Bridging The Gap Between Unstructured, Handwritten Documents And The Demand For Electronic Data

By Jean-Louis Fages, President and Chairman of the Board, A2iA Corp.

As seen in Business Solutions Magazine and ECM Connection
Original Publication Date: 2010

As the information explosion continues, the demand for electronic data increases exponentially. But what happens when much of the data is trapped in unstructured documents, or written in cursive handwriting? How can you index these documents, or automate your mailroom, when these documents do not have a uniform structure and traditional recognition technology cannot access the information? And what happens when you must search or index these countless handwritten documents for tasks like e-discovery, redaction or due diligence because of new compliance or governmental regulations?

The need for the conversion of paperwork into computer-usable form is becoming more apparent each day as we enter a new era in document and information processing. It has been reported that each day in America, over one billion pages of paper-based information are manually keyed into a computer or indexed by hand. While this number continues to increase, the number of employees in the US dedicated to this work continues to decrease. We can conclude that there are two causes for this: (1) the growing number of organizations that send their documents for offshore data entry, and (2) there is an adoption of character recognition software that automates data entry tasks.

As more organizations adopt recognition technology as daily practice, the features and functionality of these tools becomes the primary factor behind productivity within an automated document-processing environment. New tools present in today's marketplace, driven by intelligent word recognition (IWR) and artificial intelligence, provide for more capabilities than ever before.

IWR differs from its predecessor, intelligent character recognition (ICR), because it recognizes data at the word or "field" level instead of at the character level, performing a deeper analysis. Utilizing this latest technology, new products on the market include OCR, ICR and IWR technology, and are capable of extracting numerous types of field-based information from a form, either constrained (machine print, hand-printed capitals), or unconstrained (freeform hand print, cursive), from a wide range of documents. In case of cursive handwriting, for each word analyzed, the system breaks down the words into a sequence of graphemes, or subparts of letters. These various curves, shapes and lines make up letters and IWR considers these various shape and groupings in order to calculate a confidence value associated with the word in question.

While IWR is not meant to replace conventional ICR and OCR systems, IWR is optimized for processing real world documents that contain mostly freeform, hard-to-recognize data, inherently unsuitable for these traditional recognition engines, and reduces the number of character errors associated with them. IWR eliminates a large percentage of the manual data entry of handwritten documents that, in the past, could only be keyed by a human.

So now that we are able to access this handwritten, unstructured data and convert it to a computer-usable form, what happens when compliance or regulations mandate we perform activities such as e-discovery, redaction or due diligence? And, taking it a step further, how can we work with this information for indexing and automation?

The Federal Rules of Civil Procedure (FCRP), for example, clearly state that all pertinent documents – no matter their form – are subject to the same identification, preservation, disclosure and production requirements. Whether machine-printed or handwritten, standardized forms or free-form blocks of text,

any and all relevant documents are required by the FRCP to be produced for inspection or other purposes, without exemption. Traditionally, searching through these documents has been a time-consuming manual process, often requiring additional resources. Great strides have been made in similar searches of electronic documents, as the data is much easier to index. However, even these technological advances cannot work with the electronic image of documents that are unstructured in nature or contain handwritten data. Unless they conform to a more standardized form and consist of printed text, finding and extracting the needed information can be challenging. Electronic images of loosely structured documents or freeform handwriting have been particularly difficult to feed into an automated solution. Time-intensive manual searches of these documents can often be less accurate than results from systems set up to similarly look through digital originals and purely electronic data.

The market's latest technology bridges this gap in functionality and, when integrated into existing systems for handling e-discovery, redaction or due diligence, gives organizations far greater capabilities when it comes to managing their documents and information. Tools proven successful for mailroom automation and redaction may also be applied to e-discovery. Expenditures can be reduced, efficiencies raised, times shortened and accuracy significantly enhanced. With the ability to recognize and transcribe freeform and even handwritten documents with as much ease as structured forms, this first-ever technology of its kind represents a much more effective way to search though digitized documents than any of the traditional methods available today. Users of this new technology can also see great time and cost savings when utilizing it for the indexing of archives or automating mailrooms / workflow applications. By accessing pertinent data previously unreachable by traditional recognition engines, the user saves both time and money by not having to utilize manual labor.

Leveraging intelligent word and character recognition technology and advanced document classification capabilities, organizations adopting the market's newest form of technology greatly expand their automated discovery of key information. Digitized documents are analyzed and classified based on their geometric layout and content characteristics, as well as document type. The documents can then be searched for predefined elements, and the results may be incorporated into pre-existing e-discovery or document management systems, providing unparalleled access to the extracted data.

Given the demands and requirements of today's market, any automated solution must be at least as accurate and error-free as the manual processes it replaces. As accuracy is equally important to successful discovery, due diligence, investigational or redaction efforts, this is an important advantage to today's new technology. By extracting selected data from the documents, users can now perform searches for pertinent information with an ease, accuracy and wide-reaching capability. Unstructured handwritten documents are no longer a bottleneck, as this technology makes a significant difference in the efficiency of tasks that were previously performed manually, and allows to make them part of an automated workflow.

Author's Biography

Jean-Louis Fages initiated A2iA's American subsidiary, A2iA Corporation, in 1999 and is actively involved in the company's international expansion. He currently serves as the General Manager of A2iA S.A. and the President and Chairman of the Board of A2iA Corporation. A2iA is a cursive extraction and recognition software company that operates one of the world's largest research centers focusing on ways to extract information from everyday paper documents that contain handwritten information. Jean-Louis received a diploma in engineering from the ESICI in Bordeaux and prior to joining A2iA in 1997, he was a business manager at Unisys France, responsible for the banking and finance unit. He also initiated some of the first ICR projects during his time at Unisys.