



“Business is a
Conversation”

– David Weinberger
The Cluetrain Manifesto:
The End of Business as Usual
By Christopher Locke, Rick Levine,
Doc Searls, and David Weinberger

Classification and Separation

White Paper
March 2008

KOFAX 

High-Volume Document Capture Demands Classification and Separation

Document automation technologies often fall victim to their own success. For example, the original layout of a typewriter keyboard, which is still used in virtually all keyboards today, was expressly designed to compensate for the shortcomings of the mechanics of the day.

Similar issues continue to surface as the speed of automation causes problems up- or down-stream from a workflow. The rated throughput speed of modern high-speed document capture devices is rarely achieved as bottlenecks in document preparation, sorting and error correction invariably slow work processes. Further, things that seem intuitive to a human can be extremely difficult to duplicate with software.

Document image capture – the process of taking paper documents and converting them to a digital image by scanning – has opened a world of process automation benefits. Documents can be transmitted electronically, rather than carried physically. Vast amounts of information can be stored in the smallest spaces, while the functions of data analysis, process workflow, search, and retrieval enjoy exponential productivity improvement. Nonetheless, with the improvements in the speed of document conversion have come a number of challenges not initially anticipated or easily overcome.

Optical Character Recognition (OCR) is an astounding accomplishment, offering the ability to extract data from forms. To the human eye, a “2” is a two and a “5” is a five; getting software to draw that same conclusion – with comparable accuracy – is no easy feat. While the recognition engines have improved a lot, having software that understands the content and context of a printed page – and can make sorting, routing or document separation decisions – is even more complex.

The High Cost of Documents

Despite the increased use of electronic transactions, paper volumes continue to increase. In its report: Summary of Worldwide Capture Software 2006–2010, Harvey Spencer Associates forecasts a continuing need to capture more information from unstructured inputs into document management systems and specific business processes. “There is an increasing need to manage businesses more efficiently and to understand the value that this software brings,” Spencer writes. “We estimate that the capture industry will grow to \$2.42 billion in 2010 – a CAGR of 16.4% – exceeding worldwide IT and ECM growth rates.”

Spencer’s research also addresses the high cost of handling documents. The researchers reported the average mail-handling performance rate for a clerical document preparation position: A skilled document handler can process between 750-1000 pages per hour – i.e. remove from an envelope, sort, stack and prepare for batch scanning. This does not account for the actual scanning itself. According to Spencer, this figure is subject to fluctuate based on the amount of repair documents may require.

These figures highlight the huge expense associated with manual processing. When multiplied by the millions of documents handled by many organizations, the costs can be staggering. While mechanical handling has progressed, allowing envelope opening and document image scanning to replace what was historically a manual operation, software advances have been more gradual.

While high-volume document scanners have made the conversion of millions of paper pages seem routine, as the volumes of documents being processed has increased, expectations for the systems have also risen. The demands on processing systems at times tax their capacities.

One challenge that has burdened high-volume document processors has been the need to distinguish individual documents from one another. This is especially difficult when dealing with groups of multi-page documents, such as those frequently found in loan processing applications, or in the typical mailroom. While it is usually a simple matter for a human to say, “this is where a loan application document ends, and here is where the credit report begins,” that determination has been much more difficult for an automated system.

Yet by necessity of sheer volume, processors have been forced to feed massive amounts of unsorted pages into document scanners, arranged by little more than the order they arrive. Page separation has been a labor-intensive, manual process either by scanning each document individually, or by inserting “separator sheets” in between each document. Another option is post-scan separation where a clerical worker must make an association of multi-page documents once the documents have been scanned (key-from-image).

In addition to being a costly and inefficient way to separate mixed documents, human labor is subject to error.

Automating Data Extraction

In most automated forms processing systems, specific data must be extracted. For example, accounts payable systems need to know the dollar amount to be paid, so these systems – whether manual or automated using Optical Character Recognition (OCR) – look for data fields such as: “Balance Due” or “Invoice Total”. The wide range of terms for specific fields, their location on an unstructured document, and the virtually unlimited number of data types, has made automating the extraction process particularly challenging. What’s more, these challenges exist even after the document type is known. (Back to the example, neither human, nor OCR can begin looking for “amount due” until they recognize the document is an invoice). Therefore, the first step in data extraction is the process of document identification.

Document Identification

In addition to serving as a means to extract data, accurate document classification also serves to identify documents that are part of daily business transactions. Whether a document is a correspondence that must be answered, an application for service, or a sales-order form, a document is used to start (or continue) a workflow and often must be routed for action.

For file management, document archiving, and record’s retention, accurately identifying and classifying documents becomes critical. A document’s place in an organization’s records’ management system is determined by its content. In fact, documents are often categorized, archived and retained in multiple repositories depending on the specific characteristics of the information they contain.

While classifying specific document types, such as accounting documents, is relatively straight-forward, in a mixed-document environment – such as a mailroom – there are apt to be a wide range of incoming items that cannot be readily identified. In heavily regulated industries, such as pharmaceuticals or petrochemical, there may be countless highly technical or highly unstructured documents that require trained (and costly) experts to accurately catalog.

Even in processing environments where documents have been pre-screened (e.g. the use of designated post office box addresses) there are usually a variety of incoming document types. Scanning these, which typically arrive in batches of varying documents types, requires a great deal of manual intervention. Either each document is scanned individually, specifying the type of document to the software each time, or the paper is pre-sorted into batches of identical documents.

Manual Document Separation

Work processes that deploy manual document separation and identification are fraught with inefficiencies. Bottlenecks slow the workflow and compromise the effectiveness of information technologies deployed downstream. Document scanners and network file sharing systems remain idle while waiting for manual operators to process documents.

Compounding these problems is the fact that all manual processes are subject to errors, which ultimately leads to the need for further – corrected – separation and identification. These “exceptions” that must be re-routed or duplicated, because they are simply lost, are the most expensive part of any document handling process.

In response to these challenges, organizations have deployed a number of techniques, most of them involving a high degree of manual labor:

- **Restrict scanning to fixed-length documents with a specific number of pages**

This requires sorting batches into identical fixed-length documents; in fact, some organizations choose to photocopy all documents prior to scanning so that the scanners are fed identically sized pages.

- **Add separator sheets between documents**

Automated capture software can reliably identify blank pages, so documents can be divided by inserting a “separator sheet” between individual documents or batches. With multiple documents and varying numbers of pages, the use of separator sheets is both manually intensive and costly because they must be inserted before scanning and, if they are to be reused, removed afterwards. In high volume operations, these costs can be staggering. For example, a typical loan file might require as many as 50 separator sheets. Multiplied across hundreds or thousands, or in some cases millions of loan files, these costs become prohibitive. It has been estimated¹ that the labor cost is 1.6 cents to insert a separator sheet and a further 1.6 cents to remove it. As the printing costs are approximately 1 cent per page, it is less expensive to leave the separator sheets in the batch and print new ones for further batches.

- **Use of Bar Coded Separator Sheets**

In some operations, separator sheets contain bar codes or patch codes to indicate the document type, enabling the capture software to, more easily, separate and classify documents. This offers more accurate document identification, but also adds cost. In fact, the added cost of just the paper and ink for bar codes can be substantial. And again, the addition of a bar code adds document preparation time when the sheets are inserted as great care must be taken to ensure the correct separator sheet is used. (Few operators can tell one bar code from another).

¹ White Paper - Auto Classification Saves Cost in High Volume Scanning Environments by Harvey Spencer Associates

Unfortunately, manual document separation, which requires a document-preparation employee to skim a document in order to accurately determine its content, is expensive, time-consuming, and error-prone. For high-volume batch processors, the initial benefit of automating document preparation comes from the labor savings of eliminating document pre-sort and separator sheets. Additional advantages result from the faster processing resulting in improved productivity and the ability to derive additional benefit from installed applications.

Automating the Document Separation Process: Vexing, Not Impossible

In response to the limitations and added expense of manual document separation, businesses are eager for capture applications that can automate the process of document identification and separation. This would allow organizations to use high-speed scanners and other capture devices at the speeds for which they were designed. By eliminating the need to separate documents pre-scan, not only is labor reduced, but work processes are streamlined. Unfortunately, for many software developers, this has been extremely difficult to accomplish. Most challenging, the systems to automatically separate documents must first successfully identify and classify individual pages.

There are two basic methods of automated document classification: image-based and text-based. Both use recognition software to analyze documents and page layouts in order to identify recurring patterns to categorize documents.

Image-Based Document Classification

Image-based classification relies on documents that have a similar appearance, as opposed to their content. In cases where there is little or no readable text, this methodology can be very effective. Image-based classification can be very fast, and, with modern algorithms, its speed is not impacted by a large number of document types.

- **Learn-by-example image classification**

Image-based classification “learns by example” by comparing the physical appearance of document pages. This method is best suited to structured documents such as forms. The software works by dividing the image of a page into a matrix of squares and then estimating the probability of finding pixels within each square. These systems have become increasingly sophisticated, analyzing not only the individual matrix segments, but the image as a whole.

In order to work effectively, this method requires recurring documents in order to “train” the software. The more examples the system encounters, the more accurate it becomes.

Previously, the added computing power required for full-page analysis would tax processors. However, with the increased computing power of modern systems, there is little speed advantage by using a zone-only approach to image classification.

- **Rules-based image classification**

With learn by example image classification, there must be a minimum number of similar documents to provide a representative sample to train the system. Further, differences in the input capture sources (e.g. different scanners or fax devices) can impact this method’s effectiveness.

Rules-based image classification relies on establishing defined characteristics for the software to look for. This method is well suited to finding bar codes, or company logos, which are always present in specific documents. This can serve as an excellent way to pre-filter standard document types, prior to deploying more sophisticated methods.

Drawbacks of rules-based image classification include the need to manually select the defined features, which can be time consuming when working with large document sets. An additional concern is the fact that processing speed can be impacted if a large number of document types are defined in this way. However, image-based processing is faster than text-based, because the latter requires full-page OCR.

Text-Based Document Classification

Text-based document classification can be based on the content of a defined zone or the whole document. Also, this identification technique can be multi-faceted, for example, once classified to be an insurance claim, further classification might be based on a client name or account number. Text-based classification ranges from simple methods that look for specific words or phrases in the document to more complex self-learning systems that improve over time.

In the more rudimentary systems, key words or phrases must be specified manually and must be adjusted manually as new keywords become necessary. In some systems the words or phrases can be given weights to improve identification accuracy. Only a limited number of words, phrases and weights can be used as they must all be specified manually, so the accuracy of this method is severely limited. The totality of all the words, phrases and weights for a particular system is called a knowledge base. It is difficult to maintain the knowledge base manually because the addition of a new key word may break existing classifications.

More sophisticated systems use self-learning techniques to automatically identify the words and phrases and their weights. One such advanced technique is called the Support Vector Machine (SVM). SVMs can build knowledge bases of tens of thousands of words automatically by self-learning, which is much larger, and therefore much more accurate, than is possible by manual means. Also, as this machine is self-learning, accuracy improves over time and maintenance is simple. One method of self-learning is to supply the SVM with an existing hierarchy of folders containing documents that are already classified, called a learning set. The SVM works through the folders and examines all the documents to build its knowledge base.

The learning set contains documents that have already been manually classified. Unfortunately, a few of the documents in the learning set will not have been accurately classified because the classification process was manual and all manual processes introduce some errors. The latest technology in document classification uses a technique called Maximum Entropy Discrimination (MED), which builds its knowledge base in the same way but is more accepting of incorrectly classified documents in the learning set.

When facing a large variety of document types, self-learning text classification can be very effective. Setup time is shorter as there is no need to customize rules.

Text classification is most effective when it has ample text to work with. Therefore, it is not appropriate for documents that contain a small amount of text. Further, to achieve good results, a representative sample (usually at least 20) of each document type must be present in order to properly train the software. Because self-learning text classification relies on full-page OCR for each image, it is not as fast as image-based techniques. Currently, it may take a few seconds per page, however as processing speeds improve, this is becoming less problematic.

Handling multi-page documents

One of the greatest impediments to automating document capture is the fact that pages must be scanned one at a time, while the documents may consist of multiple pages. It's particularly difficult for software to accurately understand where one document ends and the next begins.

As capture applications have become more sophisticated, there has been recent headway toward resolving this paradox. Simple schemes recognize the first page of the document and software can be trained by example (i.e. showing the capture software the front page of every type of document). Whenever it recognizes one of the front pages of any of the documents, it assumes a new document has begun. This classifies and separates in a single operation.

This method of intelligent document separation relies heavily on the accuracy of the page classification. In the event that one of the front pages is not recognized correctly, the problem is compounded; the second document will be appended to the first document as an attachment and will effectively disappear from the workflow.

To ensure document pages have been accurately parsed, they must be manually inspected. The sophistication of the document capture platform dictates the ease with which any corrections can be made. Working with a modular capture system, an operator retains the ability to manually split document pages as necessary. However, not all capture applications offer this functionality, in which case the document must be released to the back-end and corrected there – a more costly, error-prone and time-consuming process.

Other limitations of basic document identification surface in highly unstructured documents, where there are insufficient graphical clues or context to identify a document's first page. Further, in instances where mixed pages have been scanned out of order – a frequent situation in high-volume environments – documents will apparently have pages missing or other document's pages inserted.

Global Classification

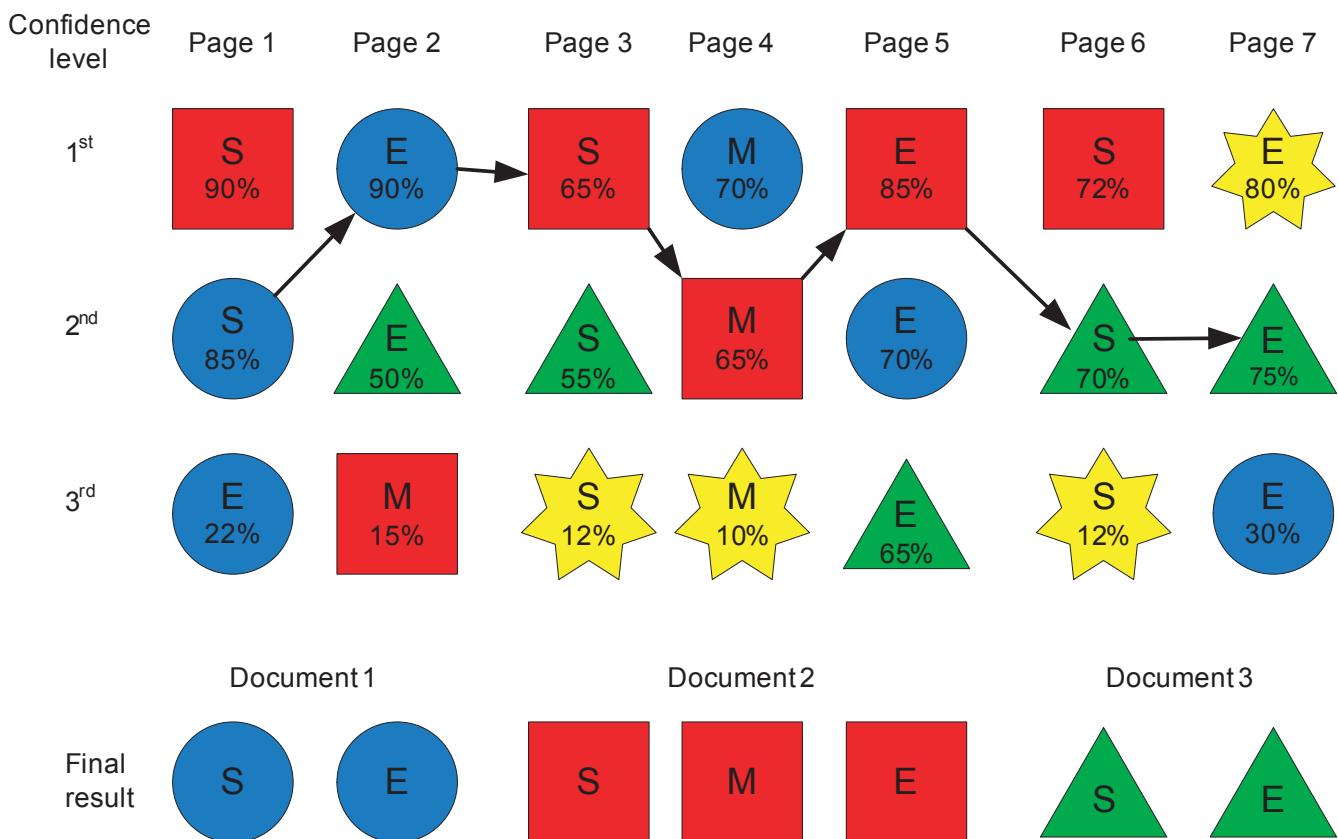
More recently, highly sophisticated systems have been developed that are deploying multiple recognition techniques to identify not only the first page in a multi-page document, but also the last and middle pages. In addition to the clues offered by individual pages, these systems constantly analyze the contiguous string of pages and how they relate to each other. The software can analyze the likelihood of each page belonging in the beginning, middle or end of a document. With each determination, the most likely placement is made – as weighted against other possibilities.

These powerful recognition technologies can identify boundaries between multiple semi-structured and unstructured documents in a single batch, replacing the traditional process of inserting separator pages and greatly improving the accuracy of the rudimentary first-page-classification schemes.

As an example, assume there are three possible document types represented by the blue circle, red square, green triangular and yellow star-shaped pages in the diagram below. Each page is scanned and classified and a confidence level associated with each possible page type. In the diagram below, the string of scanned images are shown across the diagram and three possible classifications are shown for each page, with the most confident being at the top, and the less likely below. The percentage represents the confidence level of the classification. The S represents a start page, the M a middle page, and the E an end page. There are a number of business rules in this simple example:

- All documents begin with an S page
- All documents end with an E page
- If a document is longer than 2 pages, all pages between the S and E pages are M pages
- An E page of one document is followed by an S page of another document

In the example, the business rules plus the global view of the page stream mean that the scanned pages are classified as blue circle, red square, and green triangle documents. (The pages have different shapes – squares, circles, triangles and stars – so that you can tell the difference when printed in black and white.) If only the most confident first pages had been used to classify the documents, the choice would have incorrectly been three red documents.



A Global Classification model offers a number of advantages over other methods. Because a Global Classification system looks at each page in the context of those around it, even in cases where a page has a low recognition probability, the document can still be correctly classified and separated. The system can also reorder pages that have been fed incorrectly or have been mixed during document handling.

Documents can be routed to the correct target transaction process or workflow without pre-sorting. This greatly reduces document handling costs associated with manual processing and increases the productivity of the business processes.

In Practice Today

Today Kofax's solutions are helping companies in many different industries process documents faster and more efficiently. Below is a small sample of the solutions and their immediate benefits.

Loan and Mortgage Processing

Mortgage processing is one of the most paper-intensive industries, and over the past several years Kofax has implemented solutions for more than half of the top 25 mortgage companies. One customer is realizing savings in excess of \$35,000 a day (about \$700,000 a month) thanks to a reduction in manual labor. Another mortgage customer has reduced the cost of processing each loan by \$15. (With peak processing periods handling around 30,000 loans a month, savings exceed a half million dollars a month.) These savings go straight to the bottom line.

With integrated automatic document separation technology, banks and other loan processors can increase efficiency and reduce the costs associated with processing the hundreds of thousands of mortgage and loan documents they receive from loan offices and customers.

While these benefits cut across industries, Harvey Spencer Associates cites the examples of mortgage loan processing, which contain a mix of documents, some with single pages and some with multiple pages. "Users can automatically confirm that all documents required are in the stack and that they are in the right order, even electronically fixing them when necessary," the analysts report. "When someone buys a stack of mortgages, the supporting documents have to be transferred. In the secondary mortgage market, we are talking very large numbers of pages as the buyer is likely buying several thousand mortgages as a bundle."

Document automation tools allow much more in-depth analysis of these large numbers of documents. For instance, offering the ability to electronically extract information that offers better value to the mortgagee – such as clues as to the likelihood of getting repaid.

One residential mortgage processing company handling 12 million images per month has deployed Kofax's solutions. With each customer folder containing nearly 100 pages, and as many as 80 different document types, this processor was an ideal candidate for automated document separation. Prior to automation, the volume of work required a staff of 60 people to perform the document preparation and separation. An additional staff of 16 was required for quality control.

Thanks to automated document separation, the processor has been able to significantly reduce manual processing. The same case load can now be handled with a staff of 10 people to do the document separation and preparation and three people to review.

Spencer believes that these advantages cut across industries. "Initially these deployments are being used for things like invoice processing and mailroom automation," he said. "But there is also opportunity in the ad-hoc segment –

in this case, the user does not want to spend time selecting which buttons to push – if the system makes the major decisions, you can substantially cut the time that the expensive ad hoc user spends at the scanning device.”

Government

Another paper-intensive industry is government, which has an enormous variety of document types and processes. The Utah Department of Human Services is such a government organization and is dedicated to providing child support services and support for children in care. They wanted to capture the paper documents for more than 90,000 child support cases that were stored in filing cabinets across the state. This represented 6 – 7 million pages of paper documents. Using Kofax’s advanced classification and separation technology, they were able to complete this task in 8 months instead of the expected 3 – 5 years, which was a time saving of over 80%.

Now that the files are stored electronically, multiple employees can work on the same case at the same time, helping to increase productivity and better serve customers.

ORS can now be assured that all case files are in full compliance with industry regulations and that all forms have been accurately captured. In a paper file, the case files often become a dumping ground for all kinds of unnecessary content. The Kofax solution actually automates the classification and identification of eligible documents from the case files and unnecessary and ineligible documents can be easily precluded from storage in the imaged file. In addition to keeping case files clean, this process also creates efficiencies which allow workers to get to the appropriate information in a timely manner.

Other Cases

The following cases demonstrate the levels of savings an organization deploying effective document classification, extraction and separation technology can achieve. These are actual results from Kofax customers:

- A Pathology Network substantially reduced costs related to printing separator sheets, such as paper, ink and toner by using Kofax’s automatic document separation technology for their incoming pathology specimen request forms. Its use also greatly reduced the human errors associated with manual document separation, as it automatically extracts the handwritten and printed information from the patient request form to a high degree of accuracy.
- A loan processing company was able to remove nearly all manual document preparation from their process, resulting in a reduction from 145 to 15 full-time employees. Despite the reduction, productivity increased by almost 90%. Processing time dropped from six minutes per document to less than one minute. They were ultimately able to double the number of loans processed daily;

- An information-services company processing nearly 400,000 human resources documents monthly. The service bureau sought a solution that could recognize zones of formatted pages in order to extract specific data to populate index fields. Kofax's technology has not only performed the requisite zone recognition, but it also provides the benefit of being able to recognize and automatically separate different types of forms;
- Another customer estimated that it would take 50 full-time employees to accomplish the same job manually. By deploying automated classification and separation, they were able to double capacity using only 10 people.
- A loan processing company estimates \$420,000 annual savings in reduced labor due to the reduction in manual document handling.
- Document processor nets \$100,000 annual savings in consumables alone (ink and paper) by eliminating separator sheets;
- Accuracy increase from 98% using separator sheets to 99.5% with software;
- A mortgage processor faced an eight-minute document retrieval time. They calculated they would realize complete ROI if they could reduce retrieval time to two minutes. Current document retrieval time is now less than one minute as a result of faster categorization and document retrieval provided by automation.

Document automation has come a long way since the typewriter was intentionally designed to slow productivity. While there seems to be no limit to an organization's capacity to generate paper documents, the processing of that paper has markedly improved. With the advent of intuitive document identification and separation technology, what was once one of the most expensive and time-consuming parts of a document process can now be reliably automated.

Learn More

To learn more contact Kofax at:

Phone: +1 (949) 727-1733

E-mail: info@kofax.com

Web: www.kofax.com

www.kofax.com

Copyright© 2008 Kofax, Inc. All rights reserved. Kofax is a registered trademark of Kofax, Inc. All other product names and logos are trade and service marks of their respective companies. All specifications subject to change without notice. (03.2008)

