

The Current State of OCR

By Charles W. Jackson, President, AnyDoc Software



Charles W. Jackson

Recognized as a pioneer in the software industry, Jackson has launched two successful software technology companies that have been named in *Inc. magazine's* "Inc. 500" list five times. Jackson, founder of AnyDoc Software, Inc. (formerly known as Microsystems

Technology), currently oversees and directs all aspects of the company's daily operations, including setting its strategic course in the marketplace. He also serves as its primary spokesperson and representative.

AnyDoc Software recently attended a conference of accounts payable professionals, where we sponsored a guest speaker. The speaker provided comment cards for feedback of his performance and content. On one card was a remark that floored me:

"OCR? That doesn't work."

I was stunned.

Was this merely the perception of one individual, or did it reflect the views of a significant portion of the workforce? Minus focus group data or other matrices, it's impossible to tell; however, it certainly was troubling. Troubling because not only has OCR been proven to work, but it works quite well.

Perhaps the perception of that one individual helps illustrate why only a small portion (approximately 15%) of businesses that could benefit from a solid data and document capture solution has actually taken advantage of it.

In order to bridge that gap, our industry needs to better educate the public about the strength and efficiency of OCR (and its companion, ICR) and how it can revolutionize data workflow. With any luck, people like our friend from the conference will begin to see OCR with different eyes, and those of us in the industry will have a renewed determination to inform businesses of all stripes of the benefits an OCR solution can provide.

Data Capture: Improving the Standard

It is true that OCR wasn't always as robust as it is today. In its infancy, OCR

was fraught with errors, causing many businesses to forsake data capture for tried-and-true manual data entry. However, OCR/ICR has grown into a formidable, extremely accurate technology.

Data capture has always been dependent upon convergent factors for its success. A good scanner, a reliable OCR/ICR engine and a high-end processor all work in cohesion to quickly deliver high-quality data capture from paper documents and forms. Each of these elements has improved within the past five years, thereby delivering far better data capture, far quicker than before.

For one thing, scanners are much more affordable. Recently, the prices of many mid-grade and high-end scanners have come down considerably, giving smaller businesses easier entry into the world of forms processing to handle the volume they require.

Also within the past five years, TWAIN drivers have been developed to support production scanners at rated speeds, ensuring a robust performance and compatibility between the scanners and document and data capture applications—eliminating the need for add-on "middleware" or hardware boards.

And the software designed to enhance scanner output has also made tremendous strides, delivering much cleaner digital images. Kodak's Perfect Page, Kofax's VRS® and AnyDoc's proprietary software tools were all developed to help scanners produce the best scanned images possible. They help to deskew and align the images automatically, saving precious time otherwise

dedicated to manual document preparation prior to scanning.

Character Recognition: the Strength of OCR

Of course, a clean, straight image makes data capture much easier, but OCR/ICR processing has also made quite a bit of headway in the past three years or so. OCR and ICR speed and accuracy is exponentially better than it's ever been. The technology now can process full OCR/ICR on a page in under half a second. To put it in perspective, OCR/ICR took about 45 seconds per page when we first began to develop data and document capture technology more than 15 years ago.

The systems running the OCR and ICR continue to become faster, and that obviously affects the results. The speed is useless, however, if the end-result is sloppy data. Fortunately, the OCR and ICR engines, particularly those from ScanSoft and Océ, have also gotten heartier—and the better the character recognition, the better the output. We know of businesses that consistently experience near 100% data accuracy. With these kinds of results in just a few short years, I'm excited to think about the things technology will afford us in the near future!

In recent years, a stable OCR/ICR environment has provided unique opportunities to expand the way the technology benefits business. Previously, standard data capture was dependent upon a defined template that required data fields to remain in the same location from form to form—a technology known as structured document processing. Within the past few years, a significant expansion of this technology prompted the advent of unstructured (semi-structured) document processing—a method of extracting critical data found on inconsistent locations of the same form type.

This technology gives our industry incredible opportunities to address the specific pains

"Our industry needs to better educate the public about the strength and efficiency of OCR and how it can revolutionize data workflow."

felt by the vertical markets they support—such as insurance, mortgages, healthcare, or the accounts payable department of virtually any organization. Each of these, and countless others, has critical data located on complex forms that standard, template-based forms processing tools do not handle very well. Keywords, not templates, are used to locate and extract the required data on each unstructured form type, regardless of where it may be found on the form.

In fact, our friend from the conference may not realize that OCR-based technology is available to tackle the very problems those in the accounts payable field face each day. Unstructured forms processing can automatically capture the critical invoice data, such as the invoice date, amount due, terms, purchase order number and even detail line items, that businesses need entered into their A/P systems.

It's also a very intelligent means of processing. By incorporating "fuzzy" logic into the equation, the technology seeks variances of the keywords found on the form type, such as "P.O. Number" and "PO #" for purchase order number data. Also, the more frequently a form type is processed, the more the technology "learns" where to find the data on the page and subsequently becomes faster in doing so. But the ability to do so, with precision, has happened only in the recent past.

That's not to say that unstructured forms processing itself is that new. Actually, it's been nearly a decade since the technology first emerged—albeit prematurely, perhaps. At that time, the systems supporting the technology were too costly and the results were sketchy at best. Understandably, organizations soon became skeptical that the technology had anything to offer them at all. Perhaps that unfortunate history and similar events prompted the remark we received.

Fortunately, the landscape has changed—and we need to proclaim that loud and clear. Particularly so for unstructured processing, because even some true believers in OCR doubt the technology has merit or staying power. But our experience has proven otherwise—we recently implemented our 41st successful unstructured processing solution. Additionally, we feel the market will gradually veer to unstructured processing solutions—the more businesses realize the pliability of the technology, the more they will seek ways to customize it to their way of doing business.

Distributed Capture: Re-centralizing the Decentralized Office

Perhaps the most versatile development from our industry in recent years has been distributed capture. It affords organizations

with offices across the country or around the world the freedom they've long sought to freely share their data across the enterprise. Paper forms are scanned locally, then the images get transmitted securely to the company's headquarters, where the data gets captured. Operators then verify and release this data into a backend system, where it becomes freely available within the organization for query and use.

A distributed capture solution affords multi-office organizations the freedom to select the best means for them to distribute the workflow. With workgroup or departmental scanners (that now can process 20 to 40 ppm), a broadband Internet connection and an OCR-based data and document capture solution that allows both local and

"While distributed capture offers a way to route data, more importantly it provides a highly cost-effective and secure means to do it as well."

remote capture, varying phases of the distributed capture process can be allocated to locations based on the data origination point and company need.

Perhaps the cycle begins with paper documents being scanned from branch offices worldwide. An OCR-based solution housed on servers in corporate headquarters then performs quality assurance on the freshly scanned document images. The images are processed and the data gets captured from them. Verifiers in another branch office then review the output data for any questionable characters, make corrections, and release the data to the backend system also at corporate headquarters, where the data becomes available for use. Distributed capture makes possible this or any other variable workflow process suitable to a client's needs.

While distributed capture certainly offers a convenient way to route data throughout the enterprise, more importantly it provides a highly cost-effective and secure means to do it as well. By distributing documents electronically, companies can save thousands of dollars annually in shipping costs, since there's no longer a need to mail documents from one office to another. The paper

documents stay in their office of origin as their data and images are disbursed throughout the organization. Since the documents stay put, there is little concern for them being lost or stolen in transit. And with encryption of data over an Internet connection, security is enhanced as the electronic data gets routed through the capture process.

The odd thing about distributed capture is the way it seems to have changed the organizational dynamic. Many companies with a large national or global presence have a decentralized control of their data, since each branch or regional office maintains a hold on how data within the location gets disbursed—partially due to control issues or perhaps even disorganization.

However, with the development of the distributed capture environment in the past few years, this has begun to change. A methodically performed distributed capture solution can help to re-centralize the control of data for the entire enterprise back into the hands of the corporate and/or IT structures. What was once a fractured channel of data can once again be united for the company's common good.

In Summary

I believe it is important for us to remember just how far our industry has come in just a few short years, and I think it is equally as important that we bring all of this to the table when approaching our prospective clients. We need to understand that many of them may not understand what we do, how well it works or why our solutions will make an enormous impact on their day-to-day workflow and yes, on their bottom line.

I truly believe our industry needs to educate the market on the multiple uses for a forms processing solution: efficiency improvement, cost effectiveness and compliance (think Sarbanes-Oxley, among others). And perhaps we need to expand our audience, as well. Perhaps we should reach out to the small- and medium-sized markets to ensure they too can benefit from a stable forms processing solution the way several organizations in larger markets already do.

Once we do, we'll see a vast improvement over the current 15% of qualified businesses taking advantage of a document and data capturing solution. I also believe that once we do, our friend at the recent conference will leave us a very different comment:

"OCR? Can't live without it!" ■

AnyDoc Software, Inc. offers innovative document and data capture solutions that have been the industry standard since 1989. Thousands of companies worldwide regularly rely on its software to eliminate millions of man-hours of manual data entry while improving data accuracy and productivity. AnyDoc Software provides the full spectrum of information-capture solutions from form design to document storage and retrieval with its software products OCR for AnyDoc®, CAPTUREit™ and BROKERit™. AnyDoc Software also offers complete solutions for the accounts payable and healthcare markets. For more information please visit <http://www.anydocsoftware.com>.