Paper mate

The capture and management of digitised text is big business, and businesses of all sizes can benefit from an efficient system. Roger Gann scans the options.

Illustration by Cyrus Deboo

Contents

Once upon time, bold claims were made about computers, the future and the 'paperless office'. A swift reality check confirms that computers, far from reducing the amount of paper that clutters up offices, has actually increased it. These days, thanks to fast printers and copiers, producing a high-quality document on paper has never been easier. And that's despite the enormous increase in the volume of documents that never see paper, such as email.

One solution to this ocean of paper that threatens to drown us in our offices is document management. The only problem is that this term is a movable feast and has a multitude of meanings. For this feature, I'm taking the term to refer to one of the many highly evolved software systems for organising, storing and retrieving complex digitised documents.

Whatever it is, it's big business. Spending on document management products is predicted to hit $33 billion by 2002, according to a study conducted jointly by the industry body responsible for document management, the Association for Information and Image Management International (AIIM), and IDC.

The aim of document management software is simple: to reduce the workload and improve the productivity of most offices. An Ernst and Young study indicated that electronic document management can triple processing capacity, cut staff work time by up to 50 percent, provide immediate access to decision-critical data, cut document storage space by up to 80 percent and provide fail-safe, secure systems. Document management software technology has changed beyond recognition in recent years. All document management systems capture and manage digitised text and other images, but recently developers have added to the mix specialised techniques for image capture, workflow management, text mining and formatting. Today's document management packages often include document imaging, file management, workflow management, computer output to laser disk (COLD) processes, web page publishing, forms processing, text mining and even text formatting. Highly integrated web systems are beginning to dominate the market.

Optical Character Recognition (OCR) is part and parcel of the document management solution. This process turns a digitised image of a document, essentially a picture, into editable text. More importantly, it permits searching of the contents of a document, not merely for its filename. So for example, OCR would allow you to turn an incoming fax, say a draft contract, into an editable word processor document, neatly avoiding a heap of needless re-keying. It could then be filed away electronically, to be retrieved and printed at the click of a button.

The aim of document management software is to reduce the workload of most offices

OCR technology has been in the PC domain now for many years now and while basic recognition rates are dangerously close to perfect given a good clean original, most developments in OCR software have concentrated on the thornier OCR issues such as format retention and coping with very poor original documents. Leading edge OCR technology is still to be found in commerce, with organisations such as the Royal Mail and the clearing banks devoting much effort to mastering handwriting recognition.

Just recently, the vagueness over the precise meaning of document management has been compounded by the rise of a new buzzword (or two): knowledge management, which is loosely defined as that part of information technology that seeks to give organisations stronger, more thorough control over not just information, but the way it's handled over increasingly complex networks. Examples of the genre include collaborative groupware, such as Notes or GroupWise. Eventually, given the upcoming developments in document management and especially the rise of XML, the distinction between these two standalone technologies will become very blurred indeed, prompted by the advent of the internet and intranets as universal platforms for the propagation of electronic documents and processes.

How OCR works

All OCR software works in a similar manner. After a scanner inputs a paper document, the OCR program analyses the graphic image of the document, breaks it up in a number of zones (i.e. text, numbers or graphics), identifies the letters and numbers and converts that image into a text file. This done, the software refers to its internal dictionaries to compare the words it has found with known ones, and corrects them, just like a spell-checker. It then identifies words it doesn't recognise and shows you the original scan of that word, enabling you to manually interpret the unrecognised word. Finally, the OCR program saves converted documents in most popular word processor formats. The end result is text that you can edit and incorporate into letters, reports, newsletters and even web pages. In fact, most current OCR packages can convert documents directly into HTML.

OCR can probably never deliver 100 percent recognition accuracy and while state-of-the-art packages get pretty close to this, there's still some room for error. Much hinges on the quality of the type to be scanned. For example, it would be a poor OCR package that couldn't achieve extremely high recognition scores when recognising a clean page of laser printed 12-point Times Roman text. It's the ability to cope with the more complex material, the poor quality photocopies and faxes and complicated layouts, that separates the OCR winners from the losers.

Small office **OCR**

Document management for the user with modest means.

The SOHO user can enter the rarefied world of OCR and document management for a very low admission fee. At most they'll need a cheap flatbed scanner plus some inexpensive document management software. While the very cheapest scanners are not as capable as even those with £100 price tags, they're still more than capable of acquitting themselves when it comes to document management and OCR tasks.

And the scanner's price tag also includes most of the software you'll need to get started. Most flatbed scanners ship as standard with a raft of stripped-down 'lite' image editing and OCR software and so are OCR-capable out of the box. For the SOHO user, their OCR capabilities are generally adequate. If you want more features and functionality, then these 'lite' OCR packages count as upgrade fodder, letting you purchase an otherwise expensive package like OmniPage for a wallet-friendly £79 (inc VAT) instead of a rather crisp retail price of £465 (inc VAT), which includes the entry-level PageKeeper Standard document management package.

What about document management for the SOHO user? Well, the biggest document management problem, disk storage capacity, is no longer an issue; current PCs routinely ship with 12 or 16Gb hard disks plus a plethora of removeable storage options. So if you have a lot of paper to manage, putting it on disk is more than feasible.

If you have a lot of documents to archive, a simple flatbed scanner that lacks an ADF (Automatic Document Feed) will make the scanning job a chore, and you may need to spend a couple of hundred quid to get a scanner that can take an ADF. However, the upside is attractive: on top of the convenience in having every document to hand, you can gain some living space by ditching the filing cabinet. Clearly some important documents have to be retained as originals, but most don't and so are prime candidates for the cull.

Finally, a digital filing system gives you the ability to index and search for documents electronically. Although making insertions and revisions to existing documents can be tricky, the time you save by not having to rifle through mountains of paperwork will make digitising your important documents a worthwhile office project.

The bottom line here is that is that the SOHO user with modest requirements can get a quite reasonable OCR and document management system for not a great deal of money, perhaps less than £200 all in. Xerox's budget-priced Pagis ScanWorks is a one-stop SOHO document management solution, combining scanning, photo-editing, OCR, electronic forms and document management all for around £30 ex VAT, which makes it pretty good value for money — you get MGI PhotoSuite, the Textbridge Classic OCR engine. Document management is rudimentary, however: text searching isn't available.

OCR for beginners

Most scanners ship with basic OCR packages, so-called 'LE' or 'Lite' versions of full-feature, full-price retail versions. They still feature often complex interfaces, however, and can be difficult to use, especially for the first time user.

A new product from Caere, OmniPage Wizard, is designed to address these problems. This is a no-frills OCR package aimed at the first-time user. It uses the familiar wizard program structure to lead the user through the five steps needed to OCR a document. However, it eschews the usual Windows 98 wizard look and replaces it with a completely new web-like interface comprised of large graphical icons. The only time the lush interface disappears is during the proofing stage, when the thoroughly normal OmniPage proofing dialogue pops up.

◀DOES THIS LOOK LIKE ANY OCR PACKAGE YOU'VE SEEN BEFORE? AS YOU CAN SEE, THE EMPHASIS IS ON SIMPLICITY AND EASE OF USE

Priced at £40 including VAT, OmniPage Wizard doesn't feature OmniPage's most recent recognition technology but claims to better that which is typically bundled with a scanner.

Many OmniPage features have been removed, such as zoning and training, purely to keep it all simple. Comparing it to the full-blown OmniPage Pro, its overall accuracy is only slightly lower, and it's only when you present it with poor originals does the difference between the two products show.

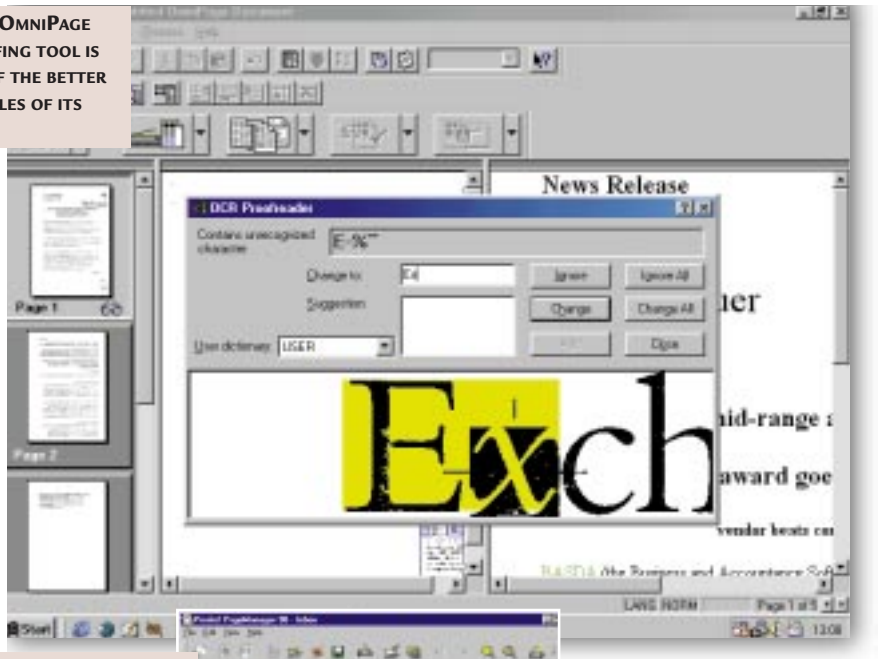
PCW DETAILS

★★★★

OmniPage Wizard
Price *£40 (£34.04 ex VAT)*
Contact *Caere UK 0171 233 6677*
www.caere.co.uk
Good Points *Simple to use. Moderate accuracy.*
Bad Points *For another £40 you can get the real thing.*
Conclusion *A genuine attempt to demystify the OCR process that succeeds.*

Small **business setup**

OCR technology can greatly benefit the larger office.



▶THE OMNIPAGE PROOFING TOOL IS ONE OF THE BETTER EXAMPLES OF ITS KIND

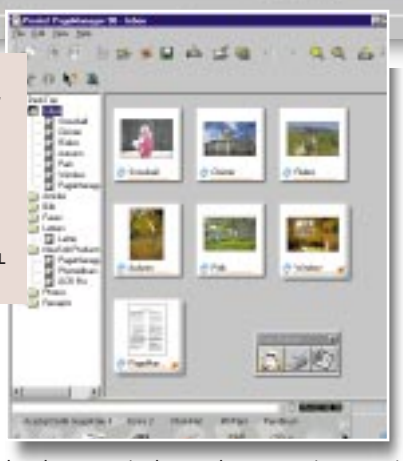
While OCR and document management may be handy for the SOHO user, larger businesses stand to reap more substantial benefits from these two technologies. You'll need a reasonably specified PC attached to a network, plus the desired OCR and document management packages. Probably the biggest difference lies in the scanner. This need not have a particularly better optical specification but it does need to be more rugged. Throughput will be crucial so the scanning speed becomes paramount, as will a SCSI or USB interface. It will almost certainly have to have an automatic document feed, an add-on that can cost more than the scanner.

At its most basic, Windows Explorer works as a rudimentary document management system. The only problem is that you can only search on the filename and not on the contents of that file. The answer here is to integrate OCR with the scanning and filing functions. For example, **Caere PageKeeper Standard** uses the OmniPage 7.0 OCR engine to sweep and index the words in the text pages you scan. The end result is an image file that you can then send to a full-blown OCR package like OmniPage Pro for OCR. But because PageKeeper Standard has already performed background OCR, you can still search on any text the image contains.

The latest version of **Presto! PageManager** has an improved OCR engine which creates a text version of a scanned document that preserves the document's original formatting. It can also output scanned documents as RTF or HTML files. It provides a wide range of annotation tools, including arrows, highlights, stamps and sticky notes.

While PageManager 98 supports Boolean and fuzzy searches, the search tool isn't easy to use. When you open one of the documents located in a search, PageManager provides no indication of the target text's location and no tools for locating it while viewing the document. Its handling of graphics images is better, though.

Not strictly in the document management field, **Enfish Tracker Pro** is a very interesting new personal information management package designed to let users index and locate data from a wide variety of sources on their hard disks. It can track everything, from the contents of email and common desktop applications, to the pages on user-specified web sites. It uses the Inso QuickView Plus viewer and so can index the text inside virtually any spreadsheet program or word-processing document.

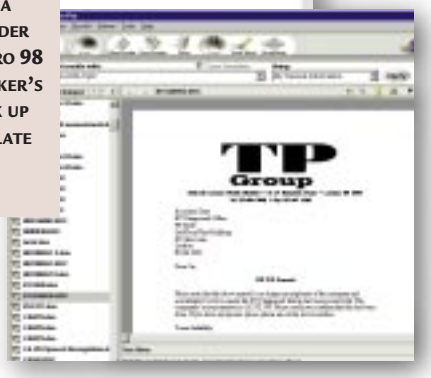


▶LIKE OTHER LOW-END DOCUMENT MANAGERS, PAGEMANAGER DOUBLES AS A PHOTO-ALBUM ORGANISER AS WELL

While document management is arguably best performed in a networking context, the opposite is probably true for OCR packages. All the current best of breed OCR packages are standalone packages, though most mesh well with networks. OCR packages have tended to be expensive software products — even today, OmniPage Pro 9.0 has an RRP of £395 (ex VAT) — but thanks to 'competitive upgrade' pricing, you shouldn't have to pay more than £80 (inc VAT) for a state-of-the-art OCR package, so they are very affordable, both for individuals and companies.

Version 9.0 of the doyen of OCR packages, OmniPage Pro, builds on the strengths of the previous version and somehow manages to increase its already impressive recognition accuracy, getting it perilously close to 100 percent in a few cases. The new version also features better support for colour, improved handling of table objects and spreadsheets, and improved format retention. A copy of PageKeeper Standard is thrown in for good measure. OmniPage Pro can operate either as a standalone application or from within most Windows word processors, thanks to its OCR Aware feature.

If you have a lot of OCR-ing to perform, it's also possible to batch acquire the images, either from disk or a scanner, and then schedule recognition at a later time, perhaps overnight, which is useful in office scenarios. Couple this with top-scoring accuracy scores, and with



▲ PROOFING IS A SIMPLE TASK UNDER TEXTBRIDGE PRO 98 ▶ ENFISH TRACKER'S ABILITY TO PICK UP DETAIL AND RELATE ITEMS IS QUITE UNCANNY

OmniPage Pro you have arguably the best OCR on a PC.

TextBridge too has consistently improved over the years. While it has always been cheaper than OmniPage Pro, it's always had to settle for second place, which is a pity for it offers strong competition. It's very simple to use, with a rather stark interface comprising just three large buttons plus a short Word/Office-style toolbar above it. The first button initiates an automatic wizard, while the other two allow you to acquire a page and recognise it, respectively.

As with OmniPage Pro, **TextBridge Pro98** goes a long way towards total automation of the entire OCR process. You can either choose to have the program perform the whole operation for you, or you can walk through the individual steps on your own. Application integration remains superior with TBP, and its Instant Access feature ties it in to all three major office suites, not just Office 97.

TextBridge Pro98 has no trouble correctly identifying both the text and graphics zones in a document. You can choose to zone a document manually if you want, but in general the auto-zoning is of a high order and fine-tuning should be unnecessary.

The results from TextBridge were similar to Presto!: it recognises most characters and it understands when a font is italic, bold, or both. Also like Presto!, TextBridge converts all coloured text to black and instead uses colours to highlight recognition problems; images are converted to black and white, too. **Presto!** OCR also supports batch scanning, useful when you have a pile of documents to scan and recognition can be deferred to a less busy time. Another potentially useful (and unique) facility is load-balancing: if you want to use older, slower PCs, you can have one PC scan the batch and another on the network recognise the resulting scanned pages.

Presto! OCR works in much the same way as its rivals — when a page is scanned and interpreted, Presto! OCR displays an image window after the scan to show the scanned page and the recognised zones to interpret. Recognition can then take place, and the result is presented for proofing in another window.

Tables are recognised as such and can be stored for exporting to a database as well as part of the document where you'll save the resulting text. Presto! has a dictionary for spell-checking, which is used only after the recognition engine is confident that it has recognised the word; the dictionary is not used to help improve recognition, but to help the user figure out whether the recognition engine correctly identified a word.

Presto!'s recognition scores were good but not on a par with OmniPage Pro, and its table handling and format retention capabilities were much weaker. File Save As options are also very limited, to RTF, HTML and ASCII text.

Overall, Presto! OCR is a competent OCR package, but with the much better OmniPage Pro available on upgrade for only a tenner more, it's hard to recommend Presto! OCR 3.0.

PCW DETAILS

★★★

Caere PageKeeper Standard
Price *£40 (£34.04 ex VAT)*
Contact *Caere UK 0171 233 6677 www.caere.com*
Good Points *Indexes text from scanned files using background OCR. Automatic indexing.*
Bad Points *Limited scanner settings; doesn't include full-blown OCR capability.*
Conclusion *Lacks power overall. Limited search facilities.*

★★★

Presto PageManager 98 Gold Edition
Price *£50 (£42.55 ex VAT)*
Contact *Guildsoft Computer Software 01752 895100 www.guildsoft.co.uk*
Good Points *Good one-stop document management solution.*
Bad Points *Searches aren't ranked.*
Conclusion *Search facilities not easy to use.*

★★★★

Enfish Tracker Pro
Price *£58.74 (£49.99 ex VAT)*
Contact *Roderick Manhattan Group 0181 875 4444 www.enfish.com*
Good Points *Powerful. Easy to use. Highly customisable*
Bad Points *Resource hungry.*
Conclusion *Sophisticated search tool.*

★★★★★

Caere OmniPage Pro 9.0
Price *£464.13 (£395 ex VAT). Upgrade £80 (£68.05 ex VAT)*
Contact *Caere UK 0171 233 6677 www.caere.com*
Good Points *Improved accuracy. Better table support.*
Bad Points *Expensive RRP. Still some way to go on format retention.*
Conclusion *Probably the best OCR you can buy.*

★★★★

ScanSoft TextBridge Pro98
Price *£81.08 (£69 ex VAT)*
Contact *ScanSoft 0118 966 8421 www.scansoft.com*
Good Points *Good recognition scores. Handles poor originals well.*
Bad Points *Weak training and zoning tools. No batch facilities.*
Conclusion *A good alternative to OmniPage Pro.*

★★★

Presto! OCR Pro 3.0
Price *£82.25 (£70 ex VAT)*
Contact *Guildsoft 01752 895100 www.guildsoft.co.uk*
Good Points *Batch processing.*
Bad Points *Weak format retention. Limited range of file formats supported.*
Conclusion *Competent, but not outstanding.*

Large **office setup**

Solutions for heavy-duty corporate publishing requirements.

Controlling documentation, particularly product information, in the larger company is not a trivial task. However, there are a wide range of heavy-duty client-server document management solutions available, such as Folio and Documentum. XML-based solutions are also beginning to surface.

ArborText's EPIC (Enterprise Product Information Chain) comprises a package of software and services for implementing XML-based corporate publishing solutions at the enterprise level. EPIC's goal is to reduce the time it takes to publish product-critical information in highly-competitive industries, where reducing the time it takes to bring products to market is crucial, e.g. car or aircraft manufacture. Because it offers beginning-to-end support for the creation, revision and publishing of complex documents, EPIC avoids many format and translation problems as documents are passed up and down the information chain.

Then there are the groupware solutions, which bridge the gap between the knowledge and document management fields. Three familiar names compete in this area, IBM/Lotus (with Notes/Domino), Novell (with GroupWise) and to a lesser extent Microsoft (with Exchange Server). These can have document handling capabilities bolted-on. **Domino.doc 2.0** is a web-based add-on for Domino that features document life-cycle management, rules-based workflow and archiving capabilities. It uses a familiar file cabinet and folder metaphor. File cabinets, or Notes database files, use Notes replication services to distribute, organise and manage documents, and related documents can be grouped together in a binder. Its Storage Manager is a back-end utility that lets administrators store documents offline in auxiliary archives, like optical disks or tapes.

By contrast, GroupWise offers full document management in its standard configuration, alongside messaging and scheduling. Users can store all types of documents and open them in all types of applications, and import individual documents or entire directory structures of documents into GroupWise libraries previously created by the administrator. Users can share individual documents with other users, or they can send email to other users and include document references that other users can save then access later. All documents may have multiple versions; and, as users update documents, they have the option to re-store them as new versions or simply update the original.

While third-party document management bolt-ons for **Exchange Server** are available, Microsoft has been a late entrant in this market and it has only recently announced future document management products. 'Tahoe' will succeed Site Server 3.0, which is currently part of the BackOffice family of server applications. This upgrade will provide document management and search capabilities, and will include approval workflow, templated publishing, document versioning, XML support for indexing documents and natural language processing, according to US reports. Tahoe will use a portal as a kind of 'knowledge desktop', allowing developers to build their own intranet portals, as well as use Tahoe's search and document management services within those portals. Another upcoming product, Polar Server, will focus on aspects of knowledge management such as document tracking, collaboration and analysis.

Production line scanning

To process serious quantities of documents, a conventional flatbed scanner, even one assisted by an ADF, really isn't up to the job. You need a dedicated, high-volume scanner. A number of companies produce heavy-duty scanners, including Xerox and Kodak and, of course, **Bell & Howell**. Its **Copiscan 4040D** is aimed at the mid-range market, at the workgroup level, and can handle 80 sides/40 pages per minute, or up to 3,000 documents a day. It carries a suitably serious price-tag — £4,500 plus VAT.

Externally, the 4040D resembles an old-fashioned laser printer. It's heavy, all-metal construction confirms its durability for high-volume scanning. As you'd expect, the device has a SCSI II interface. It uses contact image sensors (CIS) on both sides of the path to scan both sides of a document in one go. No TWAIN driver is supplied, only ISIS, but the 4040D works under Windows 3.1x, Windows 9x and Windows NT 4.0. It uses a straight-through paper path: paper is stacked at the front, and the feed can hold between 100 and 500 sheets, depending on paper thickness. Its maximum speed is only achieved in duplex mode at 200dpi. If you want 300dpi, performance drops a bit. A wide range of paper types, including NCR and thermal fax, are accommodated.

As the scanner can serve up images to an OCR program much faster than the most packages can handle, the 4040D has a deferred OCR option. Its ability to handle mixed types of documents was good, but despite my best tweaking efforts, it would still occasionally draw several pages through at once, so its paper-handling wasn't perfect, not at this price.

PCW DETAILS

★★★

Price *£5,287.50 (£4,500 ex VAT)*
Contact *Bell & Howell*
0800 783 8050
www.bellhowell.co.uk

Web solutions

Large-scale document management via the web and using the XML standard.

If many employees use remote access applications, web-based document management makes good sense because distributed client/server systems are harder to maintain than web servers. Furthermore, an inter/intranet-based system is client platform independent and provides cross-platform support. It's also scalable, not only in the enterprise but beyond it, too. A web-based system can also allow document updates in real time and get payback in less time, because more users can access the application as most end-users already have web browsers.

As a result, enterprise-wide electronic publishing over intranets is beginning to replace the traditional departmental approach. Increasingly, large corporations are turning to electronic publishing tools from companies such as Enigma, Folio and Xyvision. Ultimately web-based electronic publishing will provide companies with the ability to tie together data from separate repositories. It will be possible to create a single master table of content from multiple publications and sources on the fly. But for now, most data repositories are proprietary and not standards-based.

Help is on the way, however, in the shape of XML. The advantage of the XML standard is that it can assemble text and data from a variety of sources. That way, a user can create a complex electronic document that can still be easily searched or sent over the internet, typically using HTTP.

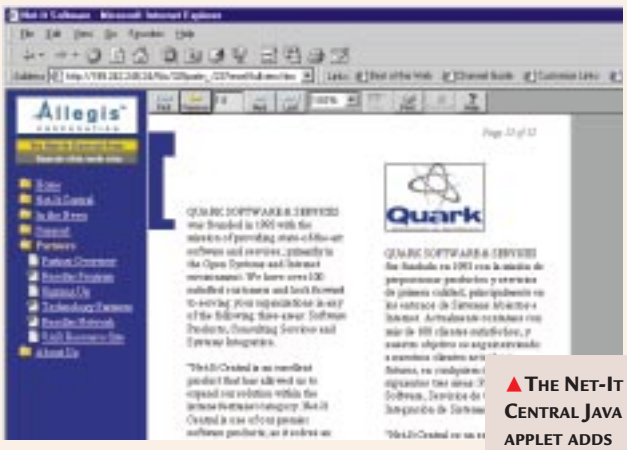
Take Insight, a document management system that includes a proprietary relational database that can drill down into nested levels of detail. For instance, Insight is deployed at the maintenance division of London Underground. A maintenance engineer can view a graphical breakdown of the parts of the train, drill down on each component: a click on the wheel pops up a graphic of the axle, which in turn can take the engineer down to component parts and identification numbers. A hyperlink can be set up to take the user into the warehouse to see if a part is available, or into instructional documentation.

The XML standard promises to revolutionise document management for the large-scale enterprise. As a result of ever-increasing regulation, most are obliged to commit large amounts of data to paper, in the form of procedure guides, products guides or maintenance manuals. Some companies produce such complex products, e.g. aircraft manufacturers, that the production of accurate manuals, which could contain hundreds of thousands of components, is a very real headache. They're also expected to keep up to date with a welter of government regulations, all of which are delivered on paper. Maintaining such a mountain of paper, keeping it current and doing it in a timely fashion, and remaini competitive, is no easy task.

Low-end web publishing solutions

For smaller companies there is a simpler way of publishing documents on their intranet or company web site. Most word processors and several OCR packages include a 'save as HTML' option, but for large-scale document management there are more efficient solutions.

Adobe Acrobat is a universal page description and viewing format found mainly on the web and yet it makes no use of HTML. It converts documents into a cross-platform portable document format (PDF) that gives users more control in page design. Your pages can feature complex text formatting with different fonts, something that's awkward if not impossible using HTML. To republish work that you've already output using a DTP package like Quark XPress, then producing a PDF file takes just a matter of seconds, using Distiller. You can add hypertext links to connect documents within your web site. A 'capture' module lets you scan and OCR straight to PDF. It's available for all three Windows formats, plus (Power) MacOS and some Unix flavours. **Acrobat Reader**, a basic viewer, is available for free download at www.adobe.com and is on the *PCW* cover CD.



▲ THE NET-IT CENTRAL JAVA APPLET ADDS NAVIGATION AND PRINTER CONTROLS TO THE TOP OF THE DOCUMENT WINDOW

Net-It Central 3.0 www.net-it.com allows small workgroups to contribute documents to a corporate intranet or extranet. It automatically converts documents created with Windows applications into Java applets, and creates a web site to display those documents. You set up one or more 'drop boxes' on a server and simply drop the document in the box. If you set up a hierarchy of drop boxes, this structure will be retained on the intranet. Conversion is more than adequate for corporate use, but doesn't offer the precision of Adobe Acrobat.