THE COMPUTER AS A TOOL FOR LEGAL RESEARCH

William B. Eldridge* and Sally F. Dennis†

The man with a new idea is a crank until the idea succeeds.

—Mark Twain

I

The Problem

The impetus for the recent surge of interest in computers as an aid in legal research is doubtless obvious to those most deeply involved in research labors. The mass of material that must be dealt with by lawyers, both scholars and practitioners, is awesome. The absolute amount of literature currently being added to our storehouse offers cause enough for concern over the inadequacies of our tools for finding the law. Even more troubling, however, is the fact that the rate of increase in the production of legal literature is climbing as fast as the absolute amount.[1]

Conceding that we are holding our own in the population explosion with the traditional "legal literature," we have also to recognize that the interrelation between law and the other social sciences is growing and may soon become respectable. Some surreptitious hand-holding with the physical sciences is even suggested by the more adventuresome jurisprudes.[2] The net result is that almost all literature may be or may become "legal," at least in some areas of research and practice. Such an undefined mass of opinion and information is well nigh impossible to talk about, let alone treat as the object of experiment. Therefore, legal literature, as the term is used in this paper, quite arbitrarily will be limited to the ordinary and well-understood types—statutes, decisions, regulations, and the literary products of persons and institutions primarily identified with the law. But we will make one encouraging aside to the interdisciplinarians: if the visions of those now engaged in the study of automated techniques for the searching of legal literature come to fruition, artificial barriers between the information stores of the discrete disciplines should crumble. A searching system meeting the requirements to be laid out in this paper should be capable of superimposition on any body of literature. Any researcher, armed with a thorough analysis and understanding of his problem, should be able to wade confidently into the literature of law, biology, sociology, anthropology, or economics (or cosmoprudence!)

* A.B. 1953, LL.B. 1956, Duke University. Project Director, Legal Research Methods and Materials, American Bar Foundation.

† B.S. 1941, Purdue University; M.S. 1948, Case Institute of Technology. Advisory Systems Engineer, International Business Machines Corporation.

[1] Layman E. Allen, Robin B. S. Brooks & Patricia A. James, Automatic Retrieval of Legal Literature: Why and How 1-22 (1962).

[2] Cowan, *Notes on the Teaching of Jurisprudence*, 15 J. Legal Ed. 1 (1962).

The proliferation of written information, sometimes dramatically labelled the "information explosion," not surprisingly has turned considerable attention to organization and identification schemes. Of course, the problem is not a new one; librarians and documentalists have struggled with it for years. Neither is it a problem unique to the law; all disciplines are being inundated by their output. Non-classified and non-hierarchical indexes, statistically prepared abstracts, descriptive word profiles, associative concept searching—these are examples of the techniques being studied. Perhaps the most significant characteristic of these current studies is that they are, for the most part, language oriented rather than subject oriented. This innovation holds great promise for the researcher. It offers the maximum in flexibility and expandability. It offers promise of the day when a researcher who learns his way around in one discipline can approach the literature of any other discipline with aplomb. It means that a fruitful research product can stem directly from study and analysis of a problem without an intervening study and analysis of indexing procedure.

The development of the new techniques of storing and retrieving literature begins with the same decisions that faced the librarians of Alexandria. Essentially these decisions can be expressed in two questions of deceptively innocent mien:

"How shall we ask questions?" and
"How shall we label and store documents?"

It will be quickly observed that literature searching services, like most other service operations, should be designed to meet the needs of their users. If it were possible logically and consistently to describe the needs of the user as he approached the store of legal writing, the task would be half done. No neat verbalization has emerged, although much effort has been expended on it. When the searcher knows what it is that he does not know, it is relatively easy to frame a question and address it to a well-labelled store of information. Thus, many times a lawyer can create a hypothetical case that will resolve his client's problem. Then by studying the method used to label the documents in the collection, he can attach imaginary labels to his hypothetical case and search for actual cases bearing the same labels. This is essentially how lawyers now perform research in case law with the existing manual tools. Given consistent high standards of editing, not-too-complex materials, and a collection of manageable size, this system will work to reasonable satisfaction—*so long as a document never changes in meaning or significance.* Even if the problem of changing significance could be solved, the method still is cumbersome. Most questions involve multiple labels, but under existing manual methods, it is possible to search for only one at a time. The ability to search for a co-occurrence of five labels at one pass of the documents alone would save immense amounts of time.

An even more difficult problem occurs where the researcher knows that there is a gap in his knowledge without knowing what it is that he does not know. Here, knowledge of the labelling system serves only to suggest possible areas which might hold the information to fill the gap. It is at this point that our conventional systems of classification most seriously and adversely affect the process of finding the law. In order to search at all, it is necessary to guess that the gap in the searcher's knowledge occurs in the field marked by this or that label. The guess is repeated in trial-and-error experiments until some label bears fruit. Two deficiencies are readily apparent. First, the waste of time and effort is prodigious. Second, there exists a perpetual dilemma in the decision to stop research; one *is never sure* that he has seen all of the data that he ought to have seen. A recent investigation indicates that this probably is true in all conventionally executed literature research.[3]

It is not enough to study how questions presently are asked by persons doing legal research. Researchers ask questions born of their familiarity with the literature storage system.[4] What we need to know is how people would ask questions if they were unlimited by a system. Without such knowledge it is impossible to formulate a responsive searching system *unless* we are able to develop a system which is capable of responding to *any* question by organizing the material with reference to the significance established by each question.

In labelling or describing documents for later retrieval, whether from a shelf file or a random access computer, it is not enough to find out what use the lawyer makes of a document, and then to label the document so that it will respond to a particular request terminology associated with that use. (On the other hand, it is very difficult to do better, unless one has ready access to a reliable crystal ball.) Decisions molded only by a present understanding of the pertinency of a document result, either immediately or gradually, in the development of classification schemes such as the Dewey Decimal or Library of Congress systems. We know that these finally become so rigid that they sacrifice the flexibility and expandability required to meet the inevitable changes in the significance of stored literature. What we need to know is the use which lawyers make of literature now: that is, the questions to which the literature is responsive, and all future uses which may be made of that literature by virtue of changed conditions and changed interpretations. Without such knowledge, we cannot formulate a permanently satisfactory storage system, *unless* we are able to develop a system capable of frequently and automatically updating itself by the application of new labels to reflect new experience and resulting significance.

Careful consideration of the nature and uses of legal literature in the light even of the cursory discussion offered above can lead us to draw up some general characteristics for an ideal system of storing and retrieving legal literature. While applied here only to legal literature, they probably will apply to all expository writing. These characteristics are ideals, which may be and probably are less than totally obtainable, but

[3] Don R. Swanson, *Interrogating a Computer in Natural Language* 2-3 (mimeo., undated), a paper presented at Congress of the International Federation of Information Processing Societies, Aug. 27-Sept. 1, 1962, available from Thompson Ramo Wooldridge, Inc., Canoga Park, California.

[4] Taylor, *The Process of Asking Questions*, 13 AM. DOCUMENTATION 391 (1962).

they express an objective. Attainment of the objective, if not now possible, will be possible eventually through technological advances, and probably the technology will have advanced before we are ready to use it. The characteristics suggested below will be used in the remainder of this paper as points of reference in the discussion of operational and projected studies of automated searching of legal literature:

1. Techniques for identifying, storing, and retrieving legal literature should be based on language rather than subject matter to the greatest extent possible. This may mean storage without indexing.

2. Such portions of the techniques, if any, as must rely on classification schemes must have a built-in capacity for automatic, large-scaled reorganization and reclassification as needed. Preferably these updating routines should be going on constantly as new data are added to the collection.

3. The system should allow for a wide variety in the searching strategies available to the researcher. For example, it is essential that a collection of judicial decisions be susceptible to searches for identical fact situations and also susceptible to searches for analogy of reasoning and for basic legal concepts.

4. The system should be at once simple enough and sophisticated enough so that the researcher can approach the store of information directly without the intervention of a human buffer. Approaching a store of literature through an indexer is like approaching the literature when it is available only in translation from an unknown language. There is an inevitable loss in the sharpness of the communication between the writer and the reader. If we create an indexless store of literature which can be questioned only by a machine operator who translates our questions into an unknown language, the same loss will occur.

While it is possible that technological limitations may include certain operations at which present-day computers are basically inefficient, we are not devoting consideration to that possibility in our discussion. There are two good reasons for this decision. First, experience would seem to indicate that supposed limitations on the capability of computers are really limitations of our ability to verbalize the operations we wish the computer to perform. When the students of learning theory, decision theory, and human thought processes in general can describe logically the processes by which humans perform mental operations, it seems likely that there will be computers to simulate such operations readily. The second reason is that the assumption of unproved limitations means that we will mold our ideal system on conjectural incapabilities. The more fruitful course seems to be to assume no limitation in the formulation of an ideal and then to back off only where absolutely necessary.

A similar consideration largely excluded from our discussion is the question of economical and efficient operation of an information system. Unquestionably there will be a time and place at which considerations of cost and efficiency will be determinative, but this is not the time or the place. Again, the most fruitful course seems to be the development of a technique responsive to all needs. Then, and only then, can realistic decisions about economic feasibility be made. We cannot foretell what the costs will be until a system has been developed, and we cannot foretell what uses can be made of the system until its capabilities have been determined.

II

THE TOOL

Why consider using a computer for selecting and organizing legal literature?

Succinctly stated, the central problem in information handling is to bring to bear the right information on the right problem at the right time, in recognizable form, and without too much excess baggage. The point we raise in looking at the computer is this: Can we expect help in stripping off the excess baggage in such a way that we can place faith in the quality of what is left when it is applied to our particular problems?

In terms of nuts and bolts, a computer consists of the following:

1. Storage—in which both the instructions to the computer and the data to be manipulated may be accommodated. Physically, storage may consist of such entities as magnetic cores, magnetized drums, magnetized disks, monomolecular films, and eventually, other things. Storage may be available to a computer at different levels of access—that is, some information may be more available and more easily manipulated than other information. When this is the case, we move the information from such less readily available storage as magnetic tape to the more available form, *e.g.*, magnetic core, when the computer actually processes it.
2. Some sort of input device or devices. These can be card readers, tape readers, typewriters, photoelectric readers, electrical signal readers, or any combinations of such elements.
3. Output devices—for example, card punches, tapewriting devices, typewriters, printers, television tubes or electrical signal emitters.
4. A control unit—which enables the computer to carry out the instructions placed in its highest access storage unit.

Functionally, a computer in its pristine form can accomplish only the following five tasks:

1. It can hold information (including its program) in cold storage.
2. It can compare two symbols, such as numbers or letters, and recognize that the two symbols are the same or not the same. This capacity, of course, is inherent also in various types of tabulating equipment.
3. It can evaluate the relative ranks of two symbols on some arbitrary scale that has been built into the machine. In practice, this usually is the numerical and alphabetical sequence with which we all are familiar, although the scale could be chosen otherwise.
4. It can assess the result of a matching or ranking and, depending on the result, change its subsequent course of action from one plan to another. In computer jargon, this amounts to "making a logical decision."
5. The last of the list is the one usually thought of first—the computer can perform arithmetic operations on numbers, namely the operations of addition, subtraction,

multiplication, and division. It can perform these operations not only on problem data but also on its own program; by this mechanism, it can modify the program.

Programming allows us to combine these functions in any manner that we choose.

The history of the computer commences in earnest a very short time ago, approximately 1952. The first applications for which the computer was shown to be an economically justifiable slave were those that made rather direct use of the basic functional capacities of the computer and needed to be carried out in large volume. Examples of such applications are payroll and invoicing calculations. The Keyword-in-Context program used to produce the American Bar Foundation's *Index to Current State Legislation* is an example of this class, although it was not one of the early ones.

The next stage in the development of computer usage was born when some engineers and scientists saw that the opportunity existed to procure numerical answers to physical problems of elaborate complexity. That class of calculation is characterized as the high complexity-low volume application. An example would be the determination of the critical size essential for fission in an atomic bomb.

For a time it appeared that all computing belonged in either one camp or the other, and computers were designated "commercial" or "scientific" depending on which use was expected for them. This condition lasted for quite a while (perhaps a year!) until it was discovered that collusion was going on between the "commercials" and the "scientifics" in industrial firms where the management had learned that money could be conserved in some operations by making use of "scientific" calculations.

There now emerged a third level of application characterized by medium complexity and medium volume. Examples of this nature are: (1) the calculation of the least cost mix in a refinery for the manufacture of that producer's spectrum of gasolines, (2) the solution of lengthy algebraic problems, such as those encountered in process equipment design, and (3) the correlation by statistical methods of data obtained from recorded observations.

As such applications grew, people began to realize that the computer embodied a unique property, which was this: It could be made to act like anything else that one desired, if one could manage somehow to define how the other thing behaved. Out of this concept grew the idea of simulation of proposed or real physical, economic, and sociological systems. In physical systems, the definitions encompassed the appropriate physical laws and relationships, and in the other cases they consisted of descriptions of environments and policies to be explored. These definitions were delivered to the computers in the form of computer programs, and the computers were then operated as if they were the systems under investigation, through which history was paraded. Such a simulation is in effect an experiment upon a system; the results in general are the average and extreme results to be expected when the real system is operated for a period of time. The attraction of the simulation is that it allows people to experiment innocuously with systems which could not be disturbed feasibly in real life because either the costs or the risks are too great.

As we study the problem of information retrieval, it may be of help to analyze how we have proceeded from the modest capacity of the rudimentary computer to the more complex applications in other fields. Essentially these problems have been attacked from two directions. From the bottom (that is, the computer) up, there have been developed and collected an assortment of schemes for combining the basic abilities of the computer to produce more sophisticated functions. As each of these small packages has been devised, tested, and approved, it has been preserved in a deck of punched cards or on magnetic tape or in some other form of reserve storage. Some of these techniques have mainly to do with facilitating programming; others are concerned with converting high-level mathematical equations to simple arithmetic.

The second direction of attack has been from the problem down. Before any of the lower-level techniques can become useful in a computer application, good problem definition is essential, and this step frequently consumes the major time and effort expended in constructing a working program. The remaining steps are to make the adjustments and fill in the blank spaces which still exist between the definition of the problem at the top and the tested tools and devices that are available for exploitation at the bottom.

In studying the problems of information retrieval, we can see that there are devices already available which may play an important part in taking us to our goal. Taking these into account, we can expand the list of computer capabilities from the five basic functions to include at least the following:

6. Combining the computer's basic matching ability with some probability arithmetic, we have available a "nearly match." For example, if the word "hereditaments" happens to arrive in a properly prepared computer without the "h," it is possible that the term could still be recognized by means of a probability caculation.
7. Some simulation of man's associative ability is possible. The simplest example of such association would be a table, which is easily plugged into a computer. A somewhat more advanced example would be a dictionary. More sophisticated forms are being developed as the need for them becomes apparent.
8. Symbolic logic can be handled readily on a computer. The essence of this technique is the banking of the logical decisions of the computer so that it can in effect reason: if A and B are true but C is not true, then D must be so, *à la* good old-fashioned syllogism. It is also possible to refine this a bit by estimating that the probability of D's being so is eighty-five per cent. A computer behaving in this way is simulating what man calls "deductive reasoning."
9. Another tool which can be useful in the organization and retrieval of information is the device that can simulate inductive reasoning, *i.e.*, recognize the form that is common to a collection of objects or ideas. This device is available in the sense that we have some knowledge about using statistics and probability as aids to classification.

A number of research groups are devoting their efforts to improving our ability to use computers in this way.

10. An element which may contribute eventually to information retrieval is language translation. The problem of identifying meanings in English is very similar to the problem of translating, say, French to English. There are no translation programs today with which everyone is satisfied, but some elements of the process have been worked out and much more research is being done.

In order to attain computer ability to organize and select information as effectively as we would desire, more work is needed both in defining carefully the nature of the problem and in developing the tools which are capable of solving the problem as it is defined. The major part of the effort now directed toward the exploration of computer applications to legal research is concentrated in this area.

III

The Experiments

At this point we turn to an examination of work actually being done or in preparation for imminent experimentation. Such a review will provide a display case for some of the problems dealt with above, and it will also comprise a fair inventory of the successes and the promise for this type of work. Fortunately, too, the work in information retrieval in the law covers a respectable band of the spectrum of theoretical approaches, enabling us to discuss the art effectively within one discipline.

A. Automated Searching of a Conventional Index

The late Robert T. Morgan, assistant professor of business law at Oklahoma State University, developed a technique known as "Point of Law."[5] In essence his technique is a mechanization of the conventional indexing method with some added advantages in searching. Trained personnel study the legal materials to produce "an analysis of each case for the particular pertinent issues which are actually decided in that decision, dicta, or other material.... We are dealing with concepts rather than words."[6] These trained personnel extract a word, a phrase, or a paragraph which identifies the issue or concept present in the material under analysis. The concept is then assigned a numerical machine code. When a sufficient body of material has been analyzed and coded, an alphabetical list of the concepts is produced by the computer. The list has been likened to a telephone directory with the concepts listed down the left side of the page and the code numbers which identify those concepts to the computer listed down the right. From the list the searcher selects those concepts which identify his problem. The code numbers are then

[5] Morgan, *The "Point of Law" Approach*, 62M Modern Uses of Logic in Law 44 (1962) [hereinafter cited as M.U.L.L.]

[6] *Id.* at 45.

presented to the computer, which searches its store of information for cases that bear all of the identifying code numbers selected by the searcher.

The Point of Law approach may be characterized as an automated and vastly accelerated West Key Number type system. The system has three operational features which distinguish it from conventional manual methods and which are very valuable. First, the system is capable of searching for numerous "concepts" at one time. In conventional searching in the law library this cannot be done, at least not among the reported decisions. Each aspect of a problem must be searched individually. The second feature is that all the law is searched in answer to each query so that when no response is received there is a fair degree of certainty that no precedent exists on that particular question.[7] The third operational advantage is that the researcher may select the type of output he wishes in answer to his requests. "[The print-out] can be, for example, citations alone. It can be citations plus headnotes, or citations plus headnotes plus full text. It can be all of these plus all of the pertinent codes and regulations that may pertain to this particular concept."[8]

In terms of our previously stated requirements, this system is unsatisfactory. It is based upon a highly subjective technique of indexing by present evaluation of the significance of legal decisions. The initial analysis step is expensive. Since, because of the mass to be analyzed, it must be accomplished by many different people, it is difficult to assure uniformity in method of treatment, in accuracy, or in depth of detail. The system does not provide for updating or reorganization of the document file. Once material is entered, it is entered forever unless it is completely removed, re-analyzed by hand and reintroduced under new code numbers. Even this would be only a compromise with present inadequacies, and not a finished solution. Of course, with the literature in machine searchable form, it is much easier to select out special areas for working over. Nevertheless, it is unrealistic to think about doing the whole file over with periodic regularity of sufficient frequency to meet the requirements of our present day rate of change. The system does allow the researcher to approach the store of information directly in his search for answers to his problems, provided he is conversant with the indexing procedure.

The point of law approach has been a valuable demonstration that the tools of legal research can be automated. It has served to draw attention to the possibilities of machine applications to legal literature. It was a stepping off place for research into automated storage and retrieval of case law. It cannot, however, be fairly characterized as more than a speeding up of our present unsophisticated methods.

[7] A recent article in the *American Bar Association Journal* relates, in the course of arguments against the utility of computers, Professor Henry Hart's well-worn tale of research for Mr. Justice Brandeis. Clerk Hart had told the Justice that he had run the digests high and low and found nothing on a requested point. "Whereupon L.D.B. fixed him with a steely eye and asked 'Have you thumbed the reports?' " According to the writer, Clerk Hart was awe-struck. Just think how L.D.B. would have reacted if the clerk had been able to look back with an equally steely eye and reply "Yes." Wiener, *Decision Prediction by Computers: Nonsense Cubed—and Worse*, 48 A.B.A.J. 1023, 1026 (1962).

[8] Morgan, *supra* note 5, at 46.

B. Automated Searching of Full Natural Text

The "Keyword in Combination" system[9] developed at the University of Pittsburgh represents an approach entirely different from that of Professor Morgan's work. In the first place, Professor Horty, who directs the work at Pittsburgh, did not set out to develop an automated legal research system. His group at the Health Law Center turned to the development of new research tools when their primary research tasks became overburdening. The necessity to perform multistate research in health law gave birth to a large-scaled effort to create an electronic searching system for statutory law. It now sometimes appears that the child has gobbled up the mother in Pittsburgh, but that is only because the searching system has attracted more notice than the continuing work in health law. Actually the system is working daily to produce the information needed in the Health Law Center's research program. In this respect the Pittsburgh work is unique. It has, almost from the beginning, been angled at the solution of specific research problems in statute law. This doubtless has been an advantage at times, since the planning was not clouded by misgivings over unrelated potential problems which never materialized. On the other hand, in the long run it may prove to be something of a disadvantage when attempts are made to apply the system to literature other than statutes.

In the Pittsburgh system, the full text of statute sections is entered into the computer via magnetic tape, with each section of the statute treated as a document and identified by a document number. The computer prepares a concordance of the entire document file, listing all the words alphabetically and noting beside each word the document numbers of each section where that word occurs. In addition to the document number, the location of the word in the original text is explicitly described by other numbers which pinpoint the sentence within the section where the word occurs and the position of the word within the sentence. While the machine is preparing the concordance, it also gathers data of considerable value in linguistic statistical studies and in the development of other programs.

Searches are accomplished by the researcher himself. He selects words which he believes should appear in a statute touching upon his problem. He may be aided in this task by a list of the words actually used in the statutes or by a thesaurus. The thesaurus acts as a spur to the imagination leading the researcher from words he has thought of to other words which might also be used. In the Pittsburgh system any

[9] For a description in more detail of the methodology of this system and the results of its applications see the following sources: Horty, *Searching Statutory Law by Computer, Interim Report No. 1 to Council on Library Resources, Inc.* (mimeo., undated); Fels and Jacobs, *Linguistic Statistics of Indexing* (mimeo., July 31, 1962) (available from Health Law Center, University of Pittsburgh); Kehl, Horty, Bacon & Mitchell, *An Information Retrieval Language for Legal Studies*, 4 COMMUNICATIONS ASS'N FOR COMPUTING MACHINERY 380-89 (1961); Horty, *The "Keyword in Combination" Approach*, 62M M.U.L.L. 54 (1962). See alsq E. JONES (ED.), LAW AND ELECTRONICS: THE CHALLENGE OF A NEW ERA 91-143 (1962).

number of artful combinations of selected words may be employed. The technique can be described best by the director himself:[10]

> Searches are initially prepared on paper, with the searcher putting down single words, or words and their synonyms, which define the words or phrases he expects to find in the docum_ents he considers relevant to his inquiry. Thus, if he wished to search for all the Pennsylvania statutes dealing with illegitimate children, he might put into one class the words "baby," "child," "foundling," "infant," "juvenile," "minor," "orphan," and so on, along with their various forms, [thereby] requiring that one of these words, at least, appear in a document for it to be considered relevant. Another class could be established containing the words "father," "mother," "parent," "unwed," "unmarried," "legitimate," and so on.
>
> To specify to the machine the relationship which must exist between the words in context, a certain operator is utilized. One such operator is the word "or," which is used within each class above to tell the machine that either "baby" or "child" or "foundling," and so on, must appear in the document for it to be considered relevant. When it is desired to tie two classes or two words together, the operators "D," "S," or "W" may be used. If as was done in the search above, it was desired that the statutory section contain at least one word from the first class of words and at least one word from the second, the operator "D" is used to indicate that at least one word in each class must appear in a relevant document. Similarly, if a tighter relationship is desired, the operator "S" would be used to indicate that representatives from each of the classes must appear in the same sentence.
>
> In a search involving illegitimate children, in addition to those documents containing representatives of the two classes stated above, it may be desired that certain documents be considered relevant if a certain single word appeared therein such as "illegitimate," "bastard," "parentage," "putative," and so on. If the document containing the phrase "born out of wedlock" is sought, the operator "W+3+3" is used. This operator requires that the word "wedlock" appear in the same sentence, no more than three words after "born."

Once framed, these searching instructions are given to the computer along with additional instructions indicating that the searcher desires the document numbers, the full citations or a complete print-out of the text of the relevant documents. As many as 500 terms can be searched simultaneously by the computer. This means that 50 different searches averaging ten terms per search could be accomplished at one pass of the tape.[11]

It is quickly apparent that the Pittsburgh system is the first to bring to bear on legal research problems computer capabilities other than speed. In the area of statutory law, where the system has been largely utilized, its effectiveness has been sharply demonstrated. Despite certain *ipse dixit* pronouncements to the contrary,[12] Professor Horty and his associates have proved that a completely unindexed body of literature can be searched effectively by a computer. The system is oriented purely

[10] Horty, *The "Keyword in Combination" Approach*, 62M M.U.L.L. 59-60 (1962).

[11] For a description of results of some actual runs of the computer and comparisons of computer results with that of faculty members using conventional tools, see Horty, *supra* note 10, at 60-62.

[12] Yehoshua Bar-Hillel, Some Theoretical Aspects of Mechanization of Literature Searching, U. S. Dep't of Commerce Tech. Report No. 3 (1960); Freed, *Prepare Now for Machine-Assisted Legal Research*, 47 A.B.A.J. 764 (1961).

toward language, relying solely upon the actual words used by the statute writers. Significance is attached only at the time of searching. This means that there is no necessity for any updating or reorganization routine. This is decidedly an advantage. In fact, the Pittsburgh system meets all our stated requirements. This suggests, since we are not totally satisfied in accepting that system as "the ideal," that our requirements may need some refinement.

Mulling over our intuitive dissatisfaction with the Pittsburgh system, we sense that its chief difficulty is that the questioner must know what it is that he does not know. At least, he must know the words which will be used to express that which he does not know, which amounts to the same thing. To rationalize our skepticism on this point we are willing to hazard several guesses:

1. The demonstrated effectiveness of the system is due partly to the nature and uses of statutory law. These are such that the searcher usually can identify the gap in his knowledge. This is certainly true of the types of searches which have been used for demonstration purposes.
2. The different nature and uses of decisional law are such that in a large portion of instances the searcher cannot identify adequately the gap in his knowledge. Especially will this be so where the research is aimed at the basic rationales which permeate all sorts of legal relationships.
3. An adequate selection of search terms will not be so easy in case law, which is not always "carefully framed in words chosen for clarity rather than literary quality."[13]

An extensive thesaurus would help to relieve some of these problems. However, it is quite easy for a subjectively created thesaurus to take on all the ills of a classified index. If a thesaurus is developed, it should incorporate the updating and self-reorganizing facility which the Pittsburgh system has avoided.

We must all, at some time, pay attention to the practicality of these proposed techniques. The Pittsburgh system is far enough along, and its operation well enough established to bring practicality into consideration. The scanning of concordances of millions upon millions of pages of closely printed legal literature in their full text is not today practicable, but new devices suitable for economical storage of immense quantities of material are being developed. The imminently expected arrival of a photoelectric print reader[14] probably will make it feasible to use full text as input. It appears likely that the critical lag exists in the development of the insight needed to harvest maximum utility from these devices.

C. Automated Searching of Full Natural Text Applied to Case Law

Professor Robert Wilson, Research Director of the Southwestern Legal Foundation, has undertaken experiments with the Horty-Pittsburgh system as applied to

[13] Horty's description of statutory language, which we accept as the legislatures' objective if not their accomplishment.

[14] A device which will transfer the printed page directly to magnetic tape without the necessity of punching cards or paper tape.

case law.[15] The material for the experimentation consists of all reported cases dealing with arbitration in five southwestern states.

Several refinements dictated by the nature of the material have been made on the Keyword-in-Combination system. The most important refinement provides for collecting all the various forms of a given word under a common "root" term. Thus "harm," "harms," "harming," "harmed," and so on, would be assigned a single numerical code. All words occurring in the natural text of the selected cases are arranged alphabetically in a list with each word followed by the "root index number" which identifies its basic common root. From this point on, searching, as described in the preceding section, is accomplished by using root index numbers instead of words. This means that in writing a search question one does not have to advert to all the possible suffixes which may appear in a document collection, since all such words will have a common root index number.

The operational value of the root index file in the selection of search terms is considerable. A greater degree of compactness of the concordance is achieved, with a consequent reduction in search time. In order to minimize the pitfalls of subjective decisions about groupings, the task of subsuming words under single code numbers is performed by hand from a complete list of words occurring in the cases. The words are not looked at in context; only orthographical similarities are considered in the groupings.

Experimentation has not developed to the point where it is possible to evaluate this refinement in the terms in which we have discussed the parent system. Further, the selection of arbitration cases will make evaluation very difficult. First, the subject matter has limited the input to a very small number of cases. Second, the artificiality of a severely limited subject matter in the input materials means that a tremendous proportion of the sophisticated work which the computer must perform in a full-size operation has been accomplished by the simple expedient of elimination.

D. Automatic Searching of Material Indexed by a Non-Conventional System

The experiment with legal literature performed at Western Reserve University Center for Documentation and Communication[16] is in some respects similar to Robert Morgan's demonstration. That is, searching is carried out via computer using an index prepared by human analysts from original text. However, the indexing method is not conventional—at least, not from the point of view of the lawyer accustomed to West's Key Number system.

The WRU indexing method, which has been tested extensively in metallurgical literature, consists of constructing a stylized statement of what it is that the reference deals with, using as much as possible the vocabulary of the original material. The

[15] Wilson, *Computer Retrieval of Case Law*, 16 S.W.L.J. 409 (1962).

[16] Melton & Bensing, *Searching Legal Literature Electronically: Results of a Test Program*, 45 MINN. L. REV. 229 (1960); Melton, *The "Semantic Coded Abstract" Approach*, 62M M.U.L.L. 48 (1962).

stylized statement is called a "telegraphic abstract." In the telegraphic abstract the words selected from the original material are tagged with codes known as "role indicators," which serve to identify the individual functions of the words and also to structure the statements into a relatively small number of patterns. Punctuation is used to segment phrases of the statement, and this device further structures the abstract. For example, the telegraphic abstract prepared for section 2-313 of the Uniform Commercial Code[17] can be translated loosely from its machineable form into English as follows: *Goods* were processed by *selling* from a *seller* to a *buyer;* an *express warranty* was defined or created in that a *bargain* occurred from a *buyer* to a *seller* conditional upon an *affirmation,* a *promise,* a *description,* and a *sample.* In this rough paraphrase, the italicized words are those taken from the original text, and the connectives are an approximation of the meaning of the role indicator codes that were used to tag the underlined words.

The analysis described up to this point is performed by individuals skilled in the literature of the field. A further analysis of the underlined words takes place via machine: the English words are looked up in the WRU Semantic Code Dictionary, and a set of codes representing various generic aspects of the original English word is substituted for the original, so that the finally coded form consists of role indicators (three-letter codes), punctuation (commas, periods, and the like), and semantic factors (four-letter codes). The telegraphic abstracts for all of the documents in a given file finally are maintained in coded form serially on magnetic tape, in readiness for computer search. At search time, a question is prepared, which resembles very much the form of a coded telegraphic abstract, and the file is searched for references that match (in the logical sense) the question.

Some comment should be made about the facilities delivered by this unique indexing method. Although the analysis is complex by comparison with other methods, the human effort required is minimized by the restrictions imposed in the format and structure. (Only a small set of role indicators actually are used, and there is a finite number of ways in which they combine naturally; this makes it easy for the indexer to make the necessary decisions.) The format also increases the resolution of searches; that is, it decreases the possibility of "false drops." On the other hand, the generic coding of the key words selected from the source document increases the system's ability to find references characterized by synonyms or terms of broader or narrower significance, and this facility of course increases the probability that all relevant documents will be found.

In summary, the WRU coding system embodies a considerable refinement in analysis over most indexing systems, and its indexing depth probability is greater than that of most conventional systems. These would be desirable attributes to incorporate into an automatic indexing system.

[17] Melton & Bensing, *supra* note 16, at 240-43.

E. Searching of Conventionally Indexed Documents, with Search Profile Augmented by Statistically Associated Index Terms

John C. Lyons, of the Graduate School of Public Law, George Washington University, with the participation of the Datatrol Corporation, has adapted Dr. Edmund Stiles' association factor[18] to the searching of files of documents dealing with antitrust problems.[19] The technique consists essentially of the following steps:

1. *Preparation of the File to Be Searched*

a) Choose (via human analysis) index terms to represent each document in the file.

b) Prepare a "Term Profile" tape using the following procedure: Compute "association factors" relating each unique index term in the entire file with each of the other terms with which it ever co-occurs in a document, using Stiles' modified chi-square formula,

$$\ln\left[\frac{\left(|fN-AB|-\frac{N}{2}\right)^2 N}{AB(N-A)(N-B)}\right] = \text{ASSOCIATION FACTOR}$$

where A is the number of documents indexed by one term, B is the number of documents indexed by a second term, f is the number of documents indexed by both terms, and N is the total number of documents in the collection. Then, for each index term prepare a list (the "term profile") of the other terms with which the given term exhibits significant association. For example, Lyons reports[20] that the term "Clayton Act Section 11" was found in this manner to be associated with the following index terms:

Term	*Association Factor*
Acquisition	3.45
Assets	3.97
Clayton Act Section 7	3.45
Economics	4.08
Legislative History	3.33
Merger	3.50
Reports	3.38
Conglomerate Acquisition	2.82
Congress	3.01
Statute	2.07

Decreasing values of the association factor correspond to decreasing co-occurrence of the pair of terms. Lyons formed a list of those terms for which the association factor was greater than 1.6, and this list became his "term profile," for one term in the system. There was; of course, a separate list corresponding to every other term, and the aggregate of lists formed the "Term Profile" tape.

c) Prepare an inverted file tape. This is the conventional inverted file, in which document numbers are grouped under term headings in such a manner that every

[18] Stiles, *The Association Factor in Information Retrieval*, 8 J. ASS'N FOR COMPUTING MACHINERY 271 (1961).

[19] An address by John C. Lyons, American Bar Association Annual Meeting, San Francisco, Cal., August 7, 1962.

[20] A paper entitled *A Search Strategy for Legal Retrieval*, distributed at the American Bar Association Annual Meeting, San Francisco, Cal., August 7, 1962.

term used in the system is followed by a list of all the documents referenced by that term.

d) Prepare a document file tape, in which the records are ordered by document number and the record entries consist of identifying information as well as a list of the index terms selected to characterize the document and a brief abstract.

2. *Execution of a Search*

a) Choose a set of index terms suitably defining the search question. Lyons uses up to four "AND" operators and an unlimited number of "OR" operators as connectives.

b) Look up the "term profile" for each term in the question. Compare term profiles, and select terms that have a high degree of coincidence among the term profiles; these form a "First Generation" list of terms, which will be used to augment the original list of search terms. First generation terms obviously are statistically associated with the original search terms, but they are not usually synonyms, since one does not ordinarily index a single document with synonymous terms. Therefore, the next step is to

c) Look up the "term profile" for each term in the First Generation profile, compare terms, and select those terms having a high degree of coincidence as the "Second Generation" profile. The Second Generation profile will contain some synonyms or near-synonyms of the original search terms. The list of original search terms, together with those in the First and Second Generation profiles form what is referred to as the "expanded list" of search terms.

d) For each term in the expanded list, compute its average association factor with all of the other terms in the expanded list. The terms which are more highly associated with the other terms in the expanded list are probably the terms most germane to the question being investigated. Select the terms with average association factors higher than some arbitrarily decided threshold for use in searching the inverted file.

e) Search the inverted file, using the list of terms collected in step (d) and use the average association factor of each term as a weight to compute probable relevancy of the documents found under that term. The net result is that each document cited at the end of the search is accompanied by a number, which is the sum of the average association factors (with respect to the question) of its index terms, and this number is used to rank the found documents in order of probable relevancy.

Both in Stiles' application of his method to physics and engineering documents and in Lyons' adaptation to legal documents, the words generated in the First and Second Generation profiles clearly bear meaningful relations to the original search terms. For example, Lyons reported that the following lists of terms were generated in a search of his antitrust file:

Original Search Terms	*First Generation Terms*	*Second Generation Terms*
Clayton Act Section 7	Alternative Source of Supply	Acquisition
		Geographic Area
Assets OR	Appreciable Segment	Geographic Market
Merger	Burden of Proof	Incipiency
	Clayton Act Sec. 11	Market Share
Vertical Acquisitions OR	Clearance Procedure	Process
	Complex Issues	Relevant Market
Vertical	Economic Concentration	Reports

Aspects of Merger
Legislative History
Steel

Economic Power
Economics
Economic Welfare
FRCP Rule 56
Nation-wide market
Stock
Substantially Lessen Competition
Conglomerate Acquisition
Congress

Summary Judgment

In effect, a robot-like mechanism has been used to take over a part of the ostensibly human job of conjuring up all of the index terms that might have been used to identify the information pertinent to a given human question. The mechanism, it will be noted, operates on words—albeit words already highly selected by human effort. The demonstration gives one hope that such mechanisms can be extended and adapted for use in the original organizing of the material.

IV

THE PROBLEM RECONSIDERED

At the end of the first section above were postulated certain characteristics of an "ideal system" for finding the law. These characteristics were derived solely from a consideration of the problems which grow out of the nature and uses of the law, out of the increasing volume and complexity of the law, and out of the increasing interrelation of law and other disciplines. The postulated characteristics were offered tentatively and with diffidence. In the light of the research discussed in section three, it will be observed that they were also offered naïvely. It should be possible now to postulate a refined set of characteristics for the ideal law-finding system.

1. That the ideal system should be based on language rather than subject matter still appears to be a sound approach. As discussed above, one way of achieving a language-based system is to store unindexed raw text. There are, however, certain obvious, and other not so obvious, drawbacks in unindexed storage. The advantages of a language-based system probably can be preserved and some of the drawbacks of indexless storage avoided if we can develop a system in which a machine constructs an index with absolute consistency and objectivity. Two types of advantage result from indexing. First, there may be advantages of economy, such as a shortening of the time lapse between the recognition of a need for information and the satisfaction of the need. Second, there are advantages in the utility of the system to the user, such as increased assurance of complete retrieval of relevant documents and reduction of the burden on the questioner to know the answer to his question. We think that language-oriented, machine-constructed indexes based on linguistic statistics hold promise for the ideal system.

2. A built-in capacity for automatic, large-scaled reorganization and updating is essential where any form of indexing or classification is used. The construction of such routines is clearly feasible for the computer when the indexes to be reorganized have been constructed originally by machine according to consistent and objective routines.

3. An examination of the research in the field of automated law-finding confirms the suggestion that there needs to be a wide variety in possible searching strategies. This is where the work already accomplished falls shortest. In the systems that have indexes, the searching strategies are indexer-oriented and thereby limited. In the indexless systems, the strategies are oriented solely toward the occurrence of a word, a particular word, and that word must be known to the questioner. Furthermore, the answer of the system can be only that the word occurred.

4. The requirement that the questioner be able to address the store of information directly is not relaxed by the introduction of an indexing system into the list of ideal characteristics. Consideration of the research now under way, however, does suggest that it is naïve to state the requirement so simply. The system should allow the questioner to address the computer directly, *and* it should not require him to describe that which he does not know. It should be enough to describe the periphery of the hole in his information, and then ask the system to fill in the blanks.

5. The research already accomplished suggests a new capability for the ideal system. There are advantages to semantic and syntactical analysis in the elucidation of meaning which should not be ignored. If we can achieve a significant portion of such analysis by automated means, we can retain those advantages while eliminating the disadvantages stemming from the subjectivity of analysis performed by humans.

V

The Plan for the ABF-IBM Pilot Experiment in Automatic Organization and Retrieval of Legal Literature

The American Bar Foundation and the International Business Machines Corporation are engaged in a joint investigation of the application of electronic information processing equipment to the handling of legal materials. This study represents a part of a larger ABF project, which deals with an analysis of legal research methods and materials in general.

The emphasis in the study is on the method of organization of material, although obviously there are many other facets that will deserve attention at some point along the way. In particular, we are trying to design a system that combines the merits of non-classification with those of classification, in an "untouched-by-human-hand" procedure.

The overall plan for the pilot experiment is to convert raw text to machine-readable form, prepare a thesaurus (index-word space, or the "association map") automatically, using half the raw material, and to index the other half automatically with reference to index-word space in such a way that volumes of index-word space representing concepts will denote each document. The documents then will be stored in a modified inverted file, in which the heading is an expression of the concept volume rather than a keyword or descriptor, and then searching will take place in an obvious manner. For the mood of the experiment we are indebted not only to the work of other groups dealing with legal materials, summarized in section three, but also to the ideas of H. P. Luhn,[21] M. E. Maron,[22] H. E. Stiles,[23] L. B. Doyle,[24] H. Borko,[25] and F. B. Baker.[26] However, our procedures will be different in detail from any of theirs. Since our plan is based in large part on unproved theory, hope, and speculation, we reserve for ourselves a disclaimer to the effect that we will feel free to switch horses at any point in the stream if it seems advisable in the light of later developments.

The steps specifically planned for the experiment are the following:

1. *Conversion of text to machine-readable form.* Text for approximately 5,800 cases taken chronologically from the Northeastern Reporter will be converted manually to punched cards and thence to magnetic tape records. We are deliberately avoiding restricting the sample material to any one field of law; we are selecting cases not because we expect the system to be restricted specifically to cases but because we consider that type of material to be a sort of middle ground between documents such as statutes and miscellaneous expository legal writing.

2. *Machine construction of a thesaurus ("index-word space").* Word frequency counts will be obtained, considering any string of letters surrounded by two spaces or any mark of punctuation to be an English word. The frequencies will be used to prepare, for each unique word that appears in the full text of 2,500 cases, an estimate of the skewness of its distribution in the file. It is expected that words that convey information (in the sense that they characterize documents with respect to this file) will have skewed distribution, while those that convey little information will be distributed approximately normally. The extreme types of distribution are illustrated in the diagrams following.

[21] Luhn, *The Automatic Creation of Literature Abstracts,* 2 IBM J. RESEARCH & DEVELOP. 159 (1958).

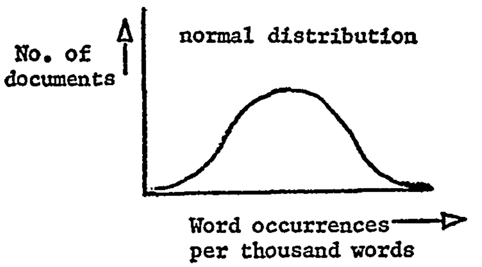
[22] Maron, *Automatic Indexing: An Experimental Enquiry,* 8 J. ASS'N FOR COMPUTING MACHINERY 404-17 (1961); Maron & Kuhns, *On Relevance, Probabilistic Indexing, and Information Retrieval,* 7 *id.* 216-44 (1960).

[23] Stiles, *supra* note 18.

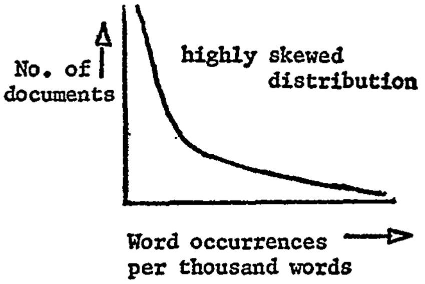
[24] Doyle, *Semantic Road Maps for Literature Searchers,* 8 *id.* 553 (1961); Doyle, *Indexing and Abstracting by Association,* 13 AM. DOCUMENTATION 378 (1962).

[25] Borko, *The Construction of an Empirically Based Mathematically Derived Classification System,* System Development Corp. Report SP-585 (mimeo.) (Oct. 1961).

[26] Baker, *Information Retrieval Based upon Latent Class Analysis,* 9 J. ASS'N FOR COMPUTING MACHINERY 512 (1962).



NONINFORMING WORD
(Type I)



INFORMING WORD
(Type II)

Style-characteristic words probably will occur in both distribution types, and we will not try to select them out. After visually inspecting a list of all the words ranked in order of skewness, we will divide the entire file of words into Type I (noninforming) and Type II (informing), based on some skewness cutoff point.

A dictionary on tape will be formed of the Type I words, and they will not be used further until the abstracting-indexing step.

For each document in the body of 2,500 cases, a list will be prepared of its Type II words, maintaining their original order within the document, but taking no account of internal borders such as sentences or paragraphs. For each Type II word an "association factor" will be calculated for every other Type II word with which it appears in any one document by computing the probability that Word A would appear this close to Word B this number of times over the entire file, if the Type II words were distributed at random. (This amounts to borrowing Stiles' idea of the association factor, but implementing it with a numerical method which takes into account nearness of the words within the document as well as the fact that they both occur in the same document.) Since the factors are probabilities, they will be numbers between zero and one. Small probabilities should indicate a high degree of association and high probabilities a lesser degree. These numbers will be used to estimate the distances between words in index-word space.

The next step is to construct from the information about distances between pairs of words an index-word space in which every word is at the correct (or approximately correct) distance from every other word in the system with which it exhibits association. The result of this operation can be visualized schematically as a sort of grid in which every word can be placed in its appropriate position by assigning it a set of coordinates. We will do this in such a way that the set of coordinates for any individual word consists of a string of six digits. (We know that John Horty found a vocabulary of the order of 20,000 significant words in his work on the statutes, and so six digits should be more than ample to assign a unique code to every informing word in our system.)

A Type II word dictionary on tape will then be prepared, the argument being the English word and the function being its 6-digit set of coordinates ("definition") in index-word space.

Physically, the tape dictionaries of the Type I and Type II words will constitute the machine-constructed thesaurus. The Type II dictionary will also be the computer's "picture" of index-word space, in which (we hope) words that appear geometrically close together will be closely associated in some way, and words that appear far apart will be not closely associated. Therefore, concepts can be represented by volumes carved out in index-word space, and these volumes may be said to correspond to the descriptors or index terms that might have been attached by humans, without their arbitrariness of classification.

The carrying out of this analysis will be a rather sizable computer job. For example, estimating that each word might be associated with fifty other words, on the average, 20,000 Type II words would require about 500,000 probabilities. The reduction of these to index-word space may have to be carried out by segmenting the total problem. On the practical side (although we are not trying at this point to be practical), it is not too costly to have to perform a horrendously large job, if it does not have to be done very often. Building of the machine thesaurus corresponds to overhauling a classification system in a humanly operated system, and of course this is a costly task, also.

3. *Abstracting and indexing.* Indexing of the remaining cases in the experiment will be performed by machine from full text, using the Type I list to discard words and the Type II list to prepare an analysis of frequencies related to index-word space. Instead of selecting specific words as indexing terms, concepts will be selected (statistically) as volumes in index-word space. A rough physical analogy to this process would be to toss pennies at the previously mentioned grid so that, for every Type II word in the source document, a penny lands at its proper slot on the grid. Where the pennies heap up in a pile, you have a concept. The circumference of the base of the pile defines the scope of the concept.

Document numbers then will be filed in document space, where the grid matches index-word space one-for-one. The document file will be maintained on tape in a form analogous to that of the customary inverted file.

A second tape document file will be maintained in order of document number, with the function for each document number consisting of the complete list of Type II words that appeared in that document; this will serve as a rough abstract at the time of search.

4. *Searching.* Searching will be carried out essentially by indexing a question presented narratively, determining the concept volumes that represent the question, and searching those volumes in document space for the relevant document numbers. Since the "edges" of the concept volumes are determined statistically, output can be listed in order of probable relevance; as an option the question could be accompanied by a request that "at least 100 references be supplied," in which case the

concept boundaries would be adjusted to provide that number. The rough abstracts contained in the second document file will be listed as an aid to the human searcher, who may then wish to eliminate some of the cites.

5. *Reorganization.* In the course of the indexing process, frequency information necessary for revision of the original thesaurus will be collected. Reorganization then will consist of recalculating the coordinates of the Type II words in index-word space.

A conceptual analogy can be drawn between the proposed system and a learning human. The system develops a sense of the meaning of words by "reading" extensively. It cannot know anything about a word that it has seen only once, but as it experiences the word repeatedly in different contexts, it begins to "catch on" to its meaning (sometimes erroneously). The more it reads, and the better the quality of the material it reads, the better informed will be the system. It will become an expert in its field, although it may prove quite ignorant when first turned loose in another's field! When this manner of ignorance becomes a significant lack, the system can be advised to read in another area to close the gap.

When the material and the system are ready, questions will be solicited from lawyers and scholars in representative practice and research. Each questioner will be asked to perform a manual search on his question, keeping careful time records. Half of the questioners will be asked to limit their manual searches to the same 5,800 cases in the Northeastern Reporter which the computer will be searching. The other half will be allowed to use anything in the entire Northeastern Reporter. In addition to providing controlled tests for the system, it is hoped that this experiment will cast some light upon the problem of our backlog. The dispositive adequacy of answers based on a complete search of a portion of the material will be measured against an incomplete search of all the material. If the former technique is equal to or better than the latter, we have then only to determine the optimum cutoff point. Perhaps we will thereby vindicate Justice Holmes' remark that "the reports of a given jurisdiction in the course of a generation take up pretty much the whole body of the law, and restate it from the present point of view."[27]

CONCLUSION

The problems raised in this paper, and other problems raised by implication, permit no easy resolution. The difficulties are as manifold and as complex as our society and our language. Having recognized this fact, it is important not to be daunted by it. Greater difficulties have been resolved in the past, and the resolutions were accomplished with lesser means than those available today. The work described here bespeaks a base of determined exploration upon which the solution may be formed.

[27] Holmes, *The Path of the Law,* 10 HARV. L. REV. 457, 458 (1897).