LEGAL TEXT RETRIEVAL SYSTEMS
THE UNSATISFACTORY STATE OF THE ART

Jon Bing

The 1985 EW Turner Memorial Lecture
given at the University of Tasmania, Hobart
22 May, 1985
on invitation from the Faculty of Law

1. The Dawn

Thirty years ago, at the Graduate School of Public Health at the University of Pittsburgh, professors John Horty and William B. Kehl started a project designed to study and improve the health statutes of the state of Pennsylvania. At approximately the same time, the University established a Data Processing and Computing Centre.

A special assignment proved to be a kind of turning point. A state legislator in Pennsylvania had a bill passed to change the phrase "retarded child" to the slightly less stigmatic phrase "exceptional child". In order to implement the bill, all locations where the phrase occurred had to be identified.

Professor Horty started out to solve this problem in the traditional way; he paid a group of students to read through the statutes and regulations, and make a note of all occurrences of the relevant phrases. It turned out that the inaccuracy was too high to be acceptable - and another group of students was hired to reread the material. Still there were errors.

A more radical method was adopted. The entire material was registered on punched cards and verified by doublepunch. When a machine-readable copy of the material was established, it became quite a trivial task to read through the material and retrieve all occurrences where the word "retarded" preceded the word "child" or variations of "child".

The result was not only a satisfactory solution to the original assignment. As a by-product, Horty got the full text of the statutes in machine-readable form. And Horty found other and more exciting ways of exploiting his material. Actually, this was the beginning of text retrieval systems which today are predominant in computerized legal information retrieval.

The first successful demonstration of a text retrieval system took place at a conference organized by the American Bar Association in 1960. Professor Horty left the university and founded the Aspen Corporation, which during the 1960s helped to launch a number of legal information systems based on the new computer technology. One of the first ventures involved the design of a system for the Staff Judge Advocate of the US Air Force at Denver, Colorado. This system was given the name LITE, as an acronym for "Legal Information Through Electronics", and was presented with an inventable flourish: "Let there be LITE!".

The system was renamed FLITE in 1975 in order to emphasize its federal responsibilities. It is still in existence, though mainly a batch-oriented system, and it plays an important part through its co-operation with the JURIS system of the US Department of Justice.

FLITE is the oldest system in existence, and it is almost a symbol of legal information retrieval: the paradox of success and resistance to new ideas.

2. The Current State of Affairs

There have been very few changes in the retrieval software or principles since Horty launched his venture. The systems are still based on an inverted file (though there have been developed more efficient file structures than those originally employed), and the retrieval is based on Boolean logic. The one great innovation originated ontside the legal applications: the introduction of on-line systems which made terminal sessions with a dialogue possible. Today, the fast feedback from the system, and the possibility of rephrasing the search request immediately are seen as essential features in a text retrieval environment.

With this major exception, text retrieval systems are very much the same as when introduced. The development of information systems making large volumes of natural language texts retrievable, has to a large extent been fueled by the legal applications. It is quite interesting to note that lawyers have been the cause of a technological development of no small scale - though lawyers traditionally are not seen as technological avant-gardists.

The result is that today most industrialized countries have some form of legal information retrieval service: See Table 1.

Table 1 - Major legal information services

Australia
CLIRS
SCALE
Belgium
CREDOC
JUSTEL
Brazil
PRODASEN
Canada
QL-SYSTEMS
Denmark
LDB
European Communities
CELEX
Finland
FINLEX

France
CREDIJ
IRETIJ
JURIS-DATA
LEXIS
SYDONI
Germany
JURIS
DATEV LEXinform
Holland
JURID
PARAC
NLEX (with CREDOC)

Ireland
ITELIS (with EUROLEX)

Italy
ITALGIURE

Mexico
UNAM-JURE
Norway
LAWDATA
Sweden
RAETTSDATA
UK
(EUROLEX discounted June 1985)
US
JURIS
LEXIS
WESTLAW
New Zealand
LEXIS

In this paper, no attempt will be made to introduce the different systems (such a review may be found in Bing et al, 1984). But it may be interesting to confront the North American and European developments.

In the US, two major commercial services operate - LEXIS and WESTLAW. These share the market, and the public services - among which the JURIS service is the foremost - do not compete with the private services outside government. The service providers are left very much alone, without involvement of government or professional bodies in influencing their policies.

Also in Canada, the QL-system is a private organization. The QL-system is, however, mainly an intermediary serving information providers which maintain data bases offered by QL to the subscribers. The government is heavily involved in the policy of legal information services through the Canadian Legal Information Council - therefore Canada may be seen as a bridge from the US to the typical European situation.

In Europe, one will find that the initiative of creating services generally originated by a professional body or the government in contrast to a commercial organisation. The first European system, CREDOC, was actually created by the Belgian notaries.

But most striking is the role played by the government, generally represented by special agencies, and typically by courts with administrative jurisdictions. Many European jurisdictions include specialized administrative courts of some sort, and such courts share certain characteristics. The case load is often very high. The previous decisions of the court itself are an important legal source, but the in-house information system was often inefficient - based on manual files and indexing by clerks. The "ideals" of the courts are those associated with justice, the rule of law, equality before the law, etc. Typically, such courts found themselves in a situation where the goals could not be reached because their own precedent decisions were not readily available, more resources in the form of extra manpower were not forthcoming, and the case load could not be reduced by the court itself. In such a situation, it was not difficult to see the promise of a better information system as a solution.

Examples of initiatives being taken by or close to such courts are abundant: The ITALGIURE initiated by the Corte Suprema di Cassazione, the French CENIJ (formerly CEDIJ) related to the Conseil d'Etat, and JURIS in Germany started by co-operating with the Bundesfinanzhof and the Bundessozialgericht, both administrative courts. The Swedish and Finnish systems were both associated with the special administrative court systems in these jurisdictions.

But these were the initiatives. Today, publishers seem to play a more active role in the development of legal information retrieval systems. It has been found that the systems are not easily contained within the restraints of public budgets and managed through the bureaucratic channels of government. Because of the initiative, European systems often started as specialized and closed systems. But due to the dynamics working on the situation after the introduction of a system, the services are emerging as more general and more open systems. And in this process publishers play an important role. Also, services are often set up as independent organisations - the links with the parent organisation are severed. This has recently been seen in Italy, Germany, and Denmark - and the tendency is also clear in other jurisdictions.

There is clearly a trend towards establishing at least one national, general and open service within each jurisdiction, and to give this service an organizational status which allows it to operate independently of the government, but often being reviewed by the government or professional bodies taking an interest in the part the service will play in providing adequate legal information services.

The result is a number of market-oriented services, generally of a national nature. The great number of smaller or specialized services has been reduced, and the situation in Europe would seem currently to fall into a rather uniform pattern of national, general and services.

3. Two "clubs" of legal information services

The result of this development is at the moment appearing to create two "clubs" of legal information service providers.

One of these clubs may be dubbed the national or "gateway" club. It is composed of national services, which, for obvious reasons, have little interest in expanding their coverage into other jurisdictions. Their operation is generally integrated with government agencies for publishing the legal gazette, documentation centres in the courts etc. And the expansion of this mode of operation into other jurisdictions is simply not feasible. There is, however, a certain demand among the customers of a national service for information on the law in other countries. This may be due to a certain legal relation between the jurisdiction (as between common law countries or the joint statutory projects of the Nordic countries), to trade between the countries, or to the presence of foreign nationals. This need is then sought to be satisfied by making a bilateral agreement with another national service. Ideally, this will make one national centre a "super-user" of the services of another national centre, and the subscribers to the "super-user" become indirectly linked to that centre, and may access its services through a leased line or other form of gateway maintained by the "super-user".

Such a network has been seen to emerge, and includes the US WESTLAW service, the Canadian QL-system, the French SYDONI, the Italian ITALGIURE, the German DATEV, the European Communities' CELEX and the Australian CLIRS. The hub of this network was the UK EUROLEX service, which in June 1985 was acquired by LEXIS. The reaction to this move was not known at the time this paper was written (July 1985), but it obviously will leave a gap that probably will be filled by some initiative.

The other "club" has only one member, though this is a major provider of legal information services, and, indeed, other types of text retrieval based services, and is LEXIS, provided by Mead Data Central.

LEXIS started out as a US service, offering competition to the traditional legal publishers, especially to the major publisher West which only some years later created its own computerized service, WESTLAW. The first international connection was made to the UK legal publisher Butterworths, which interestingly is the UK counterpart to West. Therefore, LEXIS in the UK is perhaps more similar to WESTLAW than the original LEXIS service. Next LEXIS entered France, co-operating with the publishing house Hachette, which is an outsider as far as legal publishing is concerned, but which offers LEXIS through its subsidiary TI Consulte. And in 1985 LEXIS announced that New Zealand has been added to its coverage. Also, by its acquisition of EUROLEX, LEXIS was given control of libraries of Scottish law and the law of the Republic of Ireland, which had entered into co-operation with EUROLEX, offering a service known as ITELIS. It is still too early to see whether these jurisdictions will continue to be supported by LEXIS.

LEXIS is unique in its international ventures, but it may be pointed out that EUROLEX through its Irish venture set up a similar system, serving one jurisdiction from a system located in another. Also, the Belgian CREDOC service is serving Luxembourg, and co-operating with one of the Dutch providers,

Vermande, in offering services for these jurisdiction based on their Belgian system.

It may be maintained that national services are more impressed by certain policy concerns than LEXIS - for instance, concerns for national control of legal information services, for integration of the computerized services in a broader range of services where the traditional, paper based services may also be found,and for a general improvement of the legal information services of the country, including improvement for the small legal practitioner and public agencies.

Therefore, the bisecting of legal information services, which in a dramatic way confronts LEXIS with "the rest of the world", promises spectacular developments over the next few years. Obviously, the jurisdictions of Canada and Australia will be coveted prizes in this probable conflict, and it will be interesting to see which action the CLIRS, QL-system and WESTLAW will adopt in order to consolidate their position with respect to LEXIS.

4. Performance of legal information services

Though legal information retrieval services are widespread, there is a surprising lack of acceptance among practitioners. One generally finds that use is on the average quite low - rarely more than one hour/user/month, though variations are great. Also one finds that the majority of requests are formulated with just one or two search terms, such requests making up 80-90% of the total volume.

In a recent user study commissioned by the European Communities (Butler Cox 1985), the systems are not characterized as "user friendly", but rather as "user hostile". The users reported a lack of availability and the only service whose subscribers were satisfied was LEXIS (the study only embraced EEC countries, both France and the UK have LEXIS users).

The reason for the lack of acceptance is not disclosed by the study, but analysis implies that there are two main reasons:

The first is a lack of coverage. This has traditionally been a major cause for user dissatisfaction, but has been overcome with respect to the major services. Computerized services have today typically very comprehensive databases, though also typically case law material is not documented in the full historic depth, something which, however, does not seem to bother users too much.

The second cause is insufficient performance. This will be discussed further - and in order to do so, one may start by analysing what is meant by "high performance" in a legal information service.

Figure 2 - Levels of performance

"RULE OF LAW"

- objectivity - certainty
- equity - equality

FUNCTIONAL PERFORMANCE :			AVAILABILITY	
retrieval function	: relevance : assessment	: source : function	: formal : factors	: pragmatic : factors
recall : precision :				

At the first and uppermost level of performance, one will find the objectives generally associated with a legal system as a whole. A legal information system should be designed to support the same objectives. These objectives are most visible for persons or agencies charged with a general responsibility for the legal system of a country - typical representatives are ministries of justice, courts, etc.

Providers of services, however, have had a tendency to emphasise the bottom level of retrieval performance. There is no doubt that text retrieval is extremely efficient. Measured for instance in recall and precision, providers have argued the high performance of their system as a retrieval tool. It may be maintained that the retrieval function has been over-emphasised with respect to the other aspects of functional performance.

The user is, however, mainly concerned with the middle level of performance, and this may be described as two equally important aspects.

The first is functional performance, which may be broken down into three functions. (1) The ability of the system to retrieve relevant documents efficiently, which is the function traditionally emphasised by providers. (2) The properties of a document facilitating the rapid determination of relevance (for instance titles or abstracts). (3) The ability of the system to produce the source speedily, either through the computer communication facilities of the system or through a separate document delivery system.

It may be seen that the functional performance of a computerized system is enhanced in all these three respects compared to conventional systems. Therefore, the explanation for the lack of acceptance generally may be found in the second aspect: availability.

5. Availability

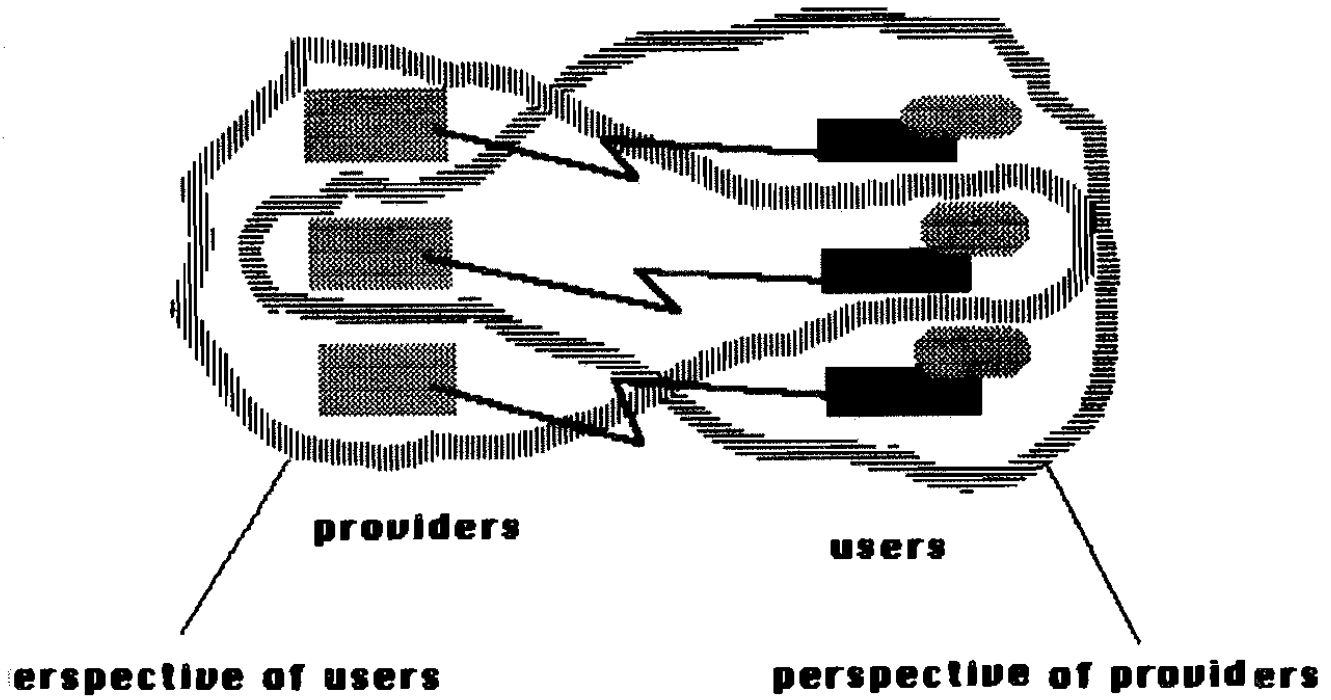
Definitions of information systems vary according to the context in which they are discussed. With respect to legal information systems, the definition of a "system" usually pivots on a provider of a service: a centre, a publisher or other organisation. In this way, it is usual to describe CLIRS, LEXIS or WESTLAW as systems.

From this perspective, a legal information system has one provider and a number of users or subscribers. Features of the system, such as database content or updating response and frequency, are quite well defined. The database of CLIRS at any given date is, for instance, the documents stored in its various text files.

But as for other system concepts, one may for different purposes amend the definition, adapting it to highlight other aspects in the relationship between the provider and the user of a service. An obvious alternative would be to let the definition pivot on the user rather than the provider. From this perspective, a legal information system has one user and a number of different providers.

This may be a perspective well suited to bring out some features of the information situation of the user, which is, of course, essential for an understanding of how legal information systems work within a jurisdiction.

Figure 3 - Different system concepts



From the perspective of the user, the provider offers information services of which he may take advantage - at a certain price. These services will be of different nature, from newsletters through journals and case reporters to monographs and, of course, computerized information systems.

In this complex situation, a number of factors may obstruct the user from obtaining the information which the user requires to address a certain legal problem. User research does give some indications of these causes, for instance the four major causes for missing information (cited from Jungjohann et al 1974, a user survey preceding the introduction of the JURIS test system in Germany):

Table 4 - Causes for missing information

Lack of time	33%
Delayed publication	21%
Missing form the library	13%
Delayed circulation	10%

Without discussing the details of this small (and ambiguous) Table - though it is tempting to point out that a common cause is the trivial fact that a source is simply missing from the library shelf - one should emphasise the major point: as much as one third of the causes are related to what the study calls "lack of time". "Time", like money, is a general way of measuring the resources at the disposal of the user. And this table only underpins what has already been stated: availability is the major problem in user-constructed information systems.

The use of any information system is associated with costs. This is obvious when the user subscribes to a journal or a computerized service; the user is then billed for the subscription fee. It is perhaps less obvious, but still quite evident, when the user browses through his own files, or looks up references in a compilation of statute law. In this case the cost is associated with the expenditure of time.

Some costs are associated with maintaining the user-constructed information system. These costs will be subscription fees, salaries to staff responsible for filing or categorizing material, costs for furnishing the library and renting space for it, costs of terminals, microform readers or other acquired equipment, etc.

Other costs are related to the work on each case. The user may spend hours in the library searching for relevant literature, or telecommunications costs and fees for accessing computerized databases may escalate.

These are variable costs which will vary from case to case. The variable costs of a case will have to be added to the calculated fraction of maintenance costs to determine the costs of information retrieval for that case.

Availability factors may be classified in different ways, but there is one distinction which is quite important - that between pragmatic and formal availability factors.

Pragmatic factors are the costs associated with purchases and fees, expenditure of time and money to access and use information systems. There are numerous different pragmatic factors.

An interesting, though trivial factor is distance. The costs associated with using a certain information service are related to the distance from the user to the place where that service may be accessed. This distance is an availability factor, only to be overcome through incurring costs - the user spends time going to the files in the neighbouring room, the next floor or the local library, or the user has to wait for a mailed request to reach a documentation centre. It may be offered as some sort of natural law of the use of legal sources that the frequency at which the source is accessed is directly related to the distance between the user's desk and the point of access.

Pragmatic factors have the common characteristic that they may be overcome by the expenditure of costs. By allocating sufficient resources, a user may always have the information made available in spite of severe pragmatic availability factors.

Not so in respect to formal availability factors. These are circumstances that determine the access to information services, but cannot be overcome by incurring costs.

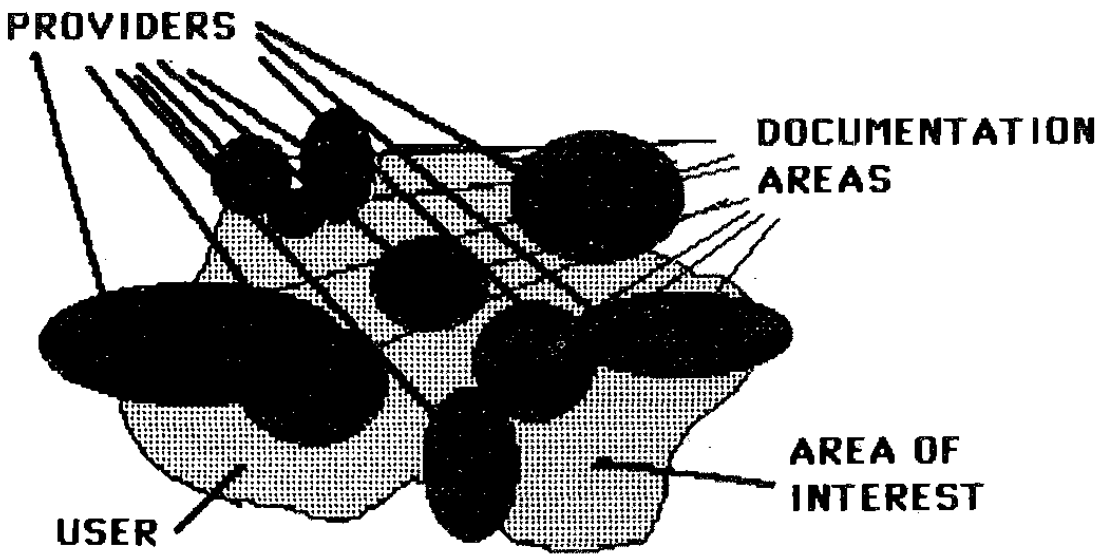
A typical example is the formal availability factor of the law of confidentiality. In many jurisdictions, decisions by public authorities are a source of law - a new decision must always take into consideration the result of prior decisions. But these decisions will generally incorporate personal information on the clients subject to the decisions. And such information, typically, will be protected by confidentiality. The lawyer working within that agency will have access to former decisions, and may argue on the basis of such decisions. But a lawyer representing a client is denied access to the files containing the prior decisions, and cannot utilize this important source of law in his own legal argument. And this availability factor cannot be overcome by incurring additional costs - it is normative, and may not be removed by user effort.

The maintenance costs are closely related to the user-constructed information system. Obviously, this information system is not designed by accident. The user has a rational motivation in acquiring or subscribing to a service.

The user will have some general idea of which future problems he may be required to respond to. These are problems corresponding to his specialization or office, and may be described as his area of interest. When assessing possible information services, the user will try to prepare for his future work, and obviously try to find services that are useful with respect to his perceived area of interest.

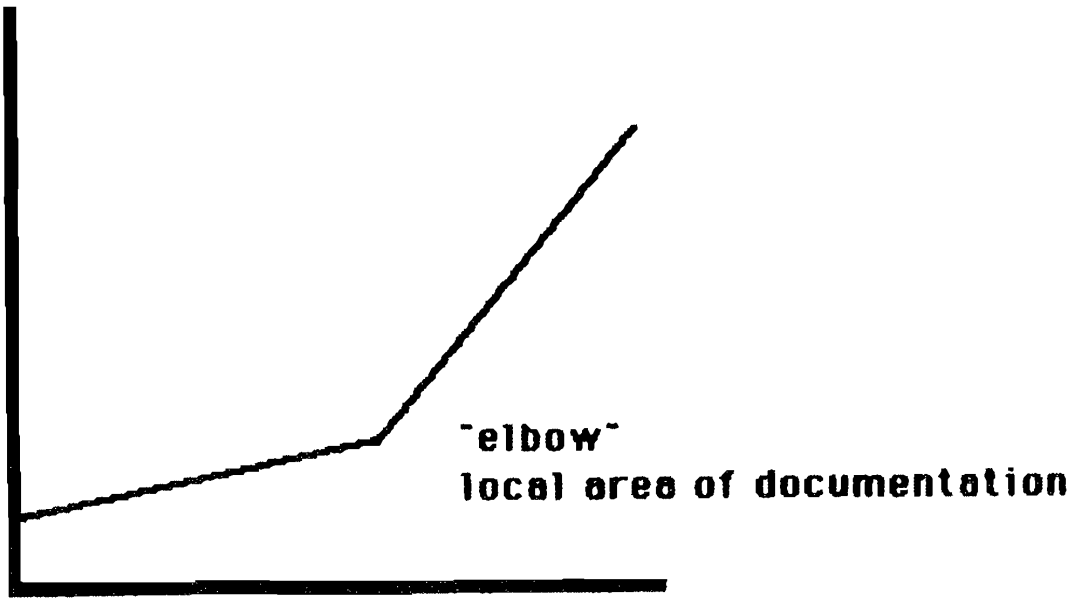
Any information service offered will have a documentation area. When acquiring legal information services, one may picture the initial assessment of the user as an attempt to identify and acquire information services with documentation areas overlapping his area of interest. By adding service to service, the documentation areas of the services provide an overlay on the area of interest.

Figure 5 - Documentation areas overlaying the area of interest



For this reason, the user acquires in a systematic way services most probably useful for his future problems. And in solving such a problem, the user will have most available those services generally most useful. Consequently, the user will typically first employ these easily available services. Only when these do not yield the necessary information to solve the problem, will the user move to other and less available systems.

Figure 6 - Typical cost curve for one case

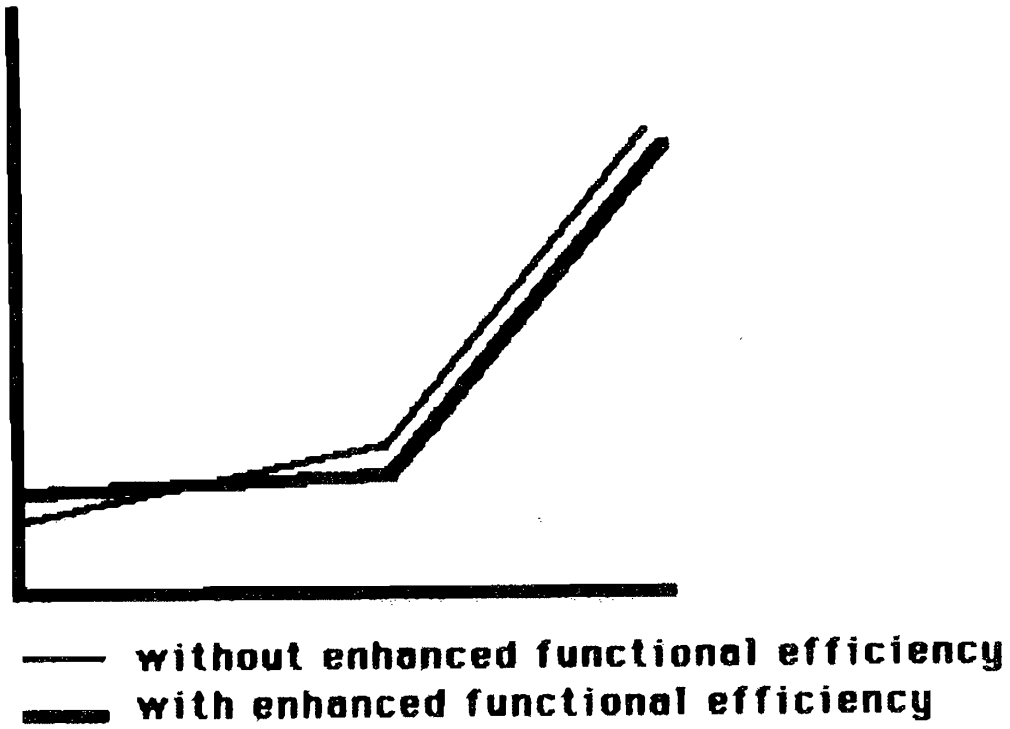


In this way, we may argue that the curve will become progressively steeper. It may also be argued that the curve typically will have an "elbow", indicating the point where the user leaves those services prepared for utilisation by prior acquisition, and accesses other services. This "elbow" may be taken as a definition of what is to be considered the local "database" of the user-constructed system.

One of the relevant factors influencing the cost curve is the functional efficiency of the user-constructed information system. As stated above, a computerized service represents, for its documentation area, a strong enhancement of functional efficiency.

The replacement or addition of a computerized service will have an impact on the cost situation as illustrated by the curve in Figure 6. If only replacing one or more existing services, the maintenance cost typically will increase, while the variable cost will be brought down. Whether it is "rational" to invest in computerized system in this situation, will be an assessment based on whether the user on average needs to access a volume of documents greater than that indicated by the intersection of the curves representing the situation with and without the enhancement.

Fig 7 – Replacing a traditional system with a system of enhanced functional performance.

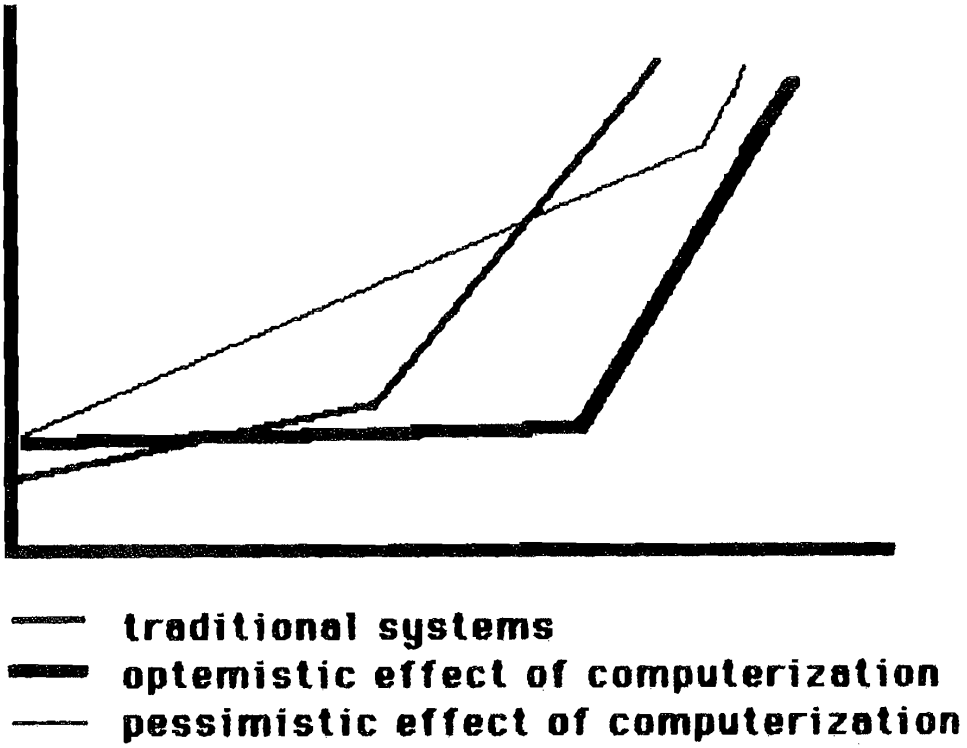


The situation is, however, rarely the simple replacement of one or more services. That may have been the case for the early specialized, closed systems, where computerization replaced manual systems within organisations such as an administrative court. But with respect to current general and open systems, the computerized service also generally increases the documentation areas of the acquired systems.

For a private practising lawyer, whose area of interest is very broad, the acceptance of computerized services may very well rely on this increased coverage of his own user-constructed system, which will incorporate the computerized service. This increase should give reduced costs, at least when accessing documents on the borderline of his area of interest.

It is doubtful whether computerized systems as a rule have achieved this result. The situation may frequently be that though the area of documentation is extended, the reduction of the variable costs is very much less than hoped for, creating a different cost situation for the user (See Figure 8).

Fig 8 – Supplementing the user-constructed system with a general computerized service, optimistic and pesimistic results.



Trying to find the reason for the pessimistic result, several factors may be listed. First, one may mention costs. The service charges of the legal information service often have curious tariff structures, partly adopted from other applications - structures which actually penalize correct use of the system (e.g the specification of a large number of synonyms, which implies a great number of disk accesses, but clearly is desirable in text retrieval). More thought should be given to the tariff structures.

And to the service charges is added the telecommunication costs. These vary greatly between jurisdictions, but are often felt to be inhibitive, especially to users far removed from the physical location of the computer facility. The introduction of packet switched networks with uniform tariffs may reduce this problem, though the tariff structures in such networks are often not designed to promote communication of large packets of text, and therefore may be felt to be inappropriate for text retrieval.

Secondly, computerized systems are not as accessible as one should like. If based on a dedicated communication terminal, this rarely is on the desk of the user; it will typically be located at a library or another room shared by several users. As mentioned above, distance is a trivial availability factor with a major impact on user behaviour.

If the users access the system through a communicating word processing system, this is often located at the desk of a secretary, adding to the problem of distance the problem of interrupting another person.

Therefore the ideal situation would be to access the service through a work station on the user's desk. Actually, the revolution in personal computers is bringing the computer onto the desk of the user, so one may eliminate distance as an important availability factor.

But there still remains the third major reason for insufficient availability, the user interface. The problem for many users starts with logging-on procedures, which involve dialling a certain number, entering of account numbers and codes, etc. It continues with the problems of using the command and search language of the system.

Traditionally, one has seen text retrieval systems as quite simple systems, requiring only a few hours of training for use. But the friendliness of the user interface may have been decreasing over the last decade. When the current text retrieval systems were introduced in the early 1970s the user interface was quite simple compared to what was required to communicate with other computerized applications in use at that time, which generally entailed use of codes and a cryptic command language governed by very strict rules. But in the last few years, microcomputers have been given very userfriendly software, employing menus, icons, mouses, touch-screens etc for communicating with the system. A comparison of the interfaces of the text retrieval interface with that of the in-house word processing system will probably not favour the former.

This is really quite a challenge, as the use of text retrieval relies on telecommunications, and the common protocols for telecommunications make it more difficult to adopt the same type of interfaces as found on micros.

Nevertheless, it is this problem we will discuss as a conclusion. We have stated that it is necessary to bring legal information retrieval services onto a new level of availability in order to secure their general acceptance. We have indicated some probable reasons for insufficient availability. The tariff structures of the providers or the telecommunication authorities are problems outside the systems themselves. Physical access to the services is mainly related to end user equipment, and is therefore also outside the systems. What is left is the user interface, and this can be improved to increase availability.

6. Improving the user interface

As mentioned above, the first text retrieval system of Horty was based on Boolean search requests. This is still the case for all major legal information retrieval systems. The user specifies which search terms should occur in a document of probable relevance, and combines these terms with Boolean operators, especially AND and OR.

This may be seen as the curse of Boolean search requests, a curse which reduces the performance of text retrieval quite unnecessarily. It is by itself somewhat puzzling why other search strategies have not become common, and that the Boolean requests are still thought to be (1) simple and (2) efficient. Both these beliefs rely on illusions.

It is very easy to demonstrate that Boolean requests are not "simple" in the sense that users employ the operators correctly. Analyses of stored search requests have shown that a surprisingly large number of requests are formulated in an inappropriate way, using the logical connectors incorrectly. It is also probable that the use of logical operators makes the users disinclined to formulate complex or long requests, and that the required use of Boolean operators therefore contributes to the observed brevity of search requests.

The efficiency also relies on an illusion. Most providers of legal information services instruct their users on how to get the answer set down to a manageable number. A Boolean request giving as the result some hundred documents is seen as unsuccessful, and the user is advised to add on further specification by ANDing search terms to the initial request.

The reason for retrieving a large number of documents may be one of two, either:

(1) the search request was too general compared to the problem. In this case we have an instance of over-recall, and it is quite appropriate to tackle this by specifying the request further, ANDing search terms until the request corresponds to the problem in specificity. In this case, the strategy recommended by the providers is quite proper.

or (2) the search request is based on a problem of a generality which makes a rather large number of documents of equally probable relevance. In this case, it is inappropriate to specify the search request further by ANDing search terms. That implies entering a Boolean lottery, where the result of having relevant documents in the answer set certainly increases, but only by excluding other documents of probable relevance.

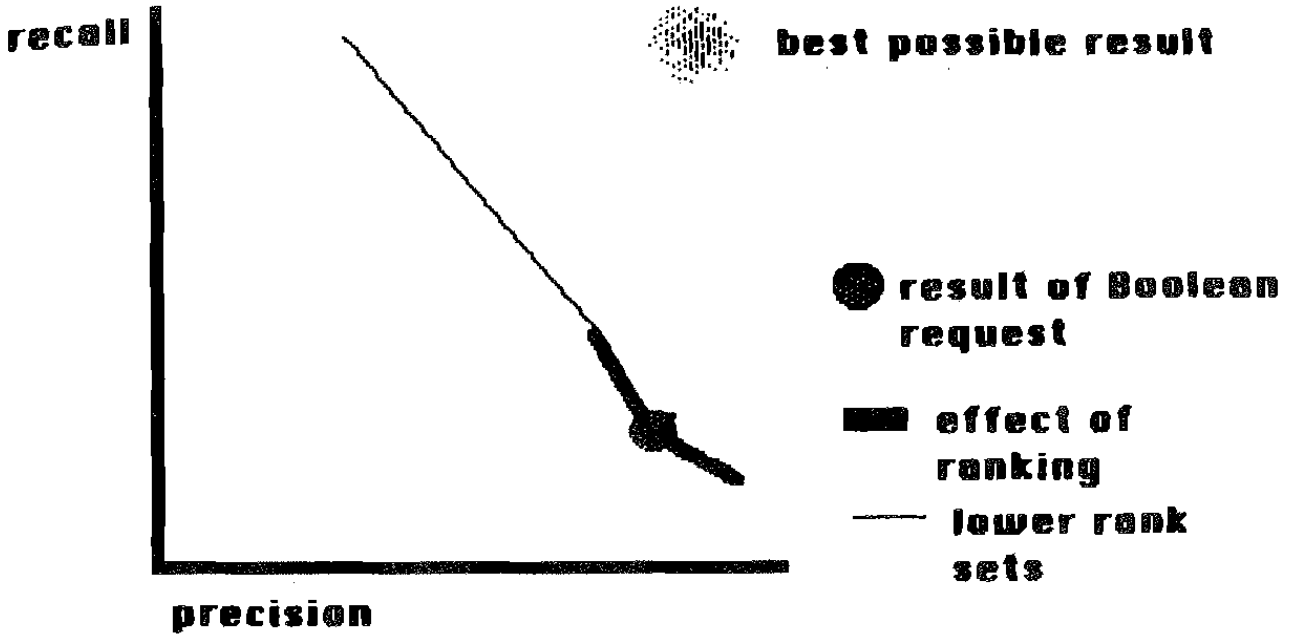
The result of this latter exercise is that the user will find some relevant documents. As there is no alternative way of retrieval, the user will be unaware of the relevant documents discarded during the search process. Therefore the illusion is created that the system is efficient and the user has retrieved all or most relevant documents.

This argument has recently been confirmed by Blair/Maron 1985, who have empirically established that lawyers over-estimate the results of their research. Blair/Maron are, however, mistaken in presenting their results as an experiment on full text retrieval systems, as their results relate to the use of Boolean strategies, and have been preceded by theoretical analysis putting forward the hypothesis their study proved.

The response to this situation has also been known for a long time, and would be to introduce a ranking function to supplement the Boolean strategy. Several ranking functions have been suggested, and some have been implemented in certain text retrieval systems. The more simplistic of these functions have not performed satisfactorily, but there are strategies which have all the power of a Boolean strategy, combined with greater ease of use and improved performance. One such strategy is the conceptor-based retrieval strategy developed by the Norwegian Research Center for Computers and Law and implemented in text retrieval systems like NOVA*STATUS and SIFT.

The conceptor-based strategy will give identical average performance to the Boolean strategy. It will, however, increase the probability of the first documents being relevant, and it will increase the probability of retrieving a larger fraction of the relevant documents in the database (increasing both precision and recall). It makes it easier for the user to formulate the search request, and encourages specification of synonyms. It can be implemented without changing any basic design principle of a text retrieval system, and may actually be added to existing systems as an alternative matching function to the one necessary for pure Boolean retrieval.

Fig 9 -- Ranked compared to Boolean results.



As indicated, the conceptor-based strategy is only one of several ranking strategies of proven value. CLIRS may be one of the first operational legal information services to have included a non-Boolean retrieval strategy, currently being developed by Lewis Pape of Computer Power, Canberra. This is based on a weighted term ranking method not dissimilar to the ones explored and promoted by Gerald Salton of Cornell University, U.S.A.

To this improvement of the retrieval function should be added greater emphasis on training. User research has proved beyond doubt that current training schemes are not sufficient. It is easy to "unlearn" the use of text retrieval systems, a holiday is all it takes. Also, the frequency of use is on average so low as to only just maintain the necessary skills for using the systems. It is doubtful whether conventional measures will suffice for amending this situation; longer courses or frequent refresher courses are not the solution, as users will not have the time, money or motivation for this type of training.

A better solution would be to address the help functions of text retrieval systems. Today, help functions are quite passive, the user has to ask for help to be assisted, and is then very often given help in the form of a general and condensed text book of instructions.

This should be replaced by an active intervention of the system. The system should monitor the dialogue, and butt in with advice when detecting that something is wrong.

The first level of intervention should address the problems of correcting mistakes and simple expansions of the search request. Analysis of stored dialogues by Norwegian users has disclosed that as much as 9% of words are misspelled. This does not mainly reflect on the fluency of lawyers in their own mother tongue, but rather on the causes for errors created by an unfamiliar keyboard and communication protocols. A misspelled word in an ANDed request will generally result in an empty answer set, the system returning with the message NO DOCUMENTS RETRIEVED, which is of little help to the user. Obviously, the system could easily return with the message: THIS WORD ...DOES NOT APPEAR IN THE DICTIONARY, PLEASE CHECK FOR MISSPELLINGS.

This is only a small example of a trivial improvement which would have obvious and positive results. The system could, however, go further. It should, for instance, check the logic of the search request, and assist the user to formulate the request correctly (and not just return with the response SYNTAX ERROR).

But leaving aside pure errors, systems should also draw on computational statistics of the vocabulary of the database, identifying, by frequency and distribution, words which are inappropriate as search terms, flagging them for the user and offering an amended request. Certain words, typically numbers, should only be allowed if part of phrases in the request. And the system should be able to recognise that a date, a section number etc was part of the request, and automatically produce alternative representations of the term ("section" being represented also as "s" and "sect").

The system should also have the possibility of automatic expansion of search terms. This is not a suggestion for incorporating a traditional thesaurus, but rather, at this level, to use simpler means, such as automatic truncation based on rules derived from computational linguistics.

These suggestions for a first level of improvement are really rather modest in programming resources, but with improved retrieval functions they may result in a considerably more friendly system.

The second level is also not ambitious: the inclusion of natural language search requests. Obviously, no natural language "understanding" is suggested; rather this is a combination of a ranking function with the automatic identification of inappropriate terms and the expansion of search terms discussed above. The result will be a very simple way of formulating search requests, completely without restriction. The result will certainly not perform as well as the result of a well-structured request by an expert user. But one has made the system more available, and made it possible for the holiday-makers to return to the retrieval system after a few weeks absence. Hopefully, the enhanced help function will then make it easier to regain a higher level of expertise, and, after a while, to construct search requests more expertly.

This second level does not require too much in terms of development and programming, but the third level is not at all easy to achieve. On this level, one would like to include conceptual tools for the user, and make the system learn from experience and be self-modifying.

The keys to the solution are two.

First, research in artificial intelligence and law may offer solutions for representing legal concepts and their internal relation. These could be used to build conceptual models of areas of law, and to represent these structures to the user. Rather than formulating a search request, the user would traverse the conceptual representation, drawing on his legal expert knowledge to identify those concepts of interest. The system would then translate the concepts into search requests and retrieve the probable relevant documents.

In doing this, the system would draw on the experience of past requests. Obviously, in an interactive environment where thousands of requests are processed each day, the requests contain a wealth of information on the problems of users, which terms these problems are expressed, in and the relations between the terms. Processing these requests, the system may make conclusions and use these to update and amend its conceptual structures.

Actually, storing the requests, processing them and linking terms may be a simple way to develop and maintain a more traditional synonym thesaurus. This is currently being explored at the Norwegian LAWDATA foundation.

The third level is only sketched here. To detail it further would entail going into the problems and possibilities of modelling legal norms. Therefore, it may be left at this stage, rather like a carrot in front of the donkey hauling the present legal information systems. It is to be hoped that in lunging for the carrot, the donkey is gradually transformed and becomes a more intelligent animal. The text retrieval system of tomorrow should be educated to the level of being a research partner of the user, discussing the legal problems in reasonable terms and assisting in finding the primary material for solving the problems of lawyers.

Bibliography

General references have been omitted. The first work cited provides a survey, and includes extensive references to the literature.

Bing, Jon/Harvold, Trygve/Fjeldvig, Tove/Svoboda, Robert (1984) Handbook of Legal Information Retrieval, North-Holland, Amsterdam.

Blair/Maron in AMC Communication March 1985.
Jungjohann, Kurt/Seidel, Ulrich/Soergel, Werner/Uhlig, Sigmar (1974) Informationsverhalter and Informationsbedarf von Juristen; Datenverarbeitung im Recht, Beiheft 2.