9th BILETA Conference
The Changing Legal Information Environment

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

On-Line Electronic Libraries

Eve Wilson

Keywords: electronic data capture - information retrieval - hypertext - world wide web - multi-layered architecture - hybrid Systems.

Abstract: This paper looks at the ways in which:

i. Documents must be stored; this can be done through capturing original electronic documents imaging intelligent character recognition functions of an electronic library and the technologies that are currently available to implement them.

ii. Information may be retrieved from the stored documents by using Boolean query, collocates, term weighting, partial matching and document ranking, document vectors, and search by copying and search by example.

iii. On-line presentation is most elegantly achieved through a hypertext interface and navigation by browsing.

The paper considers the achievements of the Worid Wide Web and HTML in linklng diverse databases and argues that interface, application, and data should be three independent components in a multi-layered system architecture. It concludes by considering how the weak integration of the retrieval by query paradigm and the navigation by browsing paradigm might be achieved using hypertext systems as front-ends to information retrieval of database management Systems.

Information explosion

Lawyers need information and adequate means to handle it, simply to be able to compete in a market place that is increasingly information driven. Better communications and increased trade means that all the information a lawyer needs can no longer be held locally or even nationally, but exists in diverse locations throughout the world. This diversity helps to explain why new information technologies using incompatible data models continue to proliferate to produce the computer equivalent of a Tower of Babel. It does very little to help us deal with the mountain of recorded information, which doubles approximately every two years. Thus, the volume and complexity of information increases and with it, the users' expectations.

Functions of a Library

From the point of view of a user, a law librarian has three main tasks:

1. to store and maintain a collection of texts;

2. to locate, or help the user to locate, information locally and in other collections;
3. to give the user access to the retrieved material.

Let us consider each of these in turn:

Storage

Traditional libraries comprise collections of books, papers and manuscripts. These are bulky and therefore expensive to store. In recent years the use of computers in document preparation has ensured that it is the electronic form of a document that should be regarded as the generic form of the document and the printed form as a specific instance of it.

Unfortunately, too few documents are currently produced in a form which is truly independent of the purpose for which they are being used. Only too frequently document content is inextricably intertwined with formatting and layout commands. While these can be stripped out, to leave the bare content, this robs the document of much implicit structural information and makes the text less suitable for use in other processes such as information retrieval, hypertext and expert systems.

Ideally, then, all electronic documents should initially be produced to conform to the highest standards of an acceptable mark-up language such as the Standardised General Markup Language (SGML)1 [ISO 1986]. All documents should belong to a well understood document class which has been clearly defined in a Document Type Definition (DTD). This should contain all the structural and content tags that are needed for any application which can be anticipated. If there arises later an application which requires additional tagging, this should be incorporated in the original DTD, the relevant tags added to the original document and the whole checked again for conformity with the DTD. Only by adhering strictly to this method can we ensure the quality of the data, and avoid the proliferation of application specific forms that require as much updating and maintenance as the original. (The case for using SGML for law documents is explained at greater length in Wilson 1993.)

Today, the original is usually a document that has been carefully produced by experienced publishers to the highest possible standards evolved over many centuries for printed documents. Frequently, this original source is then taken for another application and stripped of all but the minimum of information needed to support the new application. The result is information retrieval systems that cannot produce an adequately formatted printed document or serve as a suitable basis for hypertext conversion, and hypertext systems incapable of supporting information retrieval and bereft of indexes and tables which added so much to the printed document. We must immediately begin to regard a document as an object which includes not only information, but all the functions that can act on that information. Any of these functions can be invoked by sending an appropriate message to the self-contained document object, but the object remains unique and independent of any task which invokes some specific instance of it.

Thus, where existing standards are observed, electronic storage for new documents ought not to be a problem; however, for documents which were not originally created as electronic documents, or for which the electronic form has been lost, electronic capture is more complex. If the original printed form of the document is significant and its appearance must be preserved, an obvious solution is imaging. When a document is imaged, it is scanned by a laser. The intensity of the scan is measured in dots per inch and can be varied: the higher the intensity the better the image. However, scanning is expensive in computer storage: a single side of A4 scanned at 300 dots per inch might require as much as 8 megabytes of storage. Compression techniques such as Tagged Image File Format (TIFF) might reduce this to 20-30 kilobytes, but to maintain a document collection of any size will require a mass storage device and enough capacity in the network to allow large documents to be shunted around without degradation of the service provided for other purposes. The costs of upgrading the computer system to cope with these extra demands must be added to the basic cost of document scanning (which is currently about lOp a page).

However, a document image is by no means as versatile as a document in machine readable form. It is an entity and the text it contains cannot be manipulated or searched; it has, therefore, to be manually indexed, in the same way as a traditional paper document, to provide a machine readable document surrogate, or representative, and it requires the same intellectual effort to process: fairly low grade skills for noting the bibliographic information: author/ originator, location, title, date, recipient, references; but high-grade effort to record the content of the document and its significance in relation to other material in the collection. There is, thus, no saving of effort over paper document indexing at the document entry stage and no improvement in search techniques unless the images are held in association with an intelligent character recognition (ICR) program.

ICR is a descendant of Optical Character Recognition (OCR) a technique performed by hardware made more sensitive and accurate by the incorporation into programs of character collocation and frequency tables to aid recognition. Character recognition can be performed on the original paper document, or it can be used in association with imaging to convert documents or portions of document only when required. The advantage of ICR is that the text can be manipulated, indexed automatically and used in full text information retrieval. The advantage of using imaging with ICR is that the appearance of the original document (including marginalia and signatures) is preserved. The disadvantage of using ICR is cost: to convert a single side of A4 using ICR may cost between £1-£3 per page. Currently it is far cheaper to have the information re-typed: with double entry for data validation, re-typing can be done for approximately 68 pence per page, a major saving when large collections are involved.

Retrieval

Document retrieval is, in many ways, a Cinderella of the computing industry. The techniques used in bibliographic or free-text information retrieval by most major commercial systems have changed little in thirty years. This may be partly because information retrieval was one of the earliest non-numerical applications. Just as FORTRAN has continued to dominate scientific computing because of its early adoption in the implementation of numerical algorithms for scientists, single word document indexing with Boolean search continues to dominate document retrieval techniques. Both examples are classic illustrations of the inertia that inhibits change once a substantial investment has been made in any technology or methodology. It is a reminder that when there is an advance in technology so significant that traditional methods are re-evaluated, modified or even changed, it is worth making every effort to get it right; otherwise we may have to live with the consequences for a very long time.

More sophisticated informational retrieval methodologies have been tried and proved in small scale projects:

- term weighting in the document representative and the query: i.e. attaching a weight to index terms, either manually or automatically, to indicate the relative significance of the term in the document relative to other index terms and allowing the user to weight terms in the query according to the importance he attaches to them, [Maron 1977];
- using collocates or pairs of words in the index instead of single terms: this increases index size but makes it easier for the user to refine his query;
- partial matching: where Boolean retrieval demands that the query be satisfied exactly, partial matching uses a set of query terms and will return documents if one or more terms from the query are matched; (it is usually combined with document ranking;)
- document ranking: the returned set of documents are ordered by the system in order of computed relevance to the query and viable in any system using term weighting or partial matching;
- document vectors: weighted or unweighted document representatives and queries can be

considered as vectors in an n-dimensional document space (when n is the total number of index terms), where the angle between the document vector and the query gives a measure of the similarity between them, [Raghaven 1987];
- searching by copying: the user is shown examples of previous searches and the result sets obtained by them to help him devise his own query.

All the methods listed above, including Boolean query can be used whether the document database consists of full text or merely document representatives containing abstracted information from documents that are held manually or as electronic images. If the document representatives are concise with well defined fields, the retrieval can be handled by a conventional relational database management system (DBMS), such as Ingres or Oracle. Such systems were originally devised to handle commercial data and some kinds of scientific data that could easily be tabulated. With honourable exceptions, e.g. Spires, DBMSs are not suited to free-text data, and other systems have been specially developed to handle this. These rely predominantly on inverted word indexes into a tagged database.

Some retrieval methods that demand full text have only become tenable with the improvement in technology. For example, query by example or find another document like this one (but different from that one). These methods demand in-depth vocabulary comparisons and may become more popular as parallel processing techniques make complex multi-term search algorithms feasible for on-line work, [Stanfill 1986].

Presentation

Presentation is the area of electronic document management where computing techniques were slowest to impinge but where they are now rapidly making up for lost time. Early systems for electronic search and retrieval often returned merely a list of possibly relevant documents; these the hapless user had to assemble manually and then scan by eye to confirm or reject. It is hardly surprising that Futility Point Criterion (FCP) was a critical factor in search evaluation; (FCP is the maximum number of documents that can be returned by query before the search becomes pointless because the user cannot process the return set). Later systems with full text have tended to retrieve documents as independent entities: a user who wished to retrieve another document which had been cross-referenced in a document already retrieved, had to submit another search request to the system, possibly specifying a different database. The retrieved documents had frequently been stripped of much of their structure, formatting and the variety of fonts that rendered the printed version so much more attractive. This last is not a minor point. Comprehension and proof reading tests have shown that a reader's performance with printed material is 25-30% better than his performance when using a display screen, [Moskel 1984].

This is a serious indictment of screens as quality purveyors of information. It makes it imperative that every effort is made to ensure that screen presentations are at least as visually attractive as their paper forerunners. Even so, we may have to accept that people who need to make detailed and intensive study of long documents may want a printed copy. If an electronic system is working as it ought to work, it should be possible to print on demand any document, or section thereof, to the same quality and format that a commercial publisher would attain in book form.

Thus, what has effected a revolution in our acceptance of the computer screen as a device for presenting information stems not from the visual aesthetics of the medium but from its use as a medium for interactive browsing through a genre of computer systems known as hypertexts. With a hypertext system, information is regarded as a collection of nodes that are interconnected by links. Each information node can contain one or more link anchors which, when selected by the reader using a mouse, will display the target node for that anchor. These concepts of nodes and links are the only defining characteristics for hypertext: there is no internationally accepted standard or even a set of minimum features that a system must meet before it can be called hypertext. This is unfortunate, because the lack of such standards allows much scope for diversity and informality and obscures the clear relationship that there ought to be between the provision of an independent document type definition and the subsequent successful conversion of a document of that type into hypertext. A good hypertext system should support, not only links between individual nodes, but the logical and hierarchical structuring of aggregations of nodes. It should also be able to work on data which is in an independent standard form. Ideally this should mean that it is possible to provide a mapping between a DTD source and any target hypertext system.

HTML and World Wide Web

A step in this direction has been provided by the HyperText Markup Language (HTML). HTML is used to define a World Wide Web (WWW) document. (Note that again it is a way of making a document suitable for a specific application not a means of ensuring document independence). HTML will support a hierarchical structure of up to six levels and links to other documents. Other structural components supported include lists (unnumbered, numbered descriptive and nested) quotations, and preformatted text, i.e text in which line composition should not be changed. A minimal HTML document consists of a title, a level one heading and at least one paragraph.

<TITLE> Electronic Libraries </TITLE> <H1> Purposes of a Library </H1>

From the point of view of a user a law librarian has three main tasks:

<OL> <L1> to store and maintain a collection of texts; <U> to locate locally and in other collections; <L3> to give the user access to the retrieved material. </OL>

Let us consider each of these in turn. </P>

The WWW browsers use the structural tags for formatting: headings are displayed in a different font or colour from the main body of text. However, the real strength of HTML lies in its capacity to link a document to other documents, parts of documents or images. A link to another document is enclosed in anchor tags <A> and </A>. The anchor start tag must include a HREF attribute to specify the target file for the anchor.

The target file may be a file held locally or remotely. The format used is Uniform Resource Locater (URL), a draft standard for specifying objects on the Internet. A file reference has the following components:

scheme//host, domain[:port]/path/filename

where scheme may be: *http* (Hyper Text Transfer Protocol used by WWW servers) or non-http servers that WWW browsers can access (such as *gopher, WAIS, telnet and news).* The use of WWW is growing rapidly: usage of the CERN server doubles every four months. However, we must be aware that this is a partial and imperfect solution: a DTD to fit a document to an application not to maintain document independence.

The real power of WWW is that it demonstrates how easy it now is to link disparate databases. Distributed systems are useful because:

i. ease of access means that less information need be held locally at any one site, i.e. no-one need duplicate data merely to have access to it;

ii. where constant access is a requirement of the system, copies of essential data can be copied to other sites so that if one server fails users can continue to access the data without interruption.

Achieving task-oriented systems

Open systems and multi-layered architecture

So far in the development of information systems it has proved easier to establish the principles of open systems and client-server architecture than to establish the principle of data independence from application. WWW and similar servers have shown that an application can be happily divorced from its interface. For example, WWW can be browsed using at least the following graphical user interfaces

IXMosaic using X11/Motif

Macintosh Browser

Cello (for PC/Windows)

Browser-Editor for NeXTStep

There are also terminal based browsers. Thus, we already have a two-layered architecture.

However, all too frequently data is still formatted to suit the current application, and applications remain free-standing and independent of one another. This cannot be allowed to continue: applications must be able to talk to one another. It is a necessary prerequisite for task-based systems and can only be realised if the data is common to all applications. The proliferation of different data models and formats to suit the idiosyncratic needs of individual applications is wasteful. All tagging for content and structure is value-added to the information and all must be preserved in a master datafile. Applications might filter or ignore tags that they do not need but those tags must not be removed to accommodate the application.

Unfortunately, it is all too easy to remove information automatically: providers of information retrieval systems have, in the past, frequently deleted from their databases structural markers that they do not use; they have even removed content such as tables and images that were difficult to process or which played no part in the search process. Once such information has been removed, it is often lost for ever unless the text is manually edited. Automated reconstruction and restoration of a deformed and depleted text is never an easy option and all too frequently not even a viable option. We urgently need a layered architecture:

1. Interfaces
2. Applications
3. Data

Integrating applications

The most important service that a law library can provide for the user is to ensure the integration of related applications. In particular, information search and retrieval must be integrated with document browsing, and the provision of high quality printout on demand. This would not be hard to achieve: current systems are complementary, not incompatible. Information retrieval systems and relational database management systems are characterised by their style of data management: hypertext presentation programs by their style of user interface and browsing navigation paradigm. The information needed to produce a visually attractive hypertext display is identical with that needed for a high quality printed document. There is no reason why they cannot co-operate as long as the conceptual organisation of the data is identical. Let us consider this in slightly more detail.

Complementary systems

Information retrieval systems and database management systems are software packages for the management of data on secondary storage. They are efficient in:

1. data organisation, usually through indexes into the main data, and the manipulation of homogeneous data sets;

2. data entry update management;

3. fast search and retrieval by relational operators or query language;

4. multi-user access including handling problems of concurrency, privilege and priority in a high volume transaction system;

5. backup and recovery utilities.

The disadvantage of these traditional systems is that it is often difficult to accommodate new information with a different conceptual structure and most current systems are poor at handling multimedia.

In contrast, hypertext Systems tend to be good at:

1. non-linear browsing through heterogeneously formatted or non-tabular materials;

2. selection by pointing devices, and through graphical overviews;

3. navigation by browsing;

4. integration of multiple media (text, graphics, sound and video).

However, facilities in hypertexts for direct querying are usually limited, based on pattern matching and, consequently, slow. Other activities not usually noted for efficiency include managing large volumes of data and handling multiple users and concurrent transactions.

Weak integration

There are two ways of achieving integration. The optimum long term solution is to integrate the best of information retrieval, database management and hypertext traditions in a single object oriented system, but this must wait the next generation of computer software. A short term option is Iweak integration, where hypertext systems serve as front-ends for DBMSs or information retrieval systems. The user perceives the information as nodes and links, and browsing as the basic access paradigm; but the interface incorporates the ability to use the query language of the underlying direct retrieval system. This is especially useful for finding a starting point, or set of starting points, for a navigation.

Bibliography

ISO (1986). Information Processing Text and Office Information Systems Standard Generalised Markup Language (SGML) 150 8879 (E).

Maron. ME (1977). "On indexing retrieval and the meaning of about". *Journal of the American Society of Information Science* 28.1 pp 28-43.

Moskel, S et al (1984). *Proofreading and comprehension of text on screens and paper,* University of Maryland Computer Science Technical Report.

NCSA, (1987) A Beginner's Guide to HTML pub@ncsa.ufuc.edu. Raghavan & Deogun "Optimal Determination of User-Oriented Clusters" *Proceedings of ACMISIGIP* pp 140-146.

Stanfill, C & Kahle, B. (1986). "Parallel Free-text Search on the Connection Machine System *Communications of the ACM* 29:12 pp 1229-1239

Wilson, E (1993). "The Case for SGML: a Law Database, Hypertext and Information Retrieval" in *International Yearbook of Law Computers and Technology* V 7 Edited by K Russell, Journals Oxford Ltd, pp 59-75.