Legal Information Retrieval by Computer: Applications and Implications

Colin Tapper *

It is about fifteen years since Professor Horty and his team at the University of Pittsburgh decided that computers should be enlisted to keep track of the great and constant expansion of legal information. During these years the computer has been applied to modern statutes and to ancient cases; by universities, by professional bodies, by governments and by entrepreneurs; and everywhere from Austria to Australia. Every combination of keywords, abstracts, indices and original text has been tried. They have been stored in a variety of formats for searching in any number of ways to yield output in a myriad of different forms. Enormous sums of money have been spent and aeons of man-years consumed in the creation of notations, languages and programmes. Even in this relatively short period, it can already be said that no one can hope to be fully informed about the developments which have taken place. Certainly no one article can purport to delineate their major features. But the time is appropriate to draw back, to consider what has been achieved, to evaluate it, and to ask what should be done next.

I APPLICATIONS

The last fifteen years have witnessed a flourishing of computer techniques in their application to the law. A number of interesting experiments have been conducted and a few practical systems have been brought into operation. It is convenient to deal with each one of these three elements separately, bearing in mind that they do not exist in self-contained compartments since all operational systems necessarily apply some techniques and most have been tested experimentally.

(a) Techniques

Fifteen years ago computers were very limited in capacity and very slow in operation. Input was usually by punched card, no auxiliary storage was available, and output was usually by line

* Fellow and Tutor in Law, Magdalen College, Oxford.

printer. The computer would accept linguistic information only in upper case, and could store it only serially. It was therefore remarkably far-sighted of Professor Horty to think of using these techniques for handling the very large volume of data contained in the Pennsylvania statutes, particularly since he proposed to input the full text of those statutes, and not merely an index to them. In a sense this decision was forced upon him since his basic goal was to identify insufficiently indexed sections of the statutes which dealt with health law with a view to compiling a manual on that subject. The technique which he pioneered involved the preparation of the full text of all the statutes in their original language, and the creation within the computer of a complete concordance of all, except a few common words. The retrieval system then relied upon the detection of the occurrence and conjunction of particular words in the text to identify those parts which related to the required subject. Because of the technical limitations of the computers available to Horty, it was most economic to aggregate a number of searches together and to perform them all at one time, a technique known as batch processing. It was a remarkable achievement to make so many innovations in the techniques for finding legal information simultaneously, and it is hardly surprising that it took some time for them to become fully understood and accepted. It is only now that significant variations of method are beginning to be proposed. One reason for this delay is paradoxically that the pace of technical improvement has been increasing. As a result, a great deal of the available expertise has been devoted simply to adapting Horty's techniques in the light of these technical innovations. In a very short period of time, one has had to learn how to deal with optical character recoguition for input, random access auxiliary stores and terminal systems permitting direct interrogation of the system by the user. Optical character recoguition is a technique which speeds up input by allowing the computer to accept material directly from the typed, if not printed, page. It should be noted here that the possibility of accepting information directly from the printed page without the intervention of a human transcriber will create problems of a new order of magnitude, and will be considered later. The availability of random access auxiliary stores was a thoroughly liberating advance since it meant that many operations which had previously cluttered the central store of the computer and thus limited its performance could be performed elsewhere in the system. This resulted in an enormous increase in both capacity and efficiency, although it is interesting to note that some now feel that the transition from serial to random access has not been

wholly advantageous.[1] The most obviously revolutionary of the new factors mentioned above is the facility for the use of terminals which, together with the availability of multi-user capabilities, has necessitated a complete reappraisal of searching techniques. In the more sophisticated systems, terminals now comprise keyboard, video screen for the display of information and an associated device for preparing hard copy. The great change (not entirely in accordance with Horty's original philosophy) which this brings about is to put the user into close contact with the progress of his search while it is still going on. This enables him to control and modify it in the light of his interim results.

While it is true that in the past fifteen years linguistic applications of computers have increased in all fields and that in consequence standard programmes for such applications have become increasingly available, one of the results of Horty's initiative has been that legal systems have been in the forefront of these developments and have had to bear much of the strain. This has led to a certain distortion in emphasis. More time has been spent on getting the computer programming right in legal information retrieval systems than on getting the legal techniques right, and this may well have been unfortunate.

Professor Horty's original application was to the Pennsylvania statutes. A further area in which effort has been made is in the adaptation of the techniques to the demands of other types of legal material and to its expression in other languages. A number of different workers soon began to apply these techniques to patent applications, to administrative rulings and above all to cases. The problem was to determine how far and how effectively these methods, depending as they do upon form, could be applied to legal materials expressed in forms different from that of American state statutes. Thus when an attempt was made in the early 1960's to test Horty's techniques on British public health statutes it was found that none of his test questions were satisfactory because of the difference in the level of generality in the two systems. This is even more true of non-common law systems where the functional balance between statutes, cases and other materials is quite different, and is naturally reflected in differences in the form of those sources. One general similarity, however, is that legal information is at least everywhere very largely linguistic, and it has thus generally been unnecessary to grapple with the

[1] See Thomas, "GIPSY — General Information Processing System for Oklahoma", in American Bar Association *Automated Legal Research* (1973), 119.

problems of computer graphics, although land registration systems and systems for dealing with treaties which often incorporate maps are on the verge of such problems. But even the restriction to language provides only limited relief in relation to languages like Hebrew which differ so radically in their structure from that of English.[2]

(b) Experiments

Although Professor Horty was originally conducting an essentially practical exercise, namely the production of a manual, his thesis of the possibility of computer application had to be tested experimentally. It was necessary to discover whether or not word occurrence was an adequate surrogate for meaning, and if possible to measure the divergence between results achieved by reference to both occurrence and meaning. In this regard Professor Horty was at a disadvantage in two ways, one practical and one theoretical. The practical disadvantage stemmed from the statutory nature of his material. It is notorious that statutes are extremely difficult to index, and thus the conventional yardsticks for the retrieval of the statutory sections relevant to a given problem were more than ordinarily defective. It was indeed this very phenomenon which had given the original impetus to the idea of using computers. It followed that it was extremely difficult to measure the results of the experiment accurately, and that such resnlts could not necessarily be extrapolated to other better indexed areas. It is also the case that searches for solely statutory material are rare in legal practice, and that any testing of a system simply by reference to such searches is unlikely to be typical. The theoretical disadvantage was that Professor Horty was then unaware of the techniques devised by Professor Cleverdon at Cranfield in Great Britain for the measurement of the performance of information retrieval systems. Cleverdon's method involved the definition of and distinction between the two concepts of recall and precision. Recall connotes the proportion of relevant documents retrieved to the total number of relevant documents in the collection searched; precision connotes the proportion of relevant documents retrieved to the total number of documents retrieved. It was Cleverdon's thesis that recall and precision tend to be in inverse relation to each other. Thus it is possible to maximize recall by retrieving every

---

[2] Fraenkel, "Full Text Document Retrieval", in *Proceedings of Association for Computing Machinery Symposium on Information Storage and Retrieval* (1971).

document in the collection since if all the documents are retrieved it follows that all the members of every sub-set, including all the members of the sub-set of relevant documents, must also be retrieved. But this would simultaneously maximize imprecision. Conversely imprecision can be minimized by retrieving no document at all, since again this would entail that no irrelevant document was retrieved. This strategy, however, would simultaneously minimize recall since no relevant documents would be retrieved either.

A further experimental problem is that the use of these measurement techniques assumes an objective assessment of relevance. Professor Horty attempted to solve this by graduating retrieved documents into one of three categories of relevance, but he does not indicate how far the objectivity of the ascription of documents to categories was tested. This particular problem came into prominence in one of the early attempts to apply Horty's techniques to case law in an experiment conducted by a joint team from the American Bar Association and International Business Machines Ltd. It was found that divergent judgments between different testers was so acute that it became quite impossible to establish sufficient agreement to give any statistical validity to the results.[3]

My own experimental work attempted to utilize the Cranfield concepts for case law while avoiding the pitfalls of artificial search formulation and the subjective assessment of relevance.[4] Case law is in fact uniquely advantageous in these respects. It is possible to make good use of its characteristic feature of internal citation. Judgments normally cite other judgments which are relevant and patterns of citation can be traced easily in special indices. It is thus possible in relation to any segment of past case law to establish a list of all later cases which cite a case comprised in that segment. The facts of such cases can then be used as problems for the purpose of the experiment. This procedure has two advantages. First, since the test represents an actual decided case it is clear that the problem is real and not just an academic exercise; if enough cases are aggregated, the chances of the results being extrapolated into case law generally are high. The second advantage

---

[3] Eldridge, "An appraisal of a Case Law Retrieval Project", in *Proceedings of Computers and Law Conference at Queen's University* (1968).

[4] Tapper, *Feasibility Study of the Retrieval of Legal Information from Two Types of Natural Language Text*, Office for Scientific and Technical Information Research, Paper 5062 (1969).

is that the fact of citation itself represents an objective decision about the relevance of the cited case to the problem: it is made independently of the experiment by a professional judge in the course of his ordinary work. On the basis of these advantages it is possible to establish a firm low point on the precision scale (the proportion of such cited cases to the total number of cases retrieved) and a corresponding high point on the recall scale (the proportion of such retrieved cited cases to the total number of such cited cases). But an experiment based solely on such criteria would be somewhat arid since it would assume that the current practice of citation was not only accurate in its judgment of relevance (which is plausible), but also completely exhaustive (which is not). It is thus necessary to court the dangers of a subjective assessment of relevance. In my own work this was avoided by adopting a threefold distinction into cases relevant without doubt, cases irrelevant without doubt, and all the rest. This procedure was designed for speed, and by eliminating all doubtful cases, for accuracy where a positive judgment was expressed.[5]

(c) Operational systems

Professor Horty's system was always intended to be a working one, and after a decade of development it became the basis of a series of services offered by Aspen Systems Corporation, now a subsidiary of the American Can Company. Although these services have diversifed over the years, their heart has always lain in the exploitation of the original data base, statute law. It became apparent in the course of collecting state statutes in order to build up the data base that there were significant difficulties and delays within the various states in the compilation and up-dating of their statute books. It was also discovered that many state statutes contained anomalies and long-disregarded provisions. This inspired Aspen to offer more comprehensive services involving the collection and collation of state statutes. This made good commercial sense in the United States where state governments were substantial and creditworthy customers, and ones who might well provide further markets for statute-based services. Of course in a free enterprise economy Aspen was not long alone in the field, and rival services began to be provided, notably by the Data Retrieval Corporation of America, and by the state governments themselves, the latter often at the prompting of salesmen from International

[5] For a fuller account, see Tapper, *Computers and the Law* (1973), ch. 6.

Business Machines Ltd., which helped a number of the pioneer states.

Under the stimulation of this competition the orientation of the services changed from computer based legal information retrieval to computer based preparation and printing of state legislation, though information retrieval has never been completely forgotten.[6]

The advantage of sticking to statutes was that the data base was relatively small compared to the bulk of case law and other secondary sources such as encyclopaedias, books, articles and opinions. Thus far the problem of the accurate transcription of voluminous archival material has constituted the principal stumbling block to the transference of Horty's full text techniques to case law in a commercial environment. An obvious alternative is to use something less than the full text for case law. It can indeed be argued that statute law is unique in the austerity and authority of its language. Case law is relatively more discursive and diffuse and much more difficult to abridge either into index terms or an abstract. Index terms and abstracts do indeed regularly appear as part of the published version of a case in most series of reports, and index terms do operate as search tools within the conventional system. In the United States and in some parts of other legal systems [7] such shortened versions are linked into formal hierarchical structures like the West Publishing Company's key number system. In other codified systems of law it is possible to employ the organization of the code to supply a structure for other primary materials. The obvious tactic in these circumstances is to use such a structure or scheme of indexing as the data base for the computerization of case law. Such an approach has indeed been adopted by such various organizations as Law Research Services Incorporated of New York, and CREDOC in Belgium. Where an index is interposed between the user and the material there is less need for continuous control over the search by the user, but more need to ensure that the search terms match the search needs. These considerations lead naturally to the use of batch processing systems with the formulation of the search reviewed, if not undertaken, by the organizations responsible for devising the system.

The relative decline in the cost of handling and storing large quantities of information has, however, led others to keep to Horty's

---

[6] For a full survey, see Elkins, *Survey of the Use of Electronic Data Processing by State Legislatures* 2d ed. (1971).

[7] The maxims collected by the Supreme Court of Cassation in Italy, for example.

original philosophy, and to prepare systems for handling case law in full text. Such systems are probably furthest advanced in North America.[8] The largest and most widely available to ordinary lawyers is the LEXIS (formerly OBAR) [9] system in the states of Ohio and New York. This system incorporates a full text case law data base, permits freely structured searching in ordinary language, displays results upon a video screen situated in the lawyer's office and allows any amount of modification to the search while it is being conducted. Technically it is a very smooth running and sophisticated system. Lawyers, however, have been slow to take to it, even in Ohio where it was devised and where it has been on offer longest. A possible explanation is that the system does not yet do enough of the elementary thinking for the lawyer. For example, there is as yet no automatic thesaurus to suggest synonyms, antonyms, particularizations, generalizations, or grammatical or orthographical variations of the chosen search terms. Further, the search logic is restricted to Boolean operators supplemented by qualifiers for distance and direction of conjunction. Other systems have been designed to overcome some or all of these defects; but so far none has been offered long enough or widely enough to practising lawyers for it to be possible to determine how much more popularity they will inspire.

It seems then that in the last fifteen years much work has been done but relatively little has so far filtered through to the practising lawyer. The original techniques devised by Professor Horty have been improved and adapted, but not abandoned. Statutes are in some sense being handled by computers in a practical environment, but case law is not. In the second part of this article some implications to be drawn from this state of affairs will be canvassed and possible new directions for further research suggested.

II IMPLICATIONS

It is arguable that the lack of response by the practising profession to computers for the retrieval of legal information, which is indicated as much by its failure to develop better systems

[8] While the STATUS system developed by Dr Niblett at the Culham Laboratory in Great Britain is used for statutes, it is quite capable of handling case law and exhibits many of the features of operational North American systems.

[9] This system is described on the basis of the latest published material but it is possible that it has by now developed beyond the stage described in the text.

as by its failure to use those which are already provided, shows that there is no need for such facilities to be provided. After all, despite the direst forecasts and gloomiest projections, the legal profession still does cope with the increase in material and there is no obvious decline in its standards. It may be the case that relatively few practising lawyers do enough research to justify investment in a computer orientated service.[10] It might indeed make more sense to concentrate on the use of the computer to assist more directly in the solution of legal problems, but that is outside the scope of this article. But even if consideration is limited to legal information retrieval, there is a further stumbling block which may have contributed to the difficulty so far experienced in getting commercial services off the ground. It is customary to think in terms of *searches* for legal *information.* This is a vague and misleading formulation. To start with, most systems provide references to where information is to be found rather than information itself, but this is a quibble. Of much greater importance is the assumption that searches all share a common character and that the sort of information which will satisfy one will also satisfy another. This is fundamentally implausible. It has already been suggested that typical searches in statute law differ from typical searches in case law, and that academic searchers are often looking for something different from practitioners. Further, each different sort of lawyer at each different stage in his work in each different class of case in each different system of law will need to search in a different way. Thus there is not one common pattern of search but an almost infinite number. It is not obvious that they all have enough in common to enable any one system to satisfy them all. For example, the tax adviser who wants to know whether a particular concession is still being allowed by the revenue hardly requires the same sort of information as an academic lawyer interested in discovering whether there are any analogies to the requirement of the identity of the mental capacity required for making a contract and consenting to the dissolution of a marriage in other branches of the law. Again one sort of search may require an answer in terms of the current law of a particular jurisdiction, another may want only the law at some particular time in the past, or at some other particular place. One search might require no more than citations, another an indication of the volume of materials, another an indication of their contents, and another suggestions

[10] See *Operation Compulex,* Department of Justice, Ottawa (1972), 25, 26, app. 2, chart C.

on how to proceed. It is perhaps conceivable that systems of such generality and flexibility could be devised to supply information in many forms to a variety of searchers. But such a system would be incredibly wasteful of resources unless need was spread evenly and at random across the whole spectrum of possibilities, which is intrinsically unlikely. It is clearly necessary first to discriminate between the different sorts of searches which are made, and then to discover how much need there is for each of them. This is not, however, an easy task since current practice will not necessarily give an accurate answer. It will itself be influenced by the efficiency of currently available aids. Need may be better measured in economic terms, but even this is not an easy task when the comparison must take in lawyers in private practice, lawyers employed by the government or commercial organizations, members of the legislature, voluntary advisers, legal academics, students and private citizens. But however difficult the task may be, it is clear that it must not be shirked. We are at present in the era of trial and error at random, but the sums of money and resources involved are becoming so great that this cannot continue long.

In what follows it will be assumed that some such appraisal of need has been conducted, and that some legal information retrieval system by computer has been found to be desirable. In this event it will be necessary to improve the performance of the systems currently being offered in various ways and to understand the consequences which will follow from the availability of such improved facilities. These two elements will be considered separately.

(a) Improvements

Technical improvements will continue to be made, programmes will become more efficient and both will occur in ways which are partly the continuation of current trends, and partly unpredictable innovations. Possible improvements in data capture, searching and testing procedures will be discussed here.

(i) *Data capture*

As stated earlier, data capture has continued to constitute the most serious financial and time consuming obstacle to the development of large scale legal information retrieval systems. There are really two separate problems: the preparation of archival material and the assimilation of new material. So far as archival

material is concerned, the difficulty has resided in the inability of optical character recognition devices to deal with the wide variety of type faces to be found in the conventional law book. It is likely that these difficulties will one day be overcome. However, that will not bring all problems to an end; it will simply mean that a different set will have to be faced. At present transcription is carried out by men who apply their minds to the task and who are able to follow simple instructions. If they are asked to insert special characters at the end of each document they are capable of doing so, and even without special instruction they will be able to distinguish between hyphenation inserted by the printer to help justify the right hand margin, and ordinarily hyphenated words. The transcription device and the computer will be unable to accomplish either. All additional symbols will have to be specified and inserted at some stage in the transcription process. Similarly very precise instructions will have to be devised to accomplish any necessary deletions or modifications of the text, such as the relation of footnotes to the point in the text referred to. Further, in documents which depend for their meaning on a special format, such as schedules of repeals in statutes, or which contain tabular material or diagrams, some method of treatment must be prescribed. Finally, it will still be necessary to check the accuracy of the text as transcribed since even if it is assumed that the device will be completely accurate, it will still reproduce any typesetting mistakes in the original.

Problems with new materials are likely to be less severe, and are in some ways the converse of those described above. The danger here is that since it is contemplated that materials will be prepared both for printing in the ordinary way and for the computer input, making the input easily usable may militate to the disadvantage of the published version. Thus one might restrict the number of different typefaces to be used, alter the position of footnotes, or limit the use of tabular or diagrammatic copy. Still more unfortunate would be any tendency to restrict the range of terminology used to express the law or the prescription of particular ways of formulating it.

It should be possible to devise techniques which will both present the document in the way desired by author and reader to the extent that these are themselves compatible, while at the same time making any modification necessary for input to the computer. Since most of this information will be prepared by some form of direct entry, it will be necessary to pursue research into the most suitable form of terminal design, text editing devices and input software,

none of which is yet in a state suitable for the really sophisticated operations which will become necessary when a brute machine replaces an intelligent human being at the initial stage of the process.

(ii) *Searching techniques*

Searching techniques are at present largely geared to Boolean combinations of words occurring in natural language, with some qualification by reference to separation in numerical and logical terms. Thus a typical search will ask for references to documents which contain any one of words A, B or C, followed within the same sentence by any one of the words D, E or F, but not if any one of them is itself immediately succeeded in the same sentence by any one of the words G, H or I. It may well be the case that the need to cast their thoughts into so rebarbative a form is a prime cause for the revulsion and difficulty experienced by practising lawyers in using such systems. This has led a number of workers to suggest different forms which a computerized search could take. One which is currently available in some systems, but not universally and rarely in combination with other forms, makes use of citations. This is very familiar to lawyers who have been using citations to find references for centuries. It is true that this facility is therefore already well provided for by conventional means in some countries, and notably in the United States. But computerization would make such indices much easier to use both physically and psychologically: physically because there would then be no need to turn pages and change volumes at each level of the citation tracing process; and psychologically because the user could be certain that all entries were completely up to date. Even here, however, some changes in existing systems would be desirable to permit greater flexibility in the storage, searching and display of citations than exists at present. For example, second and third order citations could be retrieved and held in a buffer store for release upon a first order citation being indicated with a light pen or cursor, with the other first order citations themselves buffered and held ready for re-display while this took place. Similarly patterns of citations could be matched by vectoring techniques in order to display citations having comparable patterns together.

It would also be possible to mix citations of statutes and cases in any such system, thus making it possible to collect all the cases in which a particular statute or part of a statute was cited. These possibilities raise a more general question in relation to legal information retrieval. It is sometimes assumed that retrieval

operating on the full text of a document is necessarily capable of conveying all possible information about it. This can easily be shown to be false since a document's subsequent history, for example, is often of vital importance in understanding its significance and this can clearly never be found in the original text. Similarly it is often useful to add to the text a large variety of other pieces of information, some of them informing the user, and others providing possible parameters within which a search might be confined. Here too, much more research needs to be conducted into the nature of this auxiliary information in relation to different types of legal documents. It is likely that future systems will concentrate more and more on the use of terminals and either multi-access large centralized systems with remote communication or widely disseminated mini-computers with large secondary stores. In either case it would be highly advantageous to sub-divide the data base into smaller units which would normally be searched independently, though one would have to provide the facility of combining such sub-units together. This will make for a much faster and cheaper system, but its introduction must be preceded by much research in order to establish the optimum pattern for the sub-division.

The next area in which experiment is needed is that of search formulation. At present it is quite common to permit only the use of truncation techniques to assist the user in formulating his search. He is necessarily left to his own devices so far as choosing and combining terms is concerned. His most obviously useful aid is an adequate thesaurus. This will permit the user to expand his search terms so as to minimize the danger of under-recall because he has not thought of the synonym or circumlocution in fact used by a relevant document instead of the term guessed at by the user in formulating his search. The construction of such a thesaurus will be by no means an easy matter. It will obviously be needed to cater for synonyms, antonyms, particularizations and generalizations. However, all of these notions depend upon context. In some contexts "will" is a synonym of "stall", in others of "deed" and in others of "determination". "Light" may be an antonym of "dark" or of "heavy". In some contexts "shooting" is a particularization of "murder", in others of "causing an explosion". In some contexts "negligence" is a generalization of "overlook", in others of "fraud". It is extremely unlikely that any automatic procedure alone will ever be able to produce an adequate thesaurus in the light of these considerations. It will require the very greatest amount of ingenuity to compile one at all.

The converse problem is that of providing assistance to the user to help guard against the danger of under-precision. The most common device at present is to indicate the number of documents which will be retrieved on the basis of a particular search formulation. If the number is unacceptably high the user can modify the search until he has brought it down to acceptable proportions. This is useful, but by itself inadequate. It would be better if the documents were put into an order of prospective relevance so that the user could then examine documents at particular points on the list to secure some indication of what was going wrong. One way of doing this, adopted by Professor Lawford in the Quic Law system [11] and by International Business Machines Ltd. in its STAIRS system, is to allow the user to choose from among a number of algorithms depending upon permutations of the frequency of occurrence of a required term in a document and in the total set of documents, and the number of documents in the total set in which the required term appears. It is not clear that such algorithms are necessarily satisfactory, and a scientific study of the efficacy of these and others in the context of legal information retrieval would certainly be valuable. A similar result can be achieved by the mathematically more sophisticated technique of plotting an n-dimensional vector for each document on the basis of the word occurrence and importance, and then retrieving documents in terms of the degree of closeness of match to a similar vector prepared for the terms of the question.[12] It has not, however, been demonstrated yet that such a system would be applicable to documents of so varied and complex a character as that exhibited by the different series of case reports in full text, but such systems are clearly worthy of further exploration.

One final aid which might well benefit from further consideration is the provision of cueing assistance to the user during the course of his search. It would be very helpful if he could not only be given the number of documents which he would retrieve on the basis of the current search formulation, but also told how that number would be affected by modification in indicated ways. Such a system might be programmed to suggest possible deletions or additions, or the variation of particular operators, qualifiers or parameters. Further, it would be possible to make more general

---

[11] See Lawford, "QUIC/LAW", in American Bar Association *Automated Legal Research* (1973), 67.

[12] For an account of such a system see Vischer, "Das Dokumentationssystem der Unidata A.G." in *Kybernetik-Datenverarbeitung-Recht* Folge 1 (1971), 89.

advice available for display if the user were to request it, or if the searcher kept going wrong for reasons which the computer could recognize, such as repeated over-particularity leading to under-recall. Here as elsewhere there is much scope for further research into the sort of assistance which would assist potential users.

(iii) *Testing procedures*

There are at present no agreed testing procedures for legal information retrieval systems. This makes it extremely difficult to compare results which have been secured by different research teams in different parts of the world. The two concepts of precision and recall are by now well established, but they are not entirely satisfactory, partly because they are orientated towards one particular sort of search, namely that in which information is either relevant or irrelevant, rather than one in which it is more or less useful. They depend also upon it being possible to predicate with certainty of every document in the collection whether or not it is relevant, and this can usually be true only of relatively small data bases in relation to which experimentation runs the danger of being unreliable, statistically unsound or both. Neither are there agreed standards for counting results. Thus if one researcher finds all nine out of nine relevant cases in four searches, and one out of four in the fifth he might aggregate his figures and state that his recall is 90%, while another might calculate each search independently and then average out his recall as 85%. This sort of consideration is usually left unexplained in the presentation of experimental results.

In some obvious respects in which systems (especially those intended for commercial operation) should be tested, such as the time and expense of searching, there seems to have been no controlled testing at all. It is true that these factors may be difficult to measure, but that is no excuse for abstaining from any attempt to do so.

(b) Consequences

What difference would it make if all these steps were taken and a satisfactory system of legal information retrieval by computer available? Some might say that the legal system would collapse, choked to death by authority, but the concept of a *satisfactory* system is intended to gnard against this riposte. In fact the changes are quite unimaginable *in toto,* and all that is possible

here is a very brief indication of the possible impact upon legal education and the organization of the legal profession with some international implications.

Legal education would clearly have to change to accommodate the new technology. This does not mean that lawyers would have to understand computers; they would merely have to understand how to use them. It is also unlikely that a new system will be introduced for legal information retrieval before computers become much more a part of everyday life than they are today. This will probably mean that the basic principles of the computer, including elementary programming and keyboard operation, will be taught regularly in schools. This assumes that voice input will not by then have developed sufficiently to have made keyboard operation obsolete. It is quite possible that the necessity for direct keyboard entry by the legal practitioner is one of the things inhibiting the spread of conversational systems at present. Thus it is likely that in the future prospective lawyers will already be equipped with the technical and manual skills necessary for effective use of computerized systems. This will leave the law schools of the future free to concentrate upon the specifically legal side of such systems, which is just as it should be and is not at present. It is further envisaged that law schools will not need to establish special courses in the use of both systems any more than at present they have special courses in the use of indices, books and libraries. Rather, the impact of the new technology will influence the teaching of substantive law. Students will find their attention directed much more towards the ways in which rules and concepts are expressed, and will approach meaning much more from that point of view than they do at present. For those brought up in a tradition of the importance of the analysis of usage this will seem no bad thing. It is also, of course, likely that a secondary change will take place in the usage itself under the pressure of such close attention. Such emphasis will inevitably spill over into the world of legal publishing. The student book market will have to change to accommodate the greater stress on modes of expression, and new sorts of books suggesting possible search patterns will appear, although most of the change will lie in the incorporation of this change of emphasis into existing publications.

Another factor which probably plays a large part in inhibiting the acceptance of systems at present is that of cost. Quite apart from the cost of acquiring the data base there are expenses associated with securing access to a computer and with using it once access has been obtained. The legal profession is not ideally organized

to absorb these costs since it is mainly split into small units.[13] This means either that firms must subscribe to a general service, combine together into larger units for such a purpose or avail themselves of the mini-computers which are becoming increasingly common.[14] The other factor in the current structure of the legal profession which may be inimical to the development of retrieval services by computer is the distinction between solicitors and bar in many jurisdictions, either *de jure* or *de facto*. Where this distinction exists it exacerbates the division between the different types of search which may be required, and in such jurisdictions it is most unlikely that the same service would satisfy both branches of the profession. Similar considerations also apply to other segments of the legal profession, such as those in government (local and private), industry and the universities.

Finally something should be said of international implications. As the number of supra-national bodies multiplies, as treaties between states increase and as private citizens increasingly trade and travel in foreign countries, so the demand for knowledge of other legal systems will increase. It is unlikely that any system designed only to retrieve foreign or international law will be viable except perhaps in a few highly specialized areas. The most that can be done therefore is to ensure that so far as possible data prepared for national systems is compatible with that prepared for other systems. This requirement should not in any way inhibit the development of national systems, but there are many features of a computer system which are entirely arbitrary in the sense that they can be accomplished as easily one way as another. In such cases there is an obvious advantage in as many different systems as possible doing the same. Some work is already being done on this problem in Europe under the auspices of the Council of Europe, which in 1969 set up a committee to consider the harmonization of legal computer systems in member states. It would be helpful if this sort of approach could be widened, perhaps as a result of an initiative by an agency of the United Nations.

SUMMARY

This article has attempted to describe the progress made in legal information by computer during the past fifteen years, and to

---

[13] For the precise distribution in Canada see *Operation Compulex*, Department of Justice, Ottawa (1972), app. 1, graph C.

[14] See Hoffman, *Survey of Law Firm Computer Use — 1971*, (1971) Jurimetrics Journal 42, 86 for an account of these tendencies in the United States.

show how the pattern of that growth was largely determined by decisions made at the University of Pittsburgh in response to a rather special set of problems. It has been suggested that many of the difficulties in commercial application currently being experienced stem from failure to look hard or radically enough at the nature of legal searching. Some attempt has also been made to indicate directions along which thought might now be channelled in the hope of realising the benefits which computers may bestow upon the law.