9th BILETA Conference
The Changing Legal Information Environment

11th & 12th April 1994
Scarman House
University of Warwcick
Coventry

Wisps of smoke? the electronic library, new information retrieval techniques, and diminishing returns

Derek A. Sturdy

Keywords: Electronic library - Databases - Selective Dissemination of Information (SDI) - Control systems - Full Text - Abstracting - Indexing.

Abstract: This paper considers how an electronic library can be made useful to the user, the person for whom it is constructed and who, in many cases, has to pay for it. The author will attempt to outline why doing something just because it is technologically feasible can waste prodigious sums of money, while doing simple technical things on the base of properly applied human resources can remain the most effective, and much the cheapest, way to benefit the user. Utility is the yardstick throughout; however clever, a system which fails to benefit its users is assumed to be a waste of time and resources.

Summary of the argument

1. Full text should be kept in as simple an electronic form as possible. Money spent on complicated systems will actually be counter-productive, making the system hard to upgrade to future technologies.
2. The difficulties of retrieving anything from a full text database or collection (an electronic library) by software alone increase in size and complexity with the size of the database. Above a certain size, full text databases or collections are effectively unsearchable, using the yardstick of benefiting the users, without subsidiary systems. This is normally because a search request retrieves too much and irrelevancies multiply.
3. The most important element in control systems, designed to make large full text collections usable, is the human brain, used as a indexing and abstracting device.
4. The most important place to spend money in developing an electronic library is in the control system - the electronic "catalogue" - and its associated salary and software costs, and not on the actual storage of full text.

Electronic V Paper: has a consensus already emerged?

One issue is the desirability of the electronic library. Throughout this paper, just as it is assumed that utility is the criterion of excellence, so it is assumed that we all believe, in general principle, that the electronic library is the way forward. Two general rules might be adduced here to show at least where I stand on the issue:

1. People hate searching paper.

2. People hate reading a screen for long.

From which follows:

3. Search on a computer screen, then print out the retrieved full text, neatly and legibly, on decent paper, for full-text reading.

From this does not follow, of course, any requirement to store hard copy; a requirement to have the intellectual property issues, nicely tied up and sealed, is much more important and not within the scope of this paper.

The SDI analogy

To most of us, SDI means the selective dissemination of information, but this paragraph refers to the ill-starred "Star Wars", the strategic defence initiative of the latter years of the Reagan administration. Star Wars had a great deal going for it; lots of academic research grants, lots of business for high-tech companies, and glamour, and there was nothing wrong with any of that. There was even a strong sense that, just as ingenious and ugly hiking equipment emerged from the astonishingly large expenditure on the Moon programmes, so something equally of use to consumers (the people who paid the bill) might emerge from the SDI programme as well as the desired end of preventing hostile missiles from landing on democratic countries.

The arguments and developments alike concentrated on the hardware. Lots of literature emerged on the Laser v Brilliant Pebbles issue, for example. A few lone voices had pointed out the catastrophic flaw in the whole thing right from the start. This flaw was, that while the hardware was all feasible or at least conceivable, no-one had the faintest idea of how even to approach the software. The control system is, in our present state of knowledge, completely unwriteable, and even if we developed techniques for writing it, would remain untestable. A test programme, in which failure might result in millions of deaths, is not a test programme at all, but a lottery with the lives of innocents.

At the time of the major interest in Star Wars this point went largely unregarded for a significant reason. The software that controls systems is completely arcane to almost everyone, whether it controls the production of chemicals in a modern industrial plant, or the flow of letters on to a screen as you write on your computer. The complexity of getting a large industrial plant to run at all, let alone safely, profitably and consistently, is almost unbelievable to anyone who has not done it, but merely sees the occasional disaster when something goes wrong. The assumption is therefore made that, where complex software needs to be written, the *"boffins"* will somehow come up with an answer. The proposition that, even if they do, their resulting programs and systems might well be untestable, which in the Star Wars case can be grasped by an intelligent seven-year-old, is not even considered. Unfortunately, those involved with the project have every reason to let that proposition remain dormant. The result is the potential for spending large amounts of money on a system which will probably not work.

There is a danger of Star Wars syndrome developing in electronic library projects. The hardware is all there, or can be envisaged readily enough without major technological breakthroughs. Scanners, OCR devices, large storage media, fast processing, and networked terminals, are all in place in many institutions already, and their cost drops every week. The text retrieval software is also either in place, or recognisably about to be in place. If you know what you want, the argument therefore correctly runs, what need have you of storing paper? It is not quite so simple. The problem is that, just as you could never find what you needed in the miles of stacks of books and filing cabinets, you will not find what you want in the electronic library either. The question then is: Why did we spend so much money to make no difference to our lives? The missing ingredient, as in Star Wars, is in credible control systems.

Control systems

What do we mean by a control system for an electronic library? Here I suggest that we always consider first and last the information seeker approaching the terminal, who, in the case of the electronic library, will in most cases not be a trained information or library professional. Lawyers may think in concepts (striking out, restitution, champerty), in cases (Owens v Bracco, Rylands v Fletcher), in legislation (Section 740 liabilities, or other phrases nonsensical to non-experts), or just in words and phrases that sum up their need for knowledge by combining concepts with more concrete words (surety and the home). If the control system cannot present to the researcher a neat way to research in their own terms, it is not functional by our definition - the usefulness to the user. Worse still, if, when the researcher asks about champerty, so much clever junk is built into the system that about 500 records are listed for potential viewing, the researcher will turn away in despair. Lawyers know there can only be a few recent champerty cases; if they want wider concepts, they will ask for them themselves, using their own terms, their own rules, their own brains; only rarely will they ask the machine for help, because the machine is not a lawyer, but a machine.

The control system is therefore a quite different matter from the expert system. The expert system has its place, which we will return to later, in creating, updating, and modifying the control system. This place is further back in the process. The control system's job is to get the researchers as quickly as possible, and as accurately as possible, to the stored information they need, whether they know it exists or not. To do this, the control system relies on two essential ingredients: a front end, and an index. The index is the difficult issue and the one I shall address here. The front end is easy: both American and British companies are good at writing friendly, usable interfaces between novice and expert users and the databases they wish to access. Although amazingly user-hostile systems are still around, and, indeed, are still sold, no-one has to buy them and no user has to put up with them for long.

The control system and the library catalogue analogy

Our researcher, twenty years ago, had no difficulty at all about how to start. There was the catalogue - mighty volumes with pasted in scraps of paper, card indexes with the heavy thumbing on particular cards pointing to the favourites of your colleagues, - and you browsed through happily and inefficiently. Once you had a small list of material, you went to the stacks, or asked for the periodicals or books, and you followed ideas backwards from there.

Five or ten years ago, our researcher had a more serious difficulty. The library catalogue was on a computer, and was much more efficient in terms of retrieval. For instance, a subject search did not confine itself, as the bibliographies at the back of articles tend to do, to a back-slapping circle of cronies; your search was much more complete. Unfortunately, the skills required to get at the information were of an order of magnitude greater than those required to thumb the old-fashioned catalogues. So daunting were the screens, with their commands which appeared to be barked at you in the horrid green letters of the day, that researchers either relied on a few experts, or failed to find their material. Worse still, their failure to find material made them feel inadequate, though the fault was not theirs at all. The first stage of paying heavily to make things worse had been completed.

In both cases, however, the catalogue was the natural first stop for research, and the people who created it were the people who controlled whether the library was useful to its users or not. Certain rules became established in creating the old style of indexes:

- Don't let the absolute experts create the catalogue sections on their areas of expertise, or nobody else would ever be able to use those sections.
- Equally, don't let an engineer catalogue your law material or a lawyer your scientific books.
- Therefore: The best person to catalogue a law library is a librarian with a general law degree, or a lawyer with an information degree.

These simple rules are, I suggest, as valid today as they were twenty years ago. The aim of the control system, in short, should be to allow our researcher to approach the system with the same confidence with which the card index was approached before, but with the expertise of the old card indexer greatly augmented by the electronic facilities available. The system has also to deal with two widely differing levels of enquiry:

1. The specific, jargon-ridden problem (as in Section 740 liabilities);
2. The search for analogous material outside the sphere of expertise of the searcher (a mortgage point takes the search through contract law to, say, shipping, where our searcher has never practised).

Different techniques are required, but the control system should not ask the researcher to define which way the search is to be conducted. First, how can the researcher possibly know, since the distinction depends on criteria built into the control system? Secondly, why should the researcher care? Answers, please, fast, correct and complete, with no more silly questions, would be the reaction.

Full-text V Abstracting and Indexing

This is still widely debated, however strange it might appear that it can be perceived as a sensible matter for debate. I shall outline here an argument which is commonly used, and wrong, and I shall then attempt to explain the fallacies and propose a working alternative. The fallacious argument runs, if it can be condensed so simply, like this; given all the wonderful software we could write, or claim we have already written, who needs abstracting and indexing? You search for words, you use proximity, you use relevance ranking, you use inverted relevance ranking, you stroll, in short, down a cybernetic primrose path, and from the text in your on-line storage up comes the paragraph with the gist of your problem. You have to have the full-text anyway, or else how can you retrieve it when you want it; so why bother with the fallible, human function of abstracting and indexing?

This argument is, of course, perfectly sound provided your database is small. There is no point in making an elaborate index to this paper, viewed as an isolated, self-supporting document. Any simple word-processing program will allow you to search its content for almost any concept or idea that you hope or fear it might contain. It's different when you amass the papers from the complete conference. What happens when we reach the gigabytes, the terabytes, of the electronic library? The argument becomes fallacious as a function of the scale of the database.

It is at this point that we reach the corollary of the Shakespeare/Monkey proposition. Every schoolchild who has attempted any statistics or probability theory knows that, given enough time, enough typewriters and enough monkeys, sooner or later complete sonnets, complete plays, the complete Works, will emerge from the industriously, but randomly, typing primates. Another copy of Shakespeare's works can only be welcomed, and there is probably little in the way of an intellectual property issue in any other country than the USA. The corollary is much less happy. Let us first formulate the Shakespeare/Monkey proposition: The longer you type randomly, the more serious text will appear.

And now let us turn the proposition round, to find: The more serious text you have, the more random combinations occur.

We are all familiar with the difficulty of searching full-text databases for, say, items about dangerous dogs. "Malaysia seeks to muzzle UK press" is just the start of the rot. "Chancellor dogged by predecessor 5 mistakes" is followed by "HSBC lets loose the dogs of war on Midland", and "Last Alsatian coalmines to close". Who is to flag all these doggy terms, and make sure that our researcher just gets back to the issue of the postman and the Rottweiler?

The dog issue is an old chestnut of text retrieval. Unfortunately, as the full-text databases grow, the "dogs" multiply. More and more irrelevancies pile up. When there were just, say a gigabyte of "know-how" documents, a search produced only a few false drops of this kind (for example, the colourful phrase such as "the opposition were striking out for the shore of admissibility" which has nothing whatever to do with the legal sense of "striking out"). In time, the proportion of false drops to good material rises, and it goes on rising until the full-text database is effectively unsearchable. There are those who believe that some of the largest online databases available in the world have entered a state close to this already. Even though I am not yet one of them, I suspect that the false drop is an increasingly serious problem, one which grows exponentially, one which was not perceived as an issue in the early days of full text databases because it only appears with size, and one which is actually a matter of common-sense coupled with a generous measure of hindsight. In time, like the monkeys and their sonnets, a full-text database must grow to the point of being unsearchable without a separate index; and long before it reaches that point, our criterion of usefulness to the user will not be met.

The traditional answer to the false drop problem was the humanly-produced index. This is what the card-index did in the days before it was possible to show full-text documents on a screen, and it served a vital purpose in making sure that research pointed only to relevant material. The sheer volume of raw text produced for new full-text databases is, however, daunting, and so software-based solutions to indexing were and are being devised to eliminate the human aspect of the indexing. I shall argue here that they are doomed to failure, and that the human brain in the present state of our knowledge remains much the most effective, and the most cost-effective, way to produce the required indexes.

Software solutions to the false drop problem

We could argue that human indexing presents a serious theoretical difficulty. If you rely on human indexing, by definition you only get what the human produces. Given the limitations of the memory, capacity and patience of the indexer, how can you possibly get a good index? Surely an expert system could be devised which would do the job right, every time, and not rely on someone who might skimp the job in order to catch the last bus, be feeling cross or breezy, or just not be very good at the job?

The obvious answer seemed to be hypertext linking. In essence this concept is no more and no less than has been practised by information people for years, by means of a Thesaurus. A Thesaurus is a dictionary of keywords - words and phrases which have a specific meaning either to everyone, or within a discipline such as law, to lawyers. Various (not all) hypertext approaches aim to replace the thesaurus either with a machine-derived fixed list, or a variable, made-for-the-moment list of word associations, which nonetheless, at the moment of its existence, functions as a thesaurus. The difference lies in how a hypertext link formulates the criteria for the link.

An old-fashioned Thesaurus is intentionally restrictive and mostly relates to concepts -which is why lawyers find a thesaurus particularly useful. The keywords might just be listed, or more commonly, arranged in a hierarchy with "top terms", concepts which had no useful broader term, presiding over trees of narrower, more specific terms. Although a thesaurus evolves, it does so relatively slowly, and by its restrictive and directional functions it acts as a means of enforcing consistency - something which the instant hypertext associations obviously seek to avoid. Where hypertext uses, at least in theory, the free association of words based on instantly or previously prepared analysis of the actual text being searched, the conventional thesaurus approach uses the restricted keyword list to form the bridge between different slices of text or data. In most modern software, of course, the researchers can also use any term selected by themselves to form their own links, but that is separate from the concepts we are discussing here, which are machine-aided links.

Unfortunately, the text retrieval manuals hijacked the word "keyword" and distorted its meaning until the most common misconception among database users (I emphasise users, not creators or information professionals) is that, if your database contain keywords or makes use of a thesaurus, you can only search for keywords. This silliness arose because:

1. Text retrieval manuals referred to the searchers' search terms as "keys" or "keywords"; after all, they had to call them something,
2. Lesser, or inadequate, text retrieval systems only indexed bits of the data, those marked as "indexed keywords" or "manual indexed keys" or some such jargon;
3. Users came to believe, usually wrongly if they were using decent software, that they could only search for keywords, which they could only find by an intimate knowledge of some thesaurus, because they were the only words indexed by the software!
4. Users then made a further logical jump, concluding that only full-text databases had fulltext indexing, a connection which has never been true.

These misconceptions are remarkably prevalent today, when all modern text retrieval software indexes everything presented to it, and no information professional would dream of buying software for the purpose of text searching and text retrieval (as opposed to document management) that did not do so. Mention a keyword, and many casual database users assume, half-listening and half-understanding, that you can only search on keywords. Mention that a database has every word indexed, and the assumption is often made that it must be a full-text database of documents. With the arrival of new document management systems where the opposite is true, the confusion is set to spread; many document management systems, in contrast to text retrieval systems, provide only an index card to documents themselves, and do not attempt to provide full-text indexing.

The essence of the use of a thesaurus of keywords is that the words should provide conceptual bridges within a well-defined discipline. A thesaurus used by engineers, for example, would relate "construction" to getting something built. To the lawyer, however, "construction" might mean "interpretation" and be nothing whatsoever to do with large people in hard hats. In the first case, "construction" would lead to many other terms to do with getting things built. In the second, "construction" might eventually lead to semantics - no doubt to the horror of the researcher. however, the essential point is that a document which has been keyworded from a thesaurus provides a consistent, predictable, safe path to other documents about the same concept, and because of their tendency to think in concepts and issues, this is particularly valuable to lawyers.

Some hypertext structures aim to create bridges without the use of such a controlled list of keywords, The theory is that the text should be allowed to "speak for itself". The actual words, and their proximity to each other, used by writers will define the links which are actually valid. The meanings which the writers seek to convey can be distilled by what amounts to textual analysis. Whizzy software can perform the analysis, infinitely more quickly than any human, and hence, when the searcher types in the word "champerty", the multitudinous relationships which textual analysis has worked out will be ranked and presented to the searcher. This will ensure that links to related matters are at the very least made available, and that these links would be real in the sense that they were derived, not from the prejudices of a human indexer, but from an objective analysis of what writers had actually written ("text").

There is a whiff of structuralism in this argument which, I have to confess, I find unappealing, airy-fairy and unhelpful in practice. Aside from this personal prejudice, the difficulties with this theoretically pure approach prove to be:

1. The Shakespeare/Monkey problem sooner or later, every word gets an association with every other one. Elephants and porcupines are, you assume, associated easily because they are both mammals. When elephants and champerty become associated because someone carries out a questionable insurance operation at the Elephant and Castle...
2. You substitute the hidden textual prejudices of the writer for the visible prejudices of the indexer. It is much easier to recognise and, if necessary, analyse the indexers' prejudices

> because there are only a few indexers compared to many writers, and because the indexers' prejudices are set out, neatly tabulated and printed, in the Thesaurus they are using. The writers' textual prejudices are usually hidden and are only to be revealed by methods about which not everyone can possibly agree, and which most practising lawyers will have absolutely no time at all even to discuss.

Hypertext has, therefore, the option of returning to the building of links according to a set of externally defined rules. In their simplest form, these rules might be a straight instruction by the writer, or information controller, or whoever, to associate a complete document with another document or to place it in a set of related document - an important part of case management structures. more usually, the good old thesaurus is used, as it always has been, to provide links. The predictability of this procedure makes it useful; but of course, with large full-text databases, the false drops begin to multiply (as in the "striking out" example given above) as the size of the database increases, and we return to where we started. In short, if you are going to all that bother, why not produce an old-fashioned index, and stop even calling it hypertext.

The Size Problem

Just as false drops multiply with the size of the full-text database, so does the simple problem of finding room for all the material which has to be on hand with automatically indexed text. As you index your full-text database, you acquire index files and, in many cases, some sort of dictionary or term file as well. If you are to include proximity data, you will need long tables of where each word or phrase occurs. In due course, these database management files themselves occupy a serious amount of space which has to be found as well as the space for the text itself. At this stage, the natural move is to place the actual text elsewhere - e.g. on CD or other optical media, or on removable hard discs - so that retrieval of a document becomes a matter of requesting the physical mounting of some medium which normally resides in a cupboard or safe, rather than on-line.

At this point, again, you wonder why you are working in this way. Suppose your indexed entries, automatically generated from the text itself, have led you to request the mounting of, say, four different actual storage devices on the system, a process which might be queued and take considerable time; and you then discover, when you look at the text, that your finds are mostly false drops? By the time your full-text database reaches a certain size you need to be almost completely confident that the document or text you are asking your system to retrieve is indeed one which you wish to read.

In conclusion, hypertext is unsatisfactory because it demands, if it is to work as anything but a gimmick, the whole full-text material to work on, and therefore demands that the complete system incorporates either full indexing of all text in every document, or on-line availability of every document for textual analysis. All the size and false-drop problems associated with trying to work only with full text are compounded by the hypertext concept, and none of them are solved. The concept that works splendidly on a single CD will be a jazzy nightmare on a seriously large database of the kind that is meant by an electronic library.

The problem of technological progress

There has been a tendency to believe that all these matters can be resolved by the simple expedient of spending money on sufficient computing power, and on enough time and effort on software. I have attempted to demonstrate that in fact the problem is not soluble in this way because of its very nature. There is a further complication which renders the heavy money approach particularly wasteful. This complication is technological progress.

Let us imagine that you are installing your electronic library, with automatic indexing and text retrieval. The sums involved are, even today, not trifling. The software is written for a particular platform. When you obtain the hardware, it is not quite the latest - you want tried and tested components - but it is very far from being obsolete. You laugh to think how Y & Co., down the road, bought that steam-powered system - Gosh, was it just two years ago? - and then the penny drops. That's you, in two years time. In two years they haven't even got the bugs out of the software. You look ahead at your own software. The platform on which it is being developed - let us say, for argument's sake, UNIX - was much the best for the job when you bought the hardware and commissioned the software. But unfortunately, it is now old hat, because a quicker, smarter system which is easier to program has taken over the running. I do not have to continue the list, because many of us have had to shrug our shoulders and agree that we bought the best available at the time, and apologise for the fact that much better answers now cost a third of what we spent. The problem is only simple of solution if your applications are simple in concept.

First, your ingenious indexing software - written, shall we say, in C for UNIX. How long will it take to port to, and debug in, C++, or LISP, or some new wonder language, for the new platforms? Will you be stuck with obsolete equipment because you cannot translate your software, and if you cannot translate your software, you cannot read your indexes? Is your electronic library, in short, doomed to fossilise within weeks of commissioning?

Second, your actual method of storing text. Has your library system made use of compression techniques? If so, can you translate the stored text back into a string of characters and instructions that a new system will understand? Will you be locked into increasingly obsolete technology because you cannot get your wonderful electronic library to run in any other form? Let us not forget the commercial pressures on people supplying you with systems, who want your future maintenance work, and the way those pressures cannot help but subtly influence design of systems. Unless the actual full text is as close as possible to the original - a string of simple, recognisable characters that will remain readable, and exchangeable into new conventions with the minimum of fuss, you run a serious risk of losing information simply because in time it becomes unreadable.

This is, of course, the nightmare of current CD archiving. The term just has to be a misnomer. We cannot read any of the media commonly used for data storage twenty-five years ago on any of our modem equipment. Unless we have kept the old machines not only mothballed but usable, any data stored on the older media is actually irretrievably lost. What chance is there, that in twenty-five years time, there will be machines around that can read the CDs of today, or that we will have been able to keep in working order any of our current machines which can? Can you even remember what the machines of twenty-five years ago were like? Will you happily absorb the costs of keeping them running? Any archiving programme, therefore, has to include a rolling media-translation programme, which permits conversion of the data carried on the old media (CD, magnetic tape, paper tape, punched cards - remember them?) on to the current standard at the time that the first signs of obsolescence occurs. This is an expensive and serious problem.

The media-translation programme, however, is enormously simplified if the actual text has not been messed around by cunning software. As all of us who have been involved with conversions are aware, whether from one current medium to another or from one old word-processing system to a newer one, that the waste of time and money occurs principally if the data is not "flat". Flat means that the data does not contain control characters, mark-up characters, compression, displacements, or any other of the horrors with which software companies seek to tie their customers to their own product. Flat data is easy to convert, easy to rearrange, and from it one can make copies in all sorts of different forms - which is precisely why there is a strong reason for a software house to make its data anything but flat. The designer and maintainer of an electronic library, which is to be anything other than a nine-days-wonder, effectively has to ensure that all the text is stored as flat text, if there is to be any chance of economical media transfer in future years. This vital point is rarely to be found in the sales literature of those who promote text management or text retrieval systems; indeed, it is rare for them to mention exactly how they store the text at all.

The conclusion from all these cries of woe and doom is simple enough. Don't spend a lot of money on systems which you know are by definition ephemeral because the technology is moving so fast. Instead, achieve a system which is much more solid, and cheaper. This will be achieved by ensuring that you spend serious money on a very small part of the total system, the index, and spend the minimum on the major part of the system, the full text, which you maintain in as simple a format as is currently devisable.

Principles for an electronic library control system

The essential principle is, therefore, that the heavy computing work, and hence the use of both software and hardware, in concentrated on the index. The text itself, however it originates, is considered "passive". Apart from the actual storing of it, whether on line or off line, text is not processed or handled by the control system until it is actually called for. If it is called for, the system ensures that it really is, in all but a very few "rogue" cases, actually wanted.

The index section is compiled by humans because they are still much better at the qualitative judgements required to make an index work. Traditional thesaurus-plus-keyword approaches work, are known to work, and allow searchers if they wish to avoid any false drops. There remain, of course, huge areas which keywording is quite inadequate to cover, and for this reason the short abstract is absolutely essential. Computer generated abstracts are still fantasy, and, as we have seen above, hypertext can only work in the complete full-text environment which we are seeking to avoid. The whole tenor of this argument is that we must avoid all the compression , mark-ups and built-in links which are required to make any sense of automatically generated indexes or hypertext systems if they are not to be ludicrously slow.

These principles emerge from this argument:

1. Separate your index from your full text.

This may seem totally obvious, and in that case, please skip this paragraph. It is the principle on which document management systems are increasingly based. The index can pass requests for specific documents to the document management system. Each system an reside on a different platform, and each system does a quite different job.

2. Maintain your full text in the simplest possible form.

If the document originated within your organisation, ensure that its format is the currently standard one, or, more usefully, translate it for long-term storage into flat ASCII or ANSI and remove all wordprocessing mark-ups, database control characters, and anything else which will be unreadable in ten years time. If the document originates from the outside, keep it exactly as the publisher sent it to you. The publishers will want to keep their material searchable by the latest techniques, if only to ensure continued sales of new electronic products.

If the document results from a scan, chart a course between the Scylla of images and the Charybdis of OCR. The scanning problem is outside the scope of this paper, but seems likely at present to store ghastly problems for its users. No doubt advances will soon allow us to move forward from a system which offers a choice of locking your document into a technology which is evolving extremely rapidly (bit images) and which will therefore offer serious media-obsolescence problems, or a system which at best offers 99% accuracy (OCR), and therefore might well translate "we do not believe that" into "we do now believe that". At present, OCR and proof-reading seems to be the only practical alternative, but who would find the time or the funds to proceed this way I do not know.

Scanning is not yet, therefore, a technology which is obviously useful in the context of long-term storage of text in an electronic library. My own suspicion is that it is currently better to wait for improvements in the accuracy of OCR; for with OCR, provided the text is genuinely accurate, the problems of the immense hardware storage and technical obsolescence associated with bit images disappear. Enterprising publishers will no doubt kill the OCR and copyright birds with one stone by issuing text in flat ASCII/ANSI files at a cost which reflects their valuation of the intellectual property issues.

3. Index each document as it is admitted into the system.

When any item is admitted to the library, it is passed through a human indexing process just as in a conventional library at present. In addition to the usual abstracting and indexing functions, the indexing process will provide a reference which will allow the document to be uniquely referenced. This provides an admirable weeding out process as well. If the document is not worth indexing - a value judgement at which humans are good and computers bad - it is not worth keeping either.

4. Separate the functions of searching and retrieving.

This follows logically from proposition 1 above. It also follows from our original proposition that people hate searching paper and hate reading screens of full text. The researchers approach their terminals (by now, of course, on their desks) and browse until they produce a list of documents which, there is every reason to believe, are relevant. If required, the request for these documents can then be passed to the document management system, which will handle copyright. printing, and charging issues.

The structure that results in these principles is dual. First is a large, cheap, full-text database which is easily translated into a multitude of other forms, has not been expensively processed by computer, and is effectively totally independent of the control system. The control system merely knows where the document is; it does not modify it, add headers to it, mess about with its structure, or alter a single character within it. Next to the large, simple full-text collection is the control system, the heart of which is the index or collection of indices, which contains keywords, abstracts and pointers to the full-text documents. Only the index is searched by the user, and it is far smaller in size than the full-text collection. All the software and processing power money is concentrated upon it, leaving only a simple and adequate document management system to retrieve full-text items requested from the control system.

When the time comes for technological change, the issues are relatively simple. The media-updating programme for full-text is straightforward because there is no translation to do. Hardware is the only issue; software has not complicated the job. The control system should be selected from the very first for its ability to produce, quickly and simply, all its information similarly in flat ASCII or ANSI format. Headings, tags and similar mark-ups are acceptable here as they will be used in the conversion to the next system but will be orderly, composed by humans, predictable and recognisable. If the control system and associated software is not capable of exporting itself in a neutral format in a very short space of time, don't buy it, or you will lock yourself into the fossilisation trap.

Since the majority of your money will, under these circumstances, be spent on the human input, and since you will always have had an eye to the pitfalls of technological change, you will avoid the Star Wars syndrome. First, you will not spend large amounts of money on hardware, platforms and software that will be obsolete very soon after you have bought them. Second, you will have specified that simplicity is the criterion, and that any system which seeks to impose proprietary methods of data storage, index storage or text storage on you will have no place in your electronic library. Thirdly, the effort and money expanded on your human indexing and abstracting will be amply repaid because it will not have to be recreated. The 1994 work will be relevant, even if electronically unrecognisable, in 2034, because it will be incorporated outside the full text, in a system designed for technical change.

Making me most of the human resource

Having argued that humans make the only sensible indexes and write the only usable abstracts, how do we counter the arguments about fallibility, (the last bus, feeling ill, not actually very good at the job problems.) It is at this point that the expert system can assist. The job of the expert system is not to obscure, but to make transparent, the prejudices and principles of the human who is creating the indexes and abstracts. Since the searcher has to rely, by this argument, on the interposition on the humanly-produced control system, that control system must be consistent and the researcher must be able to understand, without serious difficulty, why one search strategy will yield full and complete results while another one will not.

All expert systems should seek to do is to distil the wisdom of your best indexers. To do this, they can work, within the indexes and therefore at the small scale in terms of software, hardware and financial resources, to point out to the indexes during their work how similar concepts have been treated in the past. For example, at LIR we have developed, and are continuing to develop, systems which analyse for the editors how previous treatment of particular cases and pieces of legislation differs from what is actually on the screen at the time. This is a relatively straightforward mailer of building indexes within indexes, and pointing out to an editor that in sixteen out of seventeen instances, such and such words have been used about this case. The current screen makes use of unusual terms - leaving the editor to decide whether this is because the treatment of the case is from a different viewpoint and that different terms are justified, or whether the indexing and abstracting is aberrant. In the ideal control system, the expert system which is available to help the human indexers is also available in explanatory form to the researcher (once one case has been found, its treatment and the terms used wherever else it appears in the indexes can be presented). In this way, our researcher who wants the tax aspects of Pepper v Hart will not be distracted by those documents which concentrate on issues of statutory interpretation.

Expert systems can also, and should also, apply the principles of "vocabularies". I have spent long enough in this paper on the issue of thesauri; here I would only suggest that a modern control system should be capable of using vocabularies at different levels. In the legal context, for example, the taxation expert will wish to use a specialised jargon which the property expert would find totally obscure. The two-level vocabulary - specialist and non-specialist - is a concept which is simple and which can be permitted. to some extent, to receive input from searchers as its own updating mechanism. The purpose of the two-level vocabulary remains the same as that of a thesaurus; to provide short-cuts into material that is specific in content, by means of the terms most familiar to the users.

Summary

I argue here that the size of the full-text databases in an electronic library provides its own difficulties for any hardware and software that is to make it accessible to the actual users, and that this problem is further complicated by the immense speed of technological change and the chance which that speed brings of spending a large amount of money to create an immediate fossil. I suggest that the solution to this problem is to keep the actual text in as simple an electronic form as possible, and to devise a control system, separate from the actual text, and interfacing with the users, which is based around conventional abstracting and indexing. The discussion of the possibility of performing indexing by software alone, without human intervention except during the writing of the programs, led me to conclude that this is not a practical alternative for all the reasons which make the attempt to search and index complete full text databases lack functionality and utility to the user.

By keeping the clever part of the job small in scale, the software and hardware costs are kept to a minimum. By using human brains for all the things at which it is good, as the principal input into the control system by means of abstracting and indexing, the control system will perform for the user and will be adaptable at the minimum cost in the future.

The alternative, integrated concept, which seeks to handle and make use of the entire bank of text, without the interposition of conventional, human-written indexes, seems likely to me to drown under the weight of its own material, and the result will be an unusable, expensive, and instantly obsolete behemoth. The difficulty is that such a system will work well in the early part of an electronic library project and will look wonderful at a small-scale demonstration; it is only as the library grows in size that the deficiencies will appear, and even larger sums of money will be spent to retrieve less and less that is useful.