# *The Electronic Law Library: Beware the Glitzy Fossil*

**Derek Sturdy**[1]
Legal Information Resources Ltd

How can an electronic law library be made useful to the user, the person for whom it is constructed and who, in many cases, has to pay for it? Doing something just because it is technologically feasible can waste prodigious sums of money, while doing simple technical things on the basis of properly applied human resources can remain the most effective, and much the cheapest, way to benefit the user Utility is the yardstick; however clever, a system which fails to benefit its users is a waste of time and resources This article discusses different approaches to the electronic library, condemning those ingenious and clever approaches which set out to do something just because it is technically possible, not because it is useful, and emphasising the increasing, not decreasing, role of the human brain in information management You might like to think of these as the "Look at me" or "Glitzy" approaches, contrasted with the "Plodder" approach which delivers the goods.

## *Electronic v Paper: has a consensus already emerged?*

One issue which need not detain us for long is the desirability of the electronic library; it is assumed that the electronic law library is the way forward but two general rules might be adduced here:

- People hate searching paper/
- People hate reading a screen for long From which follow:
- Search on a computer screen, then print out the retrieved full text, neatly and legibly, on decent paper, for actual full-text reading

## *Definitions and Scale*

The arguments presented here do not apply to a carefully constructed, modestly sized collection of homogeneous material (or material which has been edited into homogeneity) An individual CD-ROM database, has about 500-650 Mb of data which is all related in concept and subject and, probably, in format - the CD itself is only usable on a particular platform and is replaced, at regular intervals, by a successor This article considers a library of information which is much greater than this - an electronic library which is:

- vast, or intended to become so (>4 gigabytes to infinity)
- not at all homogeneous, but built up as material becomes available
- opportunist rather than structured

[1] Derek Sturdy is IT and Sales Director at Legal Information Resources Ltd. This is an edited version of a talk given at the 9th BILETA Conference The editor gratefully acknowledges the author's kind permission to reprint this article from (1994 Oct /Nov) *Computers and Law* 12-16

### *Control Systems*

All the data and text in the world is useless if it is not accessible to us on our own terms The thing that makes it accessible (or not, if you buy the wrong system) is the control system What do we mean by a control system for an electronic library? We should always consider, first and last, the information seeker approaching the terminal, who, in the case of the electronic law 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 (s 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 applying the criterion of 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

The control system is therefore a quite different matter from the expert system The expert system has its place, 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, 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 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

### *The Library Catalogue Analogy*

Knowing where to start was not a problem for a researcher 20 years ago - there was the catalogue, mighty volumes with pasted scraps of paper, card indexes with the heavy thumbing on particular cards pointing to the favourites of your colleagues, and you browsed happily and inefficiently through 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

The researcher had a more serious difficulty five or ten years ago The library catalogue was on a computer and was much more efficient in terms of retrieval 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 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 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 expertise, or nobody else will ever be able to use those sections
- Equally, don't let an engineer catalogue your law material, or a lawyer your scientific books
- 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 as valid today as twenty years ago. The aim of the control system, should be to allow a researcher to approach the system with the same confidence with which he approached the card index 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: the specific, jargon-ridden problem (as in s.740 liabilities) and the search for analogous material outside the sphere of expertise of the searcher. Different techniques are required, but the control system should not ask the researcher to define which way the search is to be conducted.

## *Full-text v Abstracting and Indexing*

This is still widely debated. Here is an argument which is commonly used, but which we believe is incorrect: given all the wonderful software we could write, who needs abstracting and indexing? You search for works, 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 online storage up comes the paragraph with the gist of your problem. You have to have the full-text anyway, otherwise how can you retrieve it when you want it; why bother with the fallible, human function of abstracting and indexing?

The argument is perfectly sound provided your database is small. 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.

## *False Drops*

Many will be 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's 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  With just 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  We suspect that the false drop is an increasingly serious problem, one which grows exponentially, one which was not perceived in the early days of full-text databases as an issue because it only appears with size, and one which is actually a matter of common-sense coupled with a generous measure of hindsight

## Human Indexing

The traditional answer to the false drop problem was the index produced by humans  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 indexing  We believe they are doomed to failure, and that the human brain in the present state of our knowledge remains the most effective, and much the most cost-effective, way to produce the required indexes

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

The obvious answer seemed to be hypertext linking  This concept has been practised by information people for years, by means of a thesaurus  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

## Keyword Use

An old-fashioned thesaurus is intentionally restrictive and mostly relates to concepts - which is why lawyers find it 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 work "keyword" and distorted its meaning; the most common misconception among database users is that, if your database contains keywords or makes use of a thesaurus, you can only search for keywords This arose because text retrieval manuals referred to the searcher's search terms as "keys" or "keywords" (after all, they had to call them something), lesser, or inadequate, text retrieval systems indexed only some of the data, those marked as "indexed keywords" or "manual indexed keys" and 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! Users then made a further logical jump, concluding that only full-text databases had full-text 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 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 the documents themselves and do not attempt to provide full-text indexing

### *Database Size Problems*

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 At this stage, the natural move is to place the actual text elsewhere (on CD or other optical media, or on removable hard disks) so the 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 being online 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

Hypertext is unsatisfactory in this context because, if it is to work as anything but a gimmick, it needs 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 online 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-ROM will be a nightmare on a 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  We believe 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

### SOFTWARE

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 because you want tried and tested components, but it is very far from being obsolete  You laugh to think how Y & Co bought that steam-powered system, and then the penny drops - that's you, in two years' time. The platform on which your software is being developed - let us say, for argument's sake, UNIX - was 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 programme has taken over the running  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

Your ingenious indexing software was written 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? Is your electronic library, in short, doomed to fossilise within weeks of commissioning?

### TEXT STORAGE

The method of storing text has to be considered  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 these pressures cannot help but subtly influence the 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

## ARCHIVING

We cannot read any of the media commonly used for data storage 25 years ago on any of our modern equipment Unless we have kept the old machines in a usable state, any data stored on the older media is irretrievably lost What chance is there that in 25 years 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? 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) onto the current standard at the time that the first signs of obsolescence occur This is an expensive and serious problem

## MEDIA TRANSLATION

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 (whether from one current medium to another or from one old word-processing system to a newer one) are aware 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, 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 of all sorts of different forms - which is why there is a strong temptation for a software house to make its data anything but flat The designer and maintainer of an electronic library 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 there to be any mention of how text is stored

## PRINCIPLES FOR AN ELECTRONIC LIBRARY CONTROL SYSTEM

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

The heavy computing work, and hence the use of both software and hardware, is concentrated on the index. The text itself is considered "passive". Apart from the actual storing of it, text is not processed or handled by the control system until it is actually called for Four principles emerge:

1 *Separate your index from full text*

This 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 can reside on a different platform, and each system does quite a different job.

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

If the document originated within your organisation, ensure that its format meets the current standard or translate it for long-term storage into flat ASCII or ANSI and remove all word-processing markups, database control characters and anything else which will be unreadable in ten years If the document originates from 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 article, 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)

Scanning is not yet, therefore, a technology which is obviously useful in the context of long-term storage of text in an electronic library. We suspect 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 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 to the system*

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 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 one above 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 managing system, which will handle copyright, printing, and charging issues

The structure that results from 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 full-text collection is the control system, the heart of which is the index or collection of indexes, which contains keywords, abstracts and pointers to the full-text documents Only the index is searched by the user All the software and processing power money is concentrated upon it, leaving only a simple and adequate document management system to retrieve any 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 for its ability to produce, quickly and simply, all its information 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 must 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 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 expended 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

### *Conclusion*

A system which seeks to handle and make use of the entire bank of text, without the interposition of conventional, human-written indexes, seems likely 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 ever larger sums of money will be spent to retrieve less and less that is useful

By following principles outlined above and by keeping the clever part of the job small in scale, the software and hardware costs are kept to a minimum By using the human brain 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.