

Intelligent Data Capture: A Trend Only Beginning

IMERGE Consulting, Inc

Robert Smallwood

July 2006

Robert.smallwood@imergeconsult.com

As Appeared in Km World

www.imergeconsult.com

Intelligent data capture (IDC), also sometimes referred to as intelligent document recognition or IDR, is the ability to scan documents or electronic pages that have no fixed layout and extract data from specific fields to populate a database or business system. Data must be pumped into major financial applications like Oracle E-Business Suites or mySAP Business Suite, and the sooner the data is entered, the sooner users at a company can begin the financial management process.

Using IDC, documents may be unstructured, with varying layouts, or “semi-structured” with some fixed fields but mostly varying formats. Examples of unstructured forms are contracts, resumes and letters; semi-structured documents may be Explanation of Benefits (EOB) forms, insurance claim documents, invoices and the like.

The bread-and-butter application for IDC is the processing of invoices for payment in the Accounts Payable process. Invoices may be hundreds of pages long and have thousands of line items. It is critical to process these invoices expeditiously so that proper and maximum discounts can be taken for payment within a limited timeframe. For instance, an invoice may allow a discount of 5% if paid within ten days, but the full amount is due within 30 days. When companies process thousands – or even millions – of invoices the annual the savings for early payment can be millions of dollars. Why not just pay every invoice immediately and gain the discount? Because holding on to a company’s cash for as long as possible improves cash flow and allows a company to invest it in short term financial instruments, further improving profitability.

But the biggest savings across the board from using intelligent data capture technologies is from radically reduced labor costs and compressed business process timeframes, due to the near-elimination of manual data entry.

The IDC marketplace has its roots in basic scanning and document capture, pioneered by firms like DICOM Group’s subsidiary Kofax (Kofax.com), EMC’s Captiva (captivasoftware.com) and ReadSoft (readsoft.net). Other firms such as ABBYY Software House of Moscow, Russia (abbyy.com), Brainware (brainware.com, formerly SER Solutions, Inc.) and Xerox partner Rochester, NY-based Document Strategies, Inc. (docstrategies.com) are also active in the intelligent data capture marketplace, which is dynamic and robust.

Acquisitions and transformations characterize the major players competing in this market. To get from basic document capture to IDR, Kofax acquired Neuroscript and LCI in 2005 and 2006, and Captiva acquired French-based SWT in May, 2005; and by year-end was acquired itself by storage giant EMC (EMC.com). Also last year, Readsoft acquired a 50% stake in Danish firm Consit Development ApS, to give ReadSoft Oracle expertise and an entrée into Oracle’s E-Business Suite financial system.

The larger firms in the IDC market space have only reached around \$10 million in annual revenues, so it’s not a huge market in and of itself. “This is a technology that belongs embedded as a part of other business systems,” states Brainware CEO Carl Mergele. “And we intend to continue pursuing partnerships to that end,” he adds. Readsoft US President Bob Fresneda heartily agrees, “Alliances with major business application providers are a key part of our strategy. That’s why we have gained certified status for Oracle’s E-Business Suite and SAP R/3.” He continues, “We can directly populate their payables applications and use Oracle and SAP native workflow routines. No one else can say that.” ReadSoft claims to be the U.S. leader in payables processing, with approximately 250 customers using ReadSoft for INVOICES alone.

But the competition in this market is intense, and how it plays out is not clear at this stage. ABBYY is a research-intensive organization focused on selling its technologies to VARS and integrators, with particular strength in Europe. Brainware has been awarded patents in the US and Europe which it hopes to parlay into licensing

opportunities with other firms, perhaps even quasi-competitors. Document Strategies is leveraging a close relationship with Xerox (both based in Rochester, NY) and a new Xerox nationwide services offering will be based on the Document Strategies technology set. EMC Captiva is preaching information lifecycle management, leveraging EMC's Documentum (documentum.com) unit for Enterprise Content Management (ECM) and EMC's core storage management capabilities. Kofax has probably the largest distribution channel network and ReadSoft is leveraging its Oracle and SAP relationships and aggressive marketing to penetrate the installed base of these software giants.

Now for a review of some of the key players in this market:

ABBYY centers its business around document recognition, data capture and linguistic technologies. They develop and sell AI (Artificial Intelligence) applications, and, in particular, document recognition and natural language processing applications that help people overcome language barriers in a world that is globalizing at an ever increasing pace. ABBYY's FlexiCapture Studio is an additional tool that extends capabilities of ABBYY FormReader data capture applications and FineReader Engine SDK. This tool helps to extract data from semi-structured forms and documents

FlexiCapture Studio allows creating a formalized description, called FlexiLayout, which tells ABBYY FormReader or the FineReader Engine how to look for required fields on documents with similar data but different layouts. ABBYY FlexiCapture Studio is intended for Developers, VARs and Integrators and provides an easy-to-use "front end" access to the development know-how previously used by ABBYY in-house developers in large-scale forms processing projects.

ABBYY FlexiCapture Studio allows the developer to create a FlexiLayout for extracting data from both simple and complex documents on which the location of similar fields may vary greatly.

Brainware has a strong presence in the oil & gas marketplace with customers like BP, Shell, Conoco and Halliburton. It attained this presence by leveraging its relationships with large integrators including BearingPoint (bearingpoint.com) and CapGemini (capgemini.com).

Brainware's A/P-distiller product is designed to handle large invoice volumes, and particularly shines in table extraction. The Brainware Table Extraction Engine is an intelligent engine that extracts tables and associated line item information from invoices using line pattern recognition technology. The line item detail can be validated or used to collect additional data elements through table-lookups. A/P-distiller excels by using advanced learning technology to "read" and "interpret" documents for sorting and data extraction. A/P-distiller can process multiple and changing invoice types, extract critical data and line item detail, and export data to existing financial, workflow, and enterprise content management (ECM) systems— significantly reducing the costs associated with document sorting, manual data entry, and validation.

Brainware's Order-to-Cash (OTC) distiller virtually eliminates manual data entry of sales orders and requires no changes to established business practices. It automatically sorts incoming documents, extracts all required data from a purchase order, and creates new orders in your back-end order processing system. OTC-distiller handles multi-page orders, multi-lingual documents, and an unlimited variety of purchase order formats.

Document Strategies is a research-oriented firm that uses advanced pattern recognition, neural and AI technologies to create a simple-to-use, yet powerful capture and search technology set. It has unique capabilities and strengths when recognizing unstructured documents. President Jim Bowen states, "Our customers can set up rules for unknown document types and pull out key information – without any pre-defined templates. Our AI engine "understands" the context of content in the documents, therefore finding information and patterns buried deep in thousands of pages of documents." Its largest group of customers are in the legal vertical market segment in discovery/litigation support applications.

Document Strategies has a marked strength in recognizing poor quality documents – those with as low as 50% recognition confidence rating – such as water-damaged or very old documents, and faxes that may be fourth generation. Its capture technology maintains formatting and search technology allows users to retrieve search requests quickly and to modulate the accuracy level (and therefore resultant number of hits) of searches. This is especially helpful in litigation support applications. Users may refine searches even using similar “thesaurus” terms or synonyms (e.g. look for “corporation” and also find incidences of “business.”) Due to its unique technology, it performs very fast searches, typically sub-second, which is unusual for AI technology in general.

EMC Captiva has traditionally focused on larger deals in the higher end of the marketplace. Today, the focus is moving downstream to more middle market companies, but the new opportunities that processing unstructured and semi-structured documents using IDC is bringing about a wealth of new opportunities. “Capture technology has evolved to a point so that semi-structured documents can be processed like structured forms used to be,” states EMC Captiva Director of Corporate Communications Rob Jensen. “That means we can go after new applications in large accounts.” The neural technology acquired from absorbing SWT allows Captiva software to “learn” an invoice format so that the next time an invoice from that vendor is processed, it is instantly recognized.

Beyond Documentum, EMC Captiva has partnerships with major ECM providers like IBM, FileNET, OpenText and Hyland. Its largest integration partner is EDS and it sells approximately 40% of its deals through the indirect channel. Captiva's IDC offerings are aimed at the lower page volume range and are based on SWT's b-Wize product line. They include b-Wize MAIL for to capture and classify incoming paper and electronic documents from one location; b-Wize MONEY to capture checks, payment slips, deposit forms, etc.; b-Wize INVOICE and b-Wize FORM. Additional products include InputAcce/ for Invoices and Digital Mailroom.

DICOM's Kofax unit has traditionally focused on the lower end of the marketplace and has built a vast network of channel partners – over 1,100 worldwide including healthcare giant Cerner, as well as SYSCOM, FileNET, Hummingbird and IBM Global Services. “Our market has traditionally been post-archival, but now we are getting pulled forward in the business process, capturing data to feed business systems,” says Kofax's VP of Engineering and Technology, Sameer Samat. “It's opened up a whole new set of opportunities for us.”

Kofax INDICIUS is a suite of IDC modules that provide advanced document classification, separation and extraction capabilities, enabling the automated processing of more types of business documents. Some of its target applications include Application Form Processing, Mortgage Processing, Records Management and Mailroom. In application form processing, INDICIUS automatically extracts handwritten and printed information from forms and performs validation to ensure downstream business processes run as smoothly as possible. In mortgage processing the presence of all the necessary documents is verified and free-form technology extracts any required data. In records management it automatically applies a file plan to a set of documents, reducing human error and the risk of non-compliance. Its mailroom application identifies document types within a mixed mailbag of correspondence, forms, invoices and so on, and automatically routes each piece to the correct workflow queue.

ReadSoft was one of the first to develop IDC for semi-structured and unstructured forms, with product rollouts in 1997 in Europe and 2000 in the U.S. “We believe we've got time-to-market and product maturation advantages over our competitors,” ReadSoft's Fresneda says. Of the competition's moves in the past year, he adds, “We see the EMC acquisition of Captiva as solid. It helps validate the market in intelligent data capture.”

Major ReadSoft US customers include Allstate and Sherwin-Williams and recent major contracts include Lockheed Martin and Sony of Canada. Worldwide customers include IKEA Porsche, Pfizer, BMW, Volvo and DaimlerChrysler. Approximately 70% of revenues are driven through the indirect channel, the largest partner being Chicago-based SPSS. ReadSoft has integrations and relationships with ECM providers Open Text, Filebound, Liberty IMS, Digitech, Westbrook and Hyland, and has closed deals with FileNET and IBM.

In addition to certifications with Oracle and SAP, ReadSoft also includes Microsoft's Dynamics GP (Great Plains) accounting software.

ReadSoft DOCUMENTS for Invoices can automatically post transactions with no human intervention. INVOICES was the first invoice handling application to be certified for integrated use with SAP, and with a hundred customers it is still the most widely used.

The intelligent data capture market an increasingly competitive marketplace. The competition is heating up, but there is a long way to go and a lot of territory to cover before the winners in this market are clear.

Next month, we will cover some additional players in this market.

Robert Smallwood is a Partner with IMERGE Consulting and may be reached at Robert.smallwood@imergeconsult.com.