped that the present paper may contribute to dispelling these doubts and uncertainties.

We shall undertake to indicate how machine language may be developed so that it links human thinking concerning a given problem or situation with machine operations that can serve to identify and select documents and graphic records that are of probable pertinent interest to the problem or situation in question. We are encouraged in this undertaking by a recent paper that emphasized the emergence of parallel trends of thought in the study of methods of alphabetized indexing, hierarchical classification, colon classification, and subject content analysis for machine searching. (6)

The Nature of Mechanized Searching

The goal of mechanized searching is to achieve hitherto unattainable efficiency in the practical utilization of recorded knowledge in spite of, and also because of, its expanding volume and increasing complexity. This purpose is closely akin to that pursued in the advancement of classification and we may adopt a common statement of purpose by agreeing that, for a given field of interest, "we are attempting to systematize the knowledge of our chosen field in such a way that any individual original item of information can be instantly available, and that any conceivable permutation of individual items of information can be sought for." (7) This statement of purpose may serve as a basis for discerning more clearly the importance of the development of machine language and the fact that such development "bristles with difficulties." To grapple successfully with such difficulties, it is necessary to explore their origin and their nature. Before undertaking this task, it is well to recall to mind certain facts and considerations which, though of pertinent interest to the development of machine language, do not provide needed guidance for carrying through its development.

As has been pointed out in an earlier paper in this series, the identifying and selecting operations that can be performed by properly designed searching equipment involve the matching-up of the characteristic aspects of an information requirement with the corresponding characteristic aspects of the subject content of graphic records. (8) This matching-up may constitute a complex of operations whose relationships must be clearly understood and defined in order that the programming of the searching machine, e.g., by wiring plugboards, may be conducted so as to achieve desired results. The theory of class definition, it was pointed out, provides a convenient means for expressing configurations of characteristics so that they may serve as the basis for programming machines to achieve matching and identifying operations. (9) In this way the theory of class definition may be advantageously applied in assuring successful machine operations. It is important not to overestimate the extent to which the theory of class definition controls machine searching in general and machine language in particular. It is especially important not to overlook the fact that the theory of class definition imposes no limitations on the characteristics that may be selected to designate important aspects of the subject content of graphic records. The only requirement is that the designating characteristics, regardless of their choice or nature, shall be stated in explicit and unambiguous fashion. This requirement is, of itself, insufficient to serve as a complete guide or framework for the development of machine language.

Nor is the needed guidance to be sought in the criteria proposed in an earlier paper for evaluating the operational effectiveness of a machine searching system or of documentation systems in general. (10) Such criteria are to be regarded as the analogs of horsepower, thermal efficiency, boiler water consumption, or similar factors that may be used to rate the performance of a steam engine. Such factors, though undeniably important, do not suffice to provide needed guidance in designing steam engines or machine searching systems. In developing machine language, it is not enough to consider methods for measuring and comparing the relative efficiencies of documentation systems.

There is a fundamental reason why development of machine language cannot be based solely on an analysis of machine operations. This reason is to be sought in the fact that machine language, to be effective, must provide a connecting

link between machine operations and human thinking.

Accordingly, it is necessary that the development of machine language take into consideration various fundamental aspects of the generation and use of human knowledge. In directing attention to such factors, we have limited ourselves to the field of science and technology. Otherwise, various special considerations relating to other fields of learning and professional activity might render our presentation excessively complicated and tortuous. This does not mean, however, that our observations and conclusions are necessarily invalid outside the field of science and technology. Important characteristics of human knowledge can be expected to be independent of the subject-matter field. For this reason, at least, we might anticipate that observations and conclusions made in science and technology may prove valid in other fields such as law or medicine. Additional investigations may be required to establish and demonstrate the extent of such validity in other fields.

Human Limitations and Knowledge

Human limitations constitute the most important factor that determines the character of human knowledge. (11)

Our sensory perceptions inform us, at best, imperfectly concerning the phenomena of nature. For example, electromagnetic radiation, outside a narrow band, cannot be detected at all by direct sensory perception. Our unaided senses leave us blissfully unaware of the babel of messages being transmitted by radio and television. It is easily possible for a person to be subjected to a harmful amount of gamma radiation without noting such exposure.

The invention of various instruments extends the range of perception, but without any assurance that important phenomena are not eluding detection. Fundamentally important phenomena in electromagnetism were discovered less than a century ago. The discovery and detection of nuclear phenomena is still a field of active research. In biology, the mystery of life phenomena has scarcely been penetrated at all.

In spite of these limitations on our powers of observation, we are confronted by a constantly shifting stream of phenomena with which we must cope if we are to survive. At the most primitive level, it is necessary to recognize food objects, though a given specific object may never have been seen before. Similarly, wild beasts and rampaging automobiles must be recognized, and suitable evasive tactics adopted. Without a certain rudimentary ability to recognize that an object never before seen partakes of the character of other similar objects to which attention has been previously directed, it would be impossible for us to survive. It is, in fact, no exaggeration to say that a primitive ability to classify is essential for the survival of men, and also of animals.

In man, however, the ability and the urge to interpret observed phenomena goes much further than the primitive form of classification practiced by animals. Relationships of place, situation and distance are accorded particular attention by primitive peoples. (12) Even limited awareness of such abstract relationships and their expression in language provide a basis for achieving decisive advantages in the struggle for existence. One of the most important of these advantages is the ability to communicate with other persons. By using gestures, sound signals, and spoken language, a basis is provided for group effort, which characterizes the activity of mankind, regardless of degree of civilization, in coping with the forces of nature. Furthermore, it becomes possible to transmit previously acquired information and knowledge between individuals, between groups, and between successive generations. The use of written symbols further facilitates the transmission of information and knowledge across the barriers of time and space. Such symbols, at first usually pictographic in character, tend to become more abstract in nature as civilization evolves. This urge toward precisely defined abstraction leads to the development, in mathematics and related branches of logic, of a type of symbolism for denoting carefully defined abstractions and the relationships between them. The very precision of mathematical formulation renders it unsuitable as a language for expressing many observations and conclusions, particularly those of a qualitative nature. Descriptive language still plays a dominating role in communicating in such important fields as chemistry, biology, and at least certain branches of physics. On account of the limitations of written language and mathematical symbolism, other forms of symbolic expression such as various forms of maps and diagrams are also in

widespread use to communicate various specialized kinds of information.

From this brief review certain basic facts emerge:

1. Limitations inherent in our senses and other means for observation and perception prevent us from discerning and learning everything about any naturally occurring object or event.
2. Perceptual data and observations, if they are to guide our actions in a fashion favorable to our interests, must undergo interpretation.
3. Such interpretation is closely linked with the generation and definition of concepts and the invention of various forms or symbols for communication by speech or by writing.
4. Spoken or written messages are never able to communicate our sensory perceptions in full detail. (It is notoriously difficult, for example, to describe a strange odor or taste. It is difficult to describe different shades of color, and impossible to describe them to a congenitally blind person).
5. Mathematical symbolism is, by comparison with natural language, more abstract, more precise, and less expressive. (In scientific and technical documentation, concepts and relationships are expressed not only by mathematical symbols, but also by words and by other forms of symbolic expression).
6. The incomplete and fragmentary character of any one person's observations, experiences, and abilities makes communication a necessary prerequisite to achieving group action. (This is equally true for a savage tribe engaged in hunting buffalo and for present-day scientists engaged in developing atomic energy).

This situation may, perhaps, be summarized by saying that the ability to conceptualize is essential both to coping with the infinitely varying phenomena of nature and to communicating to achieve success in group activity. Communication is based on various symbols (gestures, speech sounds, writing of various forms) and requires not only transmission of the symbols as such but, to achieve a useful purpose, the interpretation and understanding of the significance of the symbols.

These basic considerations are valid for a much broader range of human activity than present-day science and technology. Within this somewhat restricted field, the experience of many years has resulted in the rather firm establishment of a set of principles which guide the generation and application of concepts. These principles, taken together, constitute the basis of the scientific method. (14) Although the precise formulation of these principles continues to provide philosophers with subject matter for verbal controversy, there can be no doubt that a very wide range of phenomena can be referred to by abstract concepts such as "energy," "force," "atom," "melting point," etc. These concepts and others of like nature have proved extremely useful in analyzing the phenomena of nature, in correlating observations, in communicating the results of observation and correlation, and in predicting and controlling, to an extensive degree at least, the behavior of material objects and processes involving them. Not only are observations — either direct, through our sensory organs, or indirect, through instruments — advantageously interpreted, correlated, and communicated in terms of such concepts, but also the investigation of relationships between concepts has been found to be an advantageous and rewarding undertaking. Thus are constructed those monumental edifices of intellectual achievement known as scientific theory, whose practical utility and power are attested by the very existence of our technological civilization.

In considering what this means in terms of documentation methods, it is well to keep an additional fact in mind: Scientific concepts and theories are not constructed once and for all time to come. Rather, they are like the street system of a large city. They are constructed in the light of present knowledge and practical requirements but with an eye to probable future developments. Nevertheless, new observations — and here we must include more precise measurements, as well as new ideas for achieving correlation — have resulted repeatedly in an older theory being superseded by a more refined and penetrating analysis better able to interpret and correlate a broader range of observations. (15) Thus our ability to predict and to control is extended, perfected, and rendered more efficient. A later section of this paper is devoted to further discussion of documentation and the role of concepts and symbols.

Documentation, Its Purposes and Limitations

From the foregoing, it is perhaps evident that scientific and technical documentation is called upon to serve a number of purposes which are closely inter-related. As a consequence, they are sometimes confused, with consequent misunderstandings as to the nature of documentation methods, their scope, and their limitations.

Let us recall to mind that the basic purpose of documentation methods is to facilitate the efficient use of documents and other graphic materials that are prepared for the purpose of recording, with the aid of appropriate concepts and symbols, those observations that are regarded as sufficiently important to justify the time and effort required for recording them. It is perhaps well to emphasize again that this recording step inevitably involves both a selection of what is to be recorded and also an interpretation on the basis of currently accepted theory and concepts – or, perhaps somewhat more precisely, interpretation on the basis of a given observer's understanding of what theories and concepts are or should be currently accepted.

The general purpose of recording information is to make it available subsequently. In the simplest case, the written or graphic record serves to supplement the memory of the person who prepared the record. Far more frequent and more important is the general case in which information is transmitted to some other person. Under present day circumstances, such transmission of scientific or technical information may result in the receiver applying it advantageously with very little delay. As a consequence, documentation methods have an important newspaper function to fulfill. Here the present-day rapid rate of generation of new scientific and technical information is the source of serious problems. Even the most diligent of specialists have difficulty in reading material directly pertaining to their field of specialization. Abstracts that summarize new work must often be relied upon to provide current awareness which, in the more leisurely times of the not so distant past, was achieved by reading the full length papers. Outside an area of specialization, summary reviews must serve to provide awareness of trends and important developments.

Newly acquired scientific and technical information may not find application until considerable time has passed. As a consequence, it is advantageous to store the records of scientific and technical work for future consultation. The factual character of these two statements is well recognized. Some of their important implications are, unfortunately, often overlooked.

In the first place, a fundamental difference in modus operandi distinguishes communication involving direct transmission of a message from sender to receiver, for example, by telephone, from communication involving use of a file or library of graphic records. Direct transmission of a message may be diagrammed as shown in Figure I, where the means for accomplishing transmission may vary considerably in nature and might be exemplified not only by the telephone but by a letter, the spoken word or by visual images transmitted by television. In a simple transmission, as diagrammed in Figure I, the sender determines what symbols or images are to be transmitted. As long as the transmission is not interrupted the receiver does not determine what will be transmitted, as is sometimes annoyingly apparent when listening to the conversation of boring persons or when viewing television.

The situation is quite different when a file or library of graphic records is used to communicate information. In this form of communication the receiver must play a dynamic role. A collection of graphic records remains inert until the user takes the initiative. This requirement that the receiver take positive action may be diagrammed as shown in Figure II. Until the receiver approaches the record collection, it functions as a reservoir for receiving and accumulating incoming messages.

The number of messages in an extensive file usually precludes personally inspecting all of them each time information is needed. An attempt may be made to use a file by taking items at random, as is sometimes done when browsing. The low probability of turning up material of pertinent interest usually discourages this approach.

The virtual necessity for the user to make purposeful selections when using collections of records means that much time and effort may be conserved by processing such records in anticipation of future use. A consideration of fundamental importance in this connection is the virtual impossibility of foreseeing – in more than a general way – what situations and problems may provide opportunity for advantageous

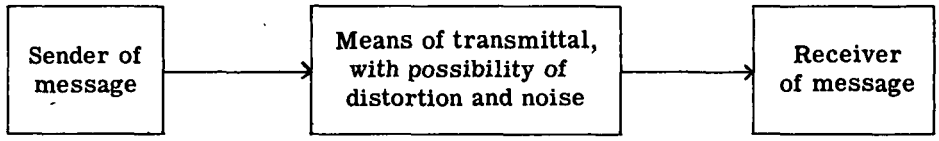

Figure I

Diagram of Direct Communication of Message

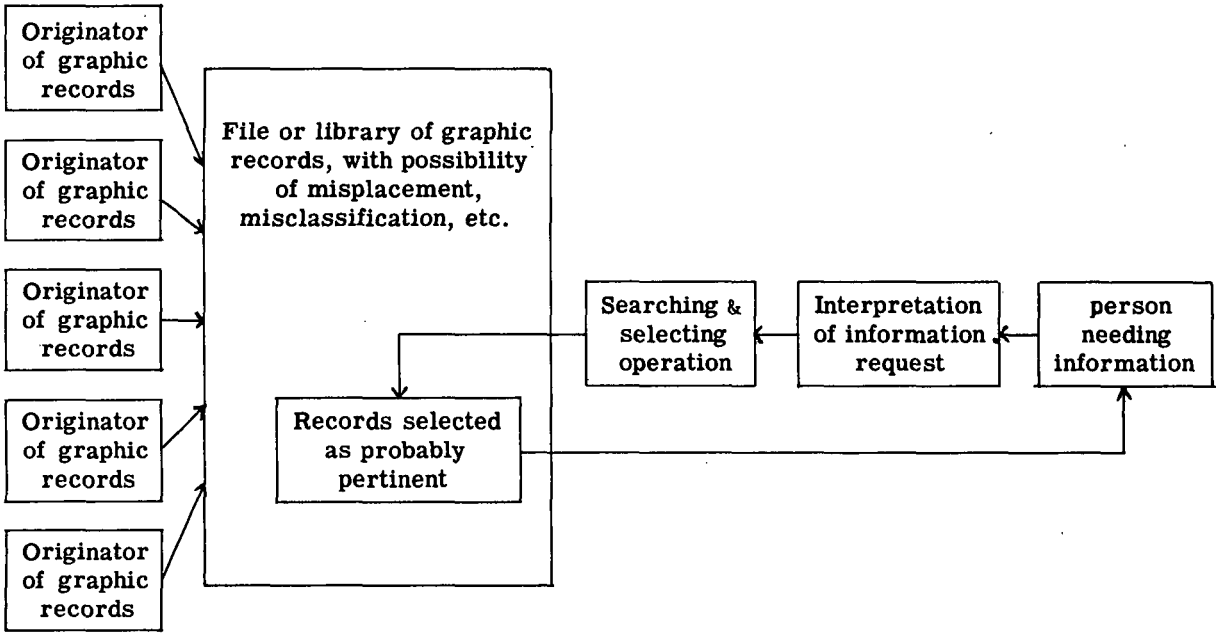

Figure II

Diagram of Communication Through a Collection of Graphic Records

application of the information stored in a given record. Nor is it possible to foresee how various pieces of information may eventually be correlated to meet future situations and new problems. This inability to foresee the future imposes limits on what can be accomplished when indexing, classifying or coding information for future retrieval. It is not possible, in short, to link future uses of information directly to the subject content of graphic records. In this sense, at least, the analysis and processing of information by documentalists can never be as complete as the users of indexes, classification systems and coding might wish.

Documentalists can scarcely do more than provide a multiplicity of useful links (or leads) between the information being processed and the currently accepted framework of scientific and technical theory. The user of a documentation system must interpret his problem in terms of the framework of theory and concepts with which the documentalist has linked up various items of information. (16)

A knowledge of basic theory and concepts must provide, therefore, the basis both for establishing and for using scientific and technical documentation systems. The newer documentation methods, by providing greater flexibility of

selectivity and by accomplishing complex selection operations with the aid of automatic equipment, can facilitate access to recorded knowledge and, in particular, its correlation. No documentation methods conceivable at the present time can, however, relieve the user of the necessity of analyzing his research assignment or similar problem as to what information would be useful or needed. For example, a research problem to extend the useful life of spark plugs by developing better insulators must be interpreted by the research man assigned to the problem with regard to what information is likely to prove helpful. Thus the research man might decide to request papers dealing with the physical and electrical properties of porcelains or he might request information on ceramic materials that are characterized by certain properties.

Furthermore, the user of an information system may find it advantageous to present a series of information requests to a comprehensive documentation system. Thus, information provided in response to an initial request may syggest that another, perhaps somewhat different, information request may turn up additional information useful in solving a given problem. As will be discussed in a subsequent paper on searching strategy, the effectiveness of a documentation system may be considerably enhanced by giving thought to the planning of a series of searching operations. This is particularly true with machine searching systems because of their greater flexibility of selectivity and resultant effectiveness of correlation.

It should also be noted that using a documentation system to satisfy a series of information requests has much in common with carrying through a series of experiments in the laboratory to acquire needed knowledge. With both ways of acquiring knowledge, an answer to one inquiry may suggest a previously unanticipated possible solution or step toward solution. It is perhaps obvious that coordination of library and laboratory research may lead the thoughtful research man first to the library, then to the laboratory, and then back to the library, with the cycle repeated several times during a lengthy or difficult assignment.

Documentation Concepts and Symbolism

As previous discussion has pointed out, we deal with the surrounding material world by taking into account various relationships such as those involving a thing and its attributes, processes and accompanying circumstances, interactions and their results, etc. Human language, in its wide range of variations has provided various sets or frameworks of concepts for observing and expressing those relationships to which attention is directed in dealing with the material world. Different languages may conceptualize various relationships in quite different ways. Thus a concept as fundamental as the past, present or future time of an action or its complete or incomplete nature may be conceptualized in different fashions in different languages. Here we have a parallel with the variations in conceptualization that may and do occur as science develops. The Einstein conception of time as described by relativity theory is quite different from the absolute time on which Newton based both his laws of motion and also their application to describing and predicting planetary and similar motions observed in our solar system.(18)

The task confronting documentalists would be much simpler and easier than it is, if the concepts basic to science, technology, and other fields of learning could be sharply defined once and for all time. The situation with regard to the conceptual framework of science might be compared with the street system in New York City. Repairs and changes are going on continually, yet, during an interval of a number of years, there may be few really important changes in many neighborhoods. Now and then, new conditions — in science this means new observations and new ideas for correlating observations — compel a very considerable change in an old, previously well-established neighborhood. And in the suburbs — in science on the outskirts of knowledge — building is going on at a rapid rate with tentative plans being proposed, many of which are never carried out. Hypothetical proposals for sweeping changes may also be advanced for reorganizing old neighborhoods and, now and then, such rebuilding of street systems and of scientific theory does in fact occur.

It is, of course, not enough for documentalists to cope with those situations in science for which the conceptual framework is well-established and seems unlikely to change. Documentalists must also work with the papers written by pioneers who are exploring new fields of science and developing new branches in technology. The available conceptual tools may leave much to be desired, just as pioneers working experimentally

may be handicapped by lack of laboratory equipment. It is no exaggeration to say that the lack of well-defined and generally accepted concepts may confront documentalists with difficult problems. In this connection some of the principles of information theory may, perhaps, be helpful in guiding our efforts to achieve the best possible results. This possibility will be considered further in a subsequent paper in this series.

Inadequacies, changes and uncertainties as to concepts are not the only source of difficulties for the conscientious documentalist. If we are to work with concepts we must represent them by certain signs or symbols. There is always danger, as Korzybski has so eloquently pointed out of confusing a sign or symbol either with a concept that is represented or with the external reality to which a given concept may be applicable. (19) Meaningful symbols denote concepts and, at best, provide a mapping of concepts. For example, a molecular structural formula of an organic compound as well as the corresponding space model are mappings of the present concept that chemists have of the compound in question. The concept of the structure of a chemical compound is, however, not the substance in question. Similarly the symbol, "e," in a physics text may be used to represent an electron. But this symbol is, obviously, not the concept of an electron, nor is the concept in turn a material object. Rather, the concept is a convenient tool, though a rather sophisticated one, for interpreting, correlating, and predicting the results of experiments and observations.

In working with symbols, the documentalist is, so to speak, twice removed from the outside world which is the source of direct observational experience. Yet it is this world whose prediction and control constitute the goal of science and technology. In contributing his share to achieving this goal, it behooves the documentalist to take heed of the nature of his tools — both conceptual and symbolic. It is, for example, not only naive, but also highly inadvisable, if consistent reliability in retrieval and discrimination is to be achieved, to base a documentation system directly on the words found in reports or similar documents. Documentation cannot hope to rise above a low level of effectiveness if it is based on the words found in documents rather than on an understanding of the latter's subject content. The effectiveness of words and other symbols depends on the care and precision with which they are used. The meaning of words is subject to continuing erosion and change by virtue of more or less gradual shifts in usage. Such uncertainties with regard to words are among the reasons for widespread use of alternate symbolism; e.g., structural formulas in chemistry, wiring diagrams in electrical engineering, special symbols in various branches of mathematics. Such symbolism often has the additional advantage of more concise representation of concepts and relationships than could be achieved by using words. The efficient exploitation of special symbolism is an important factor in the construction of machine language, as will be discussed in subsequent papers in this series.

Approaches to Developing Machine Language

The preceding discussion has been directed to considering the generation and application of knowledge and the closely related purposes that documentation is called upon to serve. Such a review of fundamentals appeared necessary in view of widespread uncertainty as to what is required to develop machine language in an effective form. Such uncertainty may have been responsible for a number of approaches to machine language being proposed. The more important of these proposals might be summarized as follows:

I. *Word Indexing*

The simplest proposal is to designate the subject content of each document to which searching operations are to be directed subsequently by a set of words found in the document itself. This approach leaves out of consideration the flexibility of natural languages in general, and English in particular. As a consequence of such flexibility, it is highly probable and, in fact, virtually certain, that important aspects of subject content will be expressed differently in different documents. Synonyms and near-synonyms are not the only source of difficulty and confusion. Combinations of words may be used in one document to indicate the same aspect of subject content that is denoted by a single word in another document. In the extreme case, which is not necessarily rare, an important aspect may be left implicit in one document, but rendered explicit in others, either by some word or

combinations of words. As a consequence, word indexing, when consistently applied to a large file of complex subject matter, can be expected to lead to a low retrieval factor. As long as ineffectiveness of retrieval is unimportant or disregarded, dissatisfaction with the results provided by word indexing may be avoided.

II. *English as the Machine Language*

It is generally recognized that the complexity of English sentence structure presents extreme difficulties to programming automatic equipment so that important aspects of subject content may be recognized with a high degree of reliability. It is somewhat surprising to observe that many persons have difficulty in foreseeing and avoiding the pitfalls of word indexing, while few, if any, advocate that English or any other natural language should be adopted as a machine language for searching and correlating purposes. Perhaps the explanation is to be sought in the obvious complexities of English phrasing and sentence structure on the one hand, and in a persistent confusion of words as symbols with their meanings on the other hand.

III. *English as a Basis for Machine Language*

As noted already, it is easy to see that English is not suitable for direct use as a machine language. It is sometimes urged that the development of machine language should proceed by conducting a detailed analysis of English in order to arrive at a set of logically consistent modes of expression which would then be the machine language. Also, this proposal would look to the establishment of rules for reducing any English sentence to logically consistent form. Carrying through this program would certainly constitute a major contribution to linguistics. There are, however, reasons for questioning whether this approach would lead to the development of a machine language of optimum effectiveness. It is well known that no universal "natural" logic underlies human languages. (20) Rather, each language is based on certain postulates as to which types of observations and correlations are important. For languages pertaining to the same group, differences in basic postulates may be minor in nature. Thus, in studying sister languages of the Indo-European group such as French, German, or Russian, we do not come into contact with many strange postulates of the kind observed, for example, in the languages of the American Indians. (12,17,20) For this reason, the arbitrary nature of the basic postulates of English is apt to escape our attention, just as an English-speaking person who never learns a foreign language is likely to remain unaware of the arbitrary conventions that are involved in the use of English auxiliary verbs, prepositions, and other connectives. (21)

Attempting to derive by analysis of English a basic machine language would lead, at best, to formulation of the basic postulates underlying our language. The basic theoretical structure of chemistry, physics, or any other branch of science would neither be revealed nor taken into account in such a study of English. To insist that the postulatory structure of English, or any other natural language, be adopted as a basis for constructing a machine language is equivalent to assuming that this postulatory structure is well adapted for recording aspects of meaning in order to provide a basis for machine searching. This seems highly doubtful when we consider the ancient origin and historical evolution of English, on the one hand, and the specialized purposes to be served by machine language, on the other hand. (17,20)

IV. *Boolean Algebra as the Basis of Machine Language*

As previously discussed, the theory of class definition provides a convenient basis for planning and conducting searching operations to be performed by machine. This fact appears sometimes to mislead careless thinkers into assuming that automatic equipment can be used to search messages (e.g., abstracts of scientific papers) only if their contents have been expressed in Boolean algebra or in some other form of symbolic logic. It seems likely that this false conclusion arises from confusing the definition and planning of scanning and searching operations with the content and the organization of the messages to be searched. Actually, any aspect of subject content, that is to say, any concept or relationship between concepts, may serve as a basis for defining and carrying through machine searching and selecting operations provided only that the aspect in question is clearly and unambiguously stated. As we shall see in a subsequent paper devoted to encoded abstracts, a variety of devices can be used to express aspects of subject content in

such a way as to render them available as reference points for planning and conducting searching operations. The basic requirement that aspects of subject content be clearly and unambiguously stated makes obvious the reason why machine scanning and searching operations, defined in terms of the theory of class definition, can scarcely be expected to produce reliable and satisfactory results, if the messages being scanned are expressed in the English language. The wide variation in wording and sentence structure makes it very difficult to predict with acceptable accuracy what words or combinations involving words and other symbols will be used to express and to record various aspects of subject content.

V. *Organized Designation of Subject Content Aspects*

As previous discussion has pointed out, our ability to cope with the ever-changing, infinitely variable phenomena of observable nature is based on our organizing our perceptions and observations in terms of the features, properties, and characteristics that pertain to various observable entities and the processes and changes in which they may be involved. Such organization of perceptions inevitably involves conceptualization; that is, the formulation of concepts or constructs which are not limited to features, characteristics, and properties, but also extend to entities, processes, and relationships between them. Conceptualization provides a framework of reference to which observations may be related so that their significance may be evaluated and decisions as to appropriate action may be made. Expressing, communicating, and recording observations and their evaluations are functions of various systems of signs, of which language has been and still remains the most important in most fields of human activity.

Among primitive peoples it is, perhaps, possible that conceptualization and the related description and evaluation of phenomena may proceed at a single level equally well understood by all who constitute a given community. Uniformity of conceptualization, description, and evaluation of phenomena are impossible in our industrial civilization, based as it is on specialization and diversification of activity. Describing an automobile to a housewife so that she can operate it is an entirely different matter than the description by an automotive engineer of the same performance characteristics. In some fields of learning; e.g., physics; conceptualization has reached the point that words are being superseded to a steadily increasing degree by mathematically defined symbols. Even in those fields in which words still can be used effectively for human communication, we must take care that the concepts they designate are clearly understood. Much of our formal professional education is devoted to imparting an understanding of concepts, just as much of the mental effort of early childhood must be devoted to learning to speak.

Such — in outline form — are the considerations that have led us to base the construction of machine language on the organized designation of subject content aspects. Here the word "aspects" is used in a generic sense in include both the full range of concepts from the most specific to the most generic, and also the different relationships between them. As will be set forth in a subsequent paper, it is convenient to distinguish between two types of relationships. One of these, which we shall term "analytic," pertains to the scope of meaning of concepts and, consequently, to the definition of corresponding terminology. Thus "glucose" is a "sugar" by virtue of definition of these terms. Similarly, "sulfonation" is a "chemical reaction." The other relationship, which we shall call "synthetic," relates to assertions made with the aid of concepts and related terms. An example of such an assertion is "man bites dog." Here the synthetic relationship between "man," "bites," and "dog" is to be carefully distinguished from the analytic relationship between "dog" and "animal."

In carrying through this approach to developing machine language, it is perhaps obvious that we may select, as component units of our machine language, any concepts and relationships that may be appropriate to the purposes to be served. Furthermore, we may establish and define various relationships, both analytic and synthetic, in any way that may be appropriate. Consistency is, of course, of primary importance if machine searching is to be both reliable and efficient. Inconsistencies will lead to uncertainties in machine performance unless we compensate for inconsistencies by special precautions when programming the searching operations to be performed by the automatic equipment.

In working out this approach to machine language, as will be described in a later paper in this series, it is necessary to take certain

practical considerations into account. Some of the more important of these will now be discussed briefly.

Fundamental and Practical Factors

The fundamental factors that underlie the development of machine language relate, as we have seen, to human limitations and knowledge, to documentation, its purposes and limitations, to the role of concepts and symbolism in documentation. It is obviously impossible to treat these important subjects exhaustively within the limits of this paper. We trust, however, that this discussion may contribute, nevertheless, to clarifying uncertainties as to machine searching in general and machine language in particular.

Fundamental considerations determine the grand strategy for the development of machine language. Details of design, on the other hand, must take into account various practical considerations if machine language and machine searching are to provide optimum advantages. These practical factors are controlling not only with respect to costs of establishing and operating machine searching systems but also with respect to their performance and usefulness.

As already noted, machine language must express aspects of subject content so that they may serve as reference points for defining and conducting machine searching operations. More specifically, achievement of advantageous results requires that these reference points be set up in such a fashion that information requests of practical importance, regardless of their scope, can be readily serviced. Furthermore, it is not enough that the machine shall direct attention to all or an acceptably large fraction of the pertinent documents (high recall factor), it is also essential that the person requesting information shall not be overwhelmed with large amounts of non-pertinent material. In other words, the "pertinency factor" must also be kept at an acceptably high level. (10) These considerations mean, in practical terms, that the development of machine language must take into account not only the necessity of recording, in explicit form, aspects of subject content that permit pertinent documents and graphic records to be identified by automatic operations. It is also necessary to take heed of the necessity of achieving rejection of documents and graphic records that are without pertinent interest to the given problem or situation that confronts the inquirer. In meeting this latter requirement, specification of synthetic relationships can prove particularly advantageous. On the other hand, making analytic relationships explicit is particularly advisable in constructing machine language so that it may provide reference points for positive identification of documents and graphic records. It would be misleading, however, to imply that one or the other of these two types of relationships serves exclusively either for identification or rejection in conducting searching and selecting operations.

It is necessary to assess with care the advantages to be achieved by designing machine language so as to make available a wide range of aspects, especially relationships, as reference points for defining and conducting searches that are directed to complex subject matter. It is quite possible to pass the point of diminishing returns by constructing machine language so that many aspects of subject content not useful in conducting searching operations are rendered explicit and recorded in form to be scanned for machine selection. In the extreme case, an excess of zeal may lead to an attempt to state explicitly all factually true aspects and all logically valid relationships. In developing machine language, care must be taken to ensure that a realistic evaluation is made of aspects of subject content that are actually useful in conducting identifying and rejecting operations. Experience indicates that surprisingly simple forms of machine language can meet rather severe requirements.

From what has been said, it is perhaps obvious that simplicity of design of machine language, commensurate with satisfactory selecting and discriminating power, is highly desirable. It may be worth noting that the principal factor that makes simplicity desirable is not the nature of machine operations. If each element of complexity in a machine language is unambiguous and explicit in character, routine operations may be set up so that the machine will not fail in performing its assigned functions. Freedom from ambiguity and clear definition of elements of complexity may be regarded as the demands that the automatic equipment makes on machine language. Highly complex sets of operations are readily performed by modern electronic devices.

The situation, however, involves another

important consideration. Before machine searching becomes operational, the important aspects of the subject content of documents and graphic records must be expressed in machine language. This basic step of interpretation by its very nature must be performed by some person who understands both the subject content of the graphic records and also how to express their important aspects in machine language. This interpretation job can be facilitated and unnecessary costs avoided by taking care that machine language is designed so as to be kept free of complexity that does not pay its way by providing useful discriminating ability.

Another possibility for coping with the cost problems that the task of interpretation presents is to coordinate the interpretation step with other processing so that additional purposes are served. Thus, to consider one possibility, the interpretation of subject content to make it amenable to machine searching can be coordinated with the generation of telegraphic style abstracts that are particularly well adapted to prompt publication of an abstract bulletin to provide up-to-the-minute current awareness of new developments. Prompt reporting is important, not only in science and technology, but also in the fields of law and legislation. A considerable measure of success has already been achieved in combining prompt abstracting with analysis of the subject matter of documents preparatory to machine searching.

The coordination of interpretation for machine searching with the generation of alphabetized subject indexes also may provide important operating economies. Similarly, when conventional classification of documents is advantageous in meeting certain needs, a coordination of processing operations may afford important economies.

Another possibility for serving multiple purposes is to conduct interpretation for machine searching in such a way that the statements of subject contents can also serve as a basis for the automatic equipment to bring these statements into interaction, either with each other, or with data externally supplied. Such interaction might involve computations of an arithmetic nature, or deductive processes based on logical relationships. Information would not only be retrieved and correlated, but would also be processed to produce new conclusions. This important realm of possibilities is outside the realm of machine literature searching and will not be considered in detail in this series.

Concluding Remarks

Our consideration of factors underlying the design of machine language has led us to consider a number of fundamental matters relating to the nature of human knowledge, the formulation of concepts, and their role both in interpreting and in communicating observational and theoretical subject matter. Various theoretical and philosophical problems encountered in the area often give rise to differences in opinion. Nevertheless, it is possible to discern an encouragingly wide area of agreement as to important fundamental principles which, together with various practical considerations, can serve to provide guidance in developing machine language. Our discussion of fundamental and practical factors can scarcely be regarded as exhaustive; yet enough may have been said, perhaps, to indicate the course we believe to be most likely to lead to success in developing machine language.

Is it realistic to feel confident of achieving success? Is it possible to develop machine language and related methodology so that machine searching systems will provide practical advantages? We believe these questions must be answered in the affirmative. Eventually, of course, the doubts of sceptics can be overcome only by large-scale operational success. As with atomic power plants, doubts may fade slowly during several years or more. We are hopeful, however, that various papers in this series, by outlining the construction and application of machine language in relatively simple form, may provide encouragement to the optimists, whose vision of the future is sometimes more realistic than the doubts of sceptics.

References

1. "Mechanized System Launches New Era for Literature Searching," Staff Report, Chemical and Engineering News *30,* 2806-2810 (1952).
2. Bernier, C. L. and Crane, E. J., "Indexing Abstracts," Ind. Eng. Chem. *40,* 725-30 (1948).
3. The importance of care in defining and using terms in constructing classification schemes becomes evident on considering the classification systems of Patent Offices, both American and foreign. Note, in particular, the "Classification Bulletins" of the U. S. Patent Office.

4. Crane, E. J., and Bernier, Charles L., "Indexing and Index-Searching," Chapter 23 in Casey, Robert S. and Perry, J. W., "Punched Cards. Their Applications in Science and Industry," Reinhold Publishing Co., New York, 1951, pp. 331-350.
5. Bush, Vannevar, et al., "Report to the Secretary of Commerce by the Advisory Committee on Applications of Machines to Patent Office Operations," Washington, D. C., December 22, 1954.
6. Vickery, B. C., "Developments in Subject Indexing," J. Documentation, *11*, 1-11 (1955).
7. Dyson, G. Malcolm, "Advances in Classification," J. Documentation, *11*, 12-18 (1955).
8. Luehrs, Fred U., Jr., Kent, Allen, Perry, J. W., and Berry, M. M., "Machine Literature Searching. VII. Machine Functions and Organization of Semantic Units," American Documentation *6*, 33-39 (1955).
9. Perry, J. W., Berry, M. M., and Kent, Allen, "Machine Literature Searching. VI. Class Definition and Code Construction," American Documentation, *5*, 238-44 (1954).
10. Kent, Allen, Berry, M. M., and Perry, J. W., "Machine Literature Searching. VIII. Operational Criteria for Designing Information Retrieval Systems," American Documentation *6*, 93-101 (1955).
11. Ayer, J. A., "Language, Truth and Logic," Victor Gollancz, London, 1948, pp. 33-44, 87-102.
12. Cassirer, Ernst, "The Philosophy of Symbolic Forms, Volume I. Language," Yale University Press, New Haven, 1953, pp. 198-215.
13. Bridgman, P. W., "The Nature of Physical Theory," Princeton University Press, Princeton, 1936, pp. 47-58.
14. Margenau, Henry, "The Nature of Physical Reality," McGraw-Hill Book Co., New York, 1950, pp. 56-101.
15. Bell, E. T., "Mathematics, Queen and Servant of Science," McGraw-Hill Book Co., New York, 1951, pp. 277-287.
16. Soule, Byron A., "Searching the Literature," Chapter 24 in Casey, Robert S., and Perry, J. W., "Punched Cards. Their Applications in Science and Industry," Reinhold Publishing Co., New York, 1951, pp. 353-365.
17. Chase, Stuart, "The Power of Words" Harcourt, Brace, and Co., New York, 1954, pp. 102-109.
18. Reichenbach, Hans, "The Rise of Scientific Philosophy" University of California Press, Berkeley and Los Angeles, 1951, pp. 144-156.
19. Korzybski, Alfred, "Science and Sanity," 3rd edition, The Institute of General Semantics, Lakeville, Conn., 1948, pp. 223-235.
20. Werkmeister, W. H., "A Philosophy of Science," Harper and Brothers, New York, 1940, pp. 108-139. Note especially pp. 137-139.
21. Gode, Alexander, and Blair, Hugh E., "Interlingua Grammer," Storm Publishers, New York, 1951, pp. 50-51.

DEFINITIONS OF DOCUMENTATION

The *ad hoc* committee of the American Documentation Institute selected to judge the results of the competition for a definition of *documentation* have announced the following awards:

"The science of ordered presentation and preservation of the records of knowledge, serving to render their contents available for rapid reference and correlation." Dr. G. Malcolm Dyson, Loughborough, England.

"The *procedure* by which the accumulated store of learning is made available for the further advancement of knowledge." Atherton Seidell.

"The art of facilitating the use of recorded, specialized knowledge through its presentation, reproduction, publication, dissemination, collection, storage, subject analysis, organization, and retrieval." Mrs. Helen L. Brownson, National Science Foundation.

The Committee reports that, though a number of excellent definitions were received it based its choices on the belief that a definition should encompass both theory and practice, emphasize the quadruple aspect of production, organization, retrieval, and dissemination, and suggest the social implications of these activities.