

Scanning and OCR

Keeping your image clean

By Ross L. Kodner

Scanning has been a sore spot in law firms for more than two decades. Why? Lawyers have viewed scanning as synonymous with optical character recognition (OCR). The problem? Even with the best OCR products, results often fall short. Many documents are not good candidates for recognition. Without a clean laser-printed source document you'll end up with gobbledygook. Your staff will tell you it would have been faster to type the document than to OCR it and clean up the resulting mess.

Instead, view scanning as a way to turn *physical paper* into *digital paper* — like photocopying the documents onto one's computer screen. When scanning as images, the process can be 20 times faster than the processing-intensive OCR approach. Further, imaged documents on screen look *precisely* like the originals: handwriting, pre-printed lines/boxes — all scan perfectly. This is a core part of the concept I call the PaperLESS Office™.

Lawyers and their staff universally have one thing in common: They are buried in an unending sea of paper. Pleadings. Correspondence. Briefs. Exhibits. Memos. Pink phone message slips. Sticky notes. You name it, paper is everywhere, chok-

ing and clogging the flow of work in both private and public law practices. Sometimes getting client work out is more an issue of managing mounds of paper than applying legal brilliance. So is there any hope at the end of the paper-lined tunnel? Yes!

For years, lawyers have been on a holy quest for the mythical and fabled "paperless office." This endlessly elusive concept is likely the greatest lie of the technology age. We're never going to be paperless, at least in the foreseeable future. We need to accept the fact that even if we reduce the amount of paper we generate, others will continue to send us paper.

Microfiche was supposed to be the answer. But microfiche isn't often used in law firms because of the general inability to access the material from the PCs we use to do our work. Scanning was the next great answer. But let's be realistic: How many of you have had bad scanning experiences?

Yes, I see you raising your hand.

Imperfect in practice

And why has scanning generally been so unsatisfactory? Because most people equate scanning with using OCR software to identify characters on a page and turn them into an editable word processing document. Good idea in concept, but far

from perfect in practice, even with the latest, greatest technology. Even with the cream of the modern OCR software crop hitting text recognition accuracy levels as high as 97 percent, that's just *not* good enough.

Victims and veterans

There are four problems here and OCR veterans and victims alike will immediately identify with all of them:

1. Think of 97 percent accuracy this way: three of every 100 characters are incorrect. With a single-spaced page containing an average 2,200 words, that's 66 errors a page. And what if one of those is nearly impossible to detect but a bet-the-case-on-it number? Not good at all.
2. OCR software tends to have a significant number of problems retaining the formatting and layout of the original scanned document. For example, you get a local state court pleading and ask your secretary to scan it. A seemingly simple request. It's a "clean" document with all appearances of being a solid candidate for an OCR scan: mainstream typestyle and original laser-printed document. What you'll likely get back could be a reformatting nightmare, with a caption that defies clean-up,

equally baffling line spacing and tab stops. OCR software tries to figure out what codes or styles to apply in the target word processing format, but it's guessing at best.

3. OCR is not terribly speedy. Even with new high-end 3.8 ghz Pentium 4 PCs, the OCR process is fairly slow, and it seems with every increase in accuracy we have a geometric leap in processing requirements. So even if you can afford one of the megabuck intelligent character recognition systems like those from Kofax that use costly processing boards, you need heavy-duty PC horsepower for adequate text recognition. Forget about those decrepit now-vintage Pentium IIIs with a paltry 256 meg of RAM that have been off-lease more than two years. Either way, with OCR there's waiting involved.

4. Finally, there's the always-wide expectation gap between what we think is OCRable and those actually are. I can't tell you how many times I've talked to folks who've said, "When I try and scan this thing, all I get is garbage. How come?" What they show you is a pre-printed state-specific divorce financial disclosure form replete with boxes and

lines. You figure it should be able to at least read the text. Wrong. We technologists must realize that the average lawyer user who expects this to work has a more legitimate claim to reality than those of us who make excuses for present technology by saying, "Well, of course it won't be right — look at all those lines and boxes — nothing can recognize those."

Bottom line? Equating scanning with OCR is a fallacy that needn't be. This is where my "PaperLESS Office™" concept comes in. First put forth in an article in *Law Office Computing* back in late 1995, here's the concept in a nutshell:

Using low-cost, high-simplicity image scanning, physical paper is turned into "digital paper." Image scanning uses a scanner to effectively photocopy your documents onto your computer system. This creates "digital paper," ideally stored in the universally readable PDF format.

Perfect image

Digital paper takes up no physical space and is manipulated easily by software on your PC. The beauty of digital paper is that it is perfect — a picture of the original document, exact in every way. Of course, you don't have editable text at this point — it's merely a picture of the document. But most of the time, that's all you need.

So now you have your digital paper/PDF, a letter from opposing counsel, a set of interrogatory responses, the C.V. of a prospective

expert witness, and a stack of hospital records. What does it accomplish having all these pictures, these pieces of non-editable text? You save time in tracking down the physical file, or rummaging through a roomful of banker's boxes. All document-search time costs money — dollars wasted whether it's lawyer or staff time. We can use the scanner for its best purpose — creating images — rather than using it for its less-than-perfect OCR ability.

And it gets better. One of the core problems in working on client files is they are always split in two locations. The documents we create are located internally on our PC systems. The client documents we receive from outside sources are stored in our paper filing systems. So, if you want to view all the correspondence in a client's file, you must look in two places — onscreen for your own documents and in the paper file to view the externally generated letters. That is, of course, if no one has that particular file in their briefcase at home.

Managing the docs

Next, whether you are using a document manager like Worldox or the document management capabilities inside case managers like TimeMatters, Amicus Attorney, PracticeMaster or ProLaw, you go to that client file's folder/directory on your system and look in the "folder" where you store the correspondence for the client. You will see document names that begin with "Letter to ..." — word-processed documents you created — and document names that

begin with "Letter from ...", which are the scanned images of externally generated documents. Now, internal and external documents are in the same convenient place. Double click and that perfect picture of your document appears in the Adobe Acrobat Reader software.

The bottom line is your client files become electronic and totally contiguous. They're all in one place. You just can't help but save all sorts of otherwise non-billable wasted time you would spend looking for things. Not to mention the ease of bringing client files home or taking them on the road to a depo or a trial without lugging back-breaking boxes of paper (and subjecting the potentially irreplaceable originals to coffee spills, misplacement and other forms of folding, spindling and mutilation).

And when you close the file, it's digital paper you can store it in a convenient byte-sized package (pun intended.) This is a far better alternative for closed-file storage than the costly space-hungry storage requirements for actual paper files, which eventually commandeer an area the size of a starter home.

Considerations

What kind of scanner should a firm deploy? What software should be used to scan, organize and search through the contents of "digital paper?" Factors to consider: (1) intended volume of documents to be scanned; (2) number of pages scanned per job; (3) budget for internal scanning vs. cost-effectiveness of outsourced scanning. As to volume, read the specifications for duty cy-

cles. Buying a \$100 scanner rated for 2,000 pages monthly when your firm needs to scan 10,000 per month will surely smoke that bargain. The scanning market roughly breaks down this way:

1) Entry-level: usually flatbed scanners without automatic document feeders, \$50-\$300. Unsuitable for law firm use because of cumbersome paper handling.

2) Portable scanners: Visioneer's Strobe XP100 is less than one pound. This smaller-than-an-egg-carton scanner can pull five imaged pages per minute into your computer for under \$200 and it has no power brick — it's powered by your PC or laptop's USB port.

3) Entry to mid-level document-fed desktop scanners: \$250-\$1,000 scanners with automatic document feeders. Suitable for lower-volume scanning up to 15,000 pages monthly. Look at Fujitsu's high-value ScanSnap series (which come bundled with a full Adobe Acrobat), Visioneer's Strobe XP450, or Xerox's category-buster, the 50-ppm Documate 252.

Instead of viewing scanners as optional peripherals, consider them an essential part of the law office desktop PC system. Scanners to create digital paper should be ticked off a PC configuration sheet in the same manner as one would select a CD-writer.

Above this level, the sky is the limit. Spend enough money and you'll end up with a 12 hp engine riding model with a pull start! (Almost.) Fujitsu, Panasonic,

Bell+Howell, Canon, Ricoh and Kodak produce scanners that push the 100 ppm mark with massive paper handling ability.

Get organized

OK, so you now have these images in your computer. What's next? Organize and search them. Document management and work-product retrieval systems are the best answer. These software systems can gently impose a file cabinet-like consistency on the way any law practice organizes both its internal and external scanned documents. Worldox is the undisputed leader in the small-firm marketplace and has been digging into the larger firm segment for several years with great success. For larger firms, Interwoven (formerly iManage) and Hummingbird Docs are popular. Most legal case management systems also incorporate document management functions that can adroitly handle scanned "electronic paper."

All document managers let you organize and search scanned image files. This presumes, of course, the images are stored in a format that actually permits content searching of what would otherwise be only a picture. Documents scanned with the latest Adobe Acrobat 7 or Adobe's Capture systems are stored in the universally viewable PDF format. PDF documents can now be "Captured image over text" documents. This means that if the software can recognize the underlying text, it may be searchable by a document manager with PDF-search capabilities. Worldox excels at such a role as part

of its overall complement of document organization, management and retrieval functions, but isn't the only tool that can accomplish this.

Think of the PDF format as an "anti-paper" tool — the answer to the nightmare that endless piles of paper represent. Just as we use anti-virus software and anti-spyware tools, PDF processors such as the market-leading Adobe Acrobat can be seen as "anti-paper" tools.

Quick tip: a common misconception is that if one scans at a higher resolution, the text recognition results will improve. In fact, often the opposite is true — lower scanner resolution settings can yield better recognition. At higher resolutions, modern scanners can actually be confused by the fibers in the paper. Set the resolution to 150-200 dpi for better text recognition results.

Paperless is never going to happen. No matter how diligently you try and reduce or eliminate the paper you generate, others will still send you paper for years to come. Learn to love your scanner by becoming PaperLESS™ in your practice. Employing a creative and common sense approach to scanning, and leveraging anti-paper PDF tools, you can transform your desktop landscape. Piles recede, billable time increases. Touch the paper less, find the paper less and you'll find more profits, more enjoyment and better client responsiveness. It's as true in 2005 as it was back in '95.

**Lawyer Ross Kodner founded
Milwaukee-based MicroLaw**

**Inc., an international legal
technology consultancy and
continuing legal education
company, in 1985. Reach him
at rkodner@microlaw.com.**