ABBYY®

White Paper

Automated Data Capture: A Bridge to E-Transactions

By Ralph Gammon – Sponsored by ABBYY USA

White Paper

Contents

The bottom line is that applying automated data capture technology represents a more cost-effective way to improve the speed and accuracy of data capture operations than hiring more manual labor.

Executive summary

Electronic transactions have multiple benefits over traditional paper-based exchanges. They are typically faster, less expensive to conduct and can even be more accurate and secure than paper. In 2009, the U.S. Treasury Department estimated that it cost nearly $3 to process a paper tax return and only $.35 to process an e-return. Invoice processing studies have shown similar cost reductions for e-invoice implementations.

That said, there is certainly no shortage of paper documents in use in transactions today. Factors such as inertia, resistance to change and the cost of converting from paper to electronic processes have ensured that billions of pieces of paper are still used in transactions annually. For example, according to Paystream Advisors, paper is the most common way for organizations to send invoices – in 2009, U.S. businesses produced more than 25 billion paper invoice documents.

In addition, according to the U.S. Healthcare Efficiency Index, there are more than seven billion paper-driven healthcare transactions that take place annually in the U.S. And the U.S. Federal government alone, not counting local and state branches, receives 49 million paper 1040 forms.

Sure, in an ideal world, these organizations could snap their fingers and make their paper transactions magically convert to electronic ones – removing considerable cost and inefficiencies from their business processes. And, albeit slowly in some cases, we are definitely trending toward more electronic transactions. But, one look at the mortgage industry, for example, where the average file size has grown to more than 200 pages, and where electronic closings are still rare, will show that we are a long way from a paperless world.

Nevertheless, organizations don't need to be helpless bystanders waiting for their customers and partners to adopt e-transactions. They can aggressively attack paper with document and automated data capture technology, which is designed to digitize paper and convert its information into electronic data. Companies that implement capture solutions can get many of the same cost savings and efficiency benefits of e-transactions.

Applying OCR for data extraction

The key to efficiently converting paper to electronic data is the application of automatic recognition technology to document images. Document images are simply pictures of paper documents. They are typically created through scanning devices (although the camera functionality on mobile computers and smartphones is increasingly being at least considered for this purpose). Automatic recognition involves utilizing technology like OCR (optical character recognition) to determine the letters and numbers appearing on an image and converting that information to an electronic format that can be understood by other computer programs.

For example, if you submitted a paper form to order this white paper, that form could have been scanned and converted to an image. Intelligent field extraction, utilizing OCR, could then have been applied to transform the "locked" information on the physical document into actionable data fed into a CRM system. Here's how that extraction process works:

1. The system first applies OCR to analyze the text and patterns of the imaged document and is able to classify it as an "order form." (Most classification systems can be customized to fit an organization's form types.)
2. Once identified, the form can be routed into a field-extraction workflow specific to the document type. This workflow produces the proper meta-data needed for integration with a specific back-end system. The image of the order form, for example, would be tagged with critical information like company name, phone, e-mail address, etc.
3. Then, the image and associated metadata are absorbed into a CRM system, where a sales person can access them for a follow-up call.

This scan-and-capture process is designed to be more efficient than key entering data into the CRM system from the paper forms.

Advantages of automated data capture

Applying automated data capture can often reduce labor costs related to key entry by 50% or more. For example, by implementing an OCR-based system for capturing data from patient billing, health insurance, and health information forms, On-Site Anesthesia Services was able to cut its labor by 75%. For complete case study go to: www.abbyytalk.com/casestudy/onsite.

Many users report that scanning and OCR applications provide them with a return on their investments in less than 18 months. A 2010 survey by AIIM (the Association for Information and Image Management) found that almost 60% of respondents reported they achieved an ROI for document scanning and data capture technologies in less than a year-and-a-half. More than 85% reported an ROI in less than three years.

In addition to reducing key entry costs, automated data entry can improve turnaround time and visibility into incoming document streams. For example, through the introduction of an OCR-driven system for capturing data from invoices, a large cable company was able to increase the early pay discounts it was taking advantage of – from just over 50% to 98% – because it was able to process its invoices faster. Because of its high volume of invoices, the cable company was able to save more than $9 million in a single year.

OCR can also be more accurate than human key entry. For the 2010 U.S. Census, for example, the error rate for key-entered data fields was reported as being 1.4%, while the error rate for fields captured through OCR was .4%[1]. Of course, the Census Bureau used techniques like voting by multiple OCR engine and database lookups to improve accuracy, but implementing these steps is still typically more cost-effective than blind-double key-entry. In this approach, organizations have two data entry clerks manually key in the data and compare the results. While it can also be used to increase accuracy, but blind double key-entry essentially doubles that amount of labor in a manual data capture process.

The bottom line is that applying automated data capture technology represents a more cost-effective way to improve the speed and accuracy of data capture operations than hiring more manual labor. This is especially true in operations that have flexible document volumes. It's more efficient to increase the volume of images running through a software application than it is to ramp up key-entry personnel. Hiring more personnel not only involves additional salary and benefits costs, but also investing in training time for possibly temporary personnel. In contrast, once a piece of software is trained on data entry, it maintains its knowledge and, in fact, will typically become better over time as adjustments are made. Essentially, the more documents an automated data capture system processes, the smarter and more accurate it gets.

Additional benefits can be achieved by utilization of the document images themselves. Images can be archived in an ECM system for improved records management and accessibility. Electronic workflows for automating tasks such as approvals and exception processing can also be applied. Combined with automated data entry, automated workflows utilizing images can accelerate ROIs, reduce manual labor and speed up document processing times even more.

Data capture options

Automated data entry can be applied in a variety of ways to scanned documents. Let's take a look at your options:

1. **Full-text recognition:** This is the most basic level of OCR. It typically involves applying character recognition to an entire scanned page. On a clean, 200 dots per inch (dpi) image with a common font of typed text, OCR engines advertise accuracy rates of more than 99%. This type of recognition is typically used for search and retrieval purposes, often in desktop or legal applications.
2. **Field-based capture on structured forms:** This is another mature area of automated data capture. It is used on standardized forms that consistently contain the same type of data in the same place. Health insurance claims forms, tax forms, surveys and service applications are examples of documents where greater than 90% field-level accuracy rates have been achieved. Templates are drawn up that tell a capture software program where to look on an image for a particular data field. OCR is then applied to each field. Rules, such as "a patient ID number contains a specific alpha-number pattern and can only be a certain number of characters long" can be applied to improve accuracy, as well as checks against databases – such as lists of addresses or names.
3. **Handprint and cursive recognition:** So far, the accuracy rates that have been quoted in this paper are for typed fonts. But, automated recognition can also be applied to handprint and cursive writing. Because accurately recognizing handprint and cursive is more difficult, rules and database lookups are typically necessary. Techniques like utilizing separate boxes for each field (or even each character) can also improve accuracy. If set up correctly, hand-print applications have been known to achieve 80%-plus field level accuracy, which can certainly provide significant cost savings for users.
4. **Intelligent document recognition (IDR):** IDR is really a set of technologies designed to automatically classify and extract data from semi-structured or unstructured documents. An invoice is an example of a semi-structured document because, while all invoices typically contain the same type of information, such as supplier, invoice number and amount, this information can appear in different locations on the document. The location of the data often depends on factors an organization receiving the documents has no control over – such as which ERP system a vendor used to create the invoice. Customer correspondence would be an example of an unstructured document. Technology such as full-text OCR, keyword search, rules and even artificial intelligence can be utilized to capture data in IDR applications.

Data capture use cases

Following are some common use cases for automated data capture:

1. **Preparing documents for search and retrieval:** This typically involves applying full-text OCR to help users find documents through search engines, whether desktop or Web-based search. For legal cases, users will often search for keywords, or even multiple keywords in proximity, to find documents they're looking for related to a specific matter.
2. **Extracting metadata for enterprise content management (ECM) systems:** Metadata enables users to perform searches on specific fields, such as "name," "account number" or "date." Extracting metadata can involve applying field-based OCR. It can also be achieved through capturing bar-codes or check box marks, which can be used to carry information describing a document. Once captured, metadata is typically exported into an ECM system where it is used to reference the related document images.
3. **Forms processing:** Forms processing involves extracting data typically to be fed into a line-of-business application like an accounting, ERP or order management system. Forms processing can be applied to structured forms utilizing templates and OCR, or semi- and un-structured forms, utilizing IDR. Rules and database look-ups can be used to increase accuracy. Integration with the line-of-business system is common.
4. **Auto-classification:** Document preparation is often one of the largest costs in a document imaging process. A 2009 TAWPI/IAPP Benchmarking Study of more than 300 sites doing document scanning found that over a three-year period, 38% of operational costs were related to document prep. Seventy-eight percent of these prep costs were attributed to labor. Auto-classification technology can be used to reduce prep costs in areas like sorting documents and inserting cover sheets. For the complete case study go to www.abbyytalk.com/casestudy/dg

Automated data capture process

Following are six steps typically included in an automated data capture process:

1. **Document prep:** Document prep involves getting paper ready for scanning. This can include removing documents from envelopes and folders, taking out staples and removing paper clips, straightening folds, taping tears, sorting documents into batches, and inserting cover sheets for identification.
2. **Scanning:** Scanning is the process of taking a picture of a document and converting it to an electronic file. Scanners designed specifically for document capture typically have feeders that can hold multiple pages to be scanned in succession. Most printer/copier devices (often referred to as MFPs) also offer document scanning. Scanners typically offer bi-tonal, color or grayscale output at resolutions from 150 dpi to 600 dpi.
3. **Image processing:** Image processing involves cleaning up scanned images with techniques such as cropping, deskewing, despeckling and even applying thresholding algorithms to create optimized bi-tonal images from color or grayscale scans. It also includes compressing and formatting images as a TIFF (Group 4), PDF or JPEG files. Image processing can be done on the scanner and/or on a PC or server in a post-scan step. In alternative environments, like mobile, some additional processing can be applied, such as removal of noise specific to camera optics, keystone effect correction, auto-edge detection and shadow removal.
4. **Data capture:** This is the step of converting the information typed or written on a page into electronic information that can be shared with business applications like accounting, ERP and ECM systems. Key entering the data from an image is one option, but the application of automatic recognition technologies like OCR and IDR is preferable because of the potential labor savings it offers.
5. **Verification/Quality Assurance:** This step ensures that captured data and documents meet standards. Images can be reviewed manually or software for automatically analyzing quality can be utilized. In character recognition applications, characters or fields that don't meet certain confidence levels (which measure how sure a software application is that it has correctly identified a character or data field) are typically looked at by a human operator for approval.
6. **Output:** This is the connection of a document and data capture application to third-party software like line-of-business and ECM systems. Application programming interfaces (APIs) can be used to write these connections. Standardized XML formatting can often be used in lieu of customized development.

Setting user expectations

One of the biggest challenges associated with automated data capture is unrealistic expectations by users. Many users expect 100% accuracy from their applications, when the purpose is actually to reduce human data entry costs and mistakes, not eliminate them. For example, if an invoice capture application can automatically capture more than 80% of the required fields at a high enough confidence level that no one has to verify them, key entry is only now needed for less than 20% of the fields. (Each field can be given a rating by the software based on how sure the capture system is that it has recognized the field correctly. Typically, higher numbers mean higher confidence.) This should create a reduction in labor (not an elimination) or a comparable increase in individual productivity to eliminate backlogs, which will more than pay for the cost of the software. Oftentimes, systems can be optimized so that 90% or more of the fields in a given application can be captured automatically, producing an even greater ROI.

There are best practices that can be employed to improve the accuracy for automated character recognition. This includes creating form designs that are both respondent and processing friendly. The forms should be designed not to intimidate or confuse a respondent, but they also should be designed with colors that are easy to drop-out, large enough fields so that separate marks don't run together and clear anchor points and registration marks so that recognition software can establish a consistent frame of reference.

The quality of printers and thickness of paper should also be considered. If respondents are going to be downloading a form from a Web site and printing it at home, there needs to be more margin for error than if a professional printing service is being used. Also, a double-sided form may need to be printed on thicker stock to prevent text and graphics bleeding through to the other side.

Dictionaries and database look-ups should also be used in the data capture process when possible, along with rules, such as "a Social Security number is nine digits long." It's also important that users check their accuracy at the field level vs. the character level to get a true picture of how well their data capture system is working.

Calculating the cost of downstream mistakes is another factor that can come into play. Every data capture system, whether based on manual keying or automated recognition, is going to have errors. Efficiently managing the system involves calculating what it is worth to prevent those errors.

Mistakes on data captured from customer satisfaction surveys, for example, are not going to be as costly as mistakes from money transfer forms. So, in the customer satisfaction application, confidence levels for passable data could be set lower so that there is less manual intervention. In the money transfer application, confidence levels would be set higher and more money invested in QA. At some point, there is a diminishing return on QA investments when compared to the cost of an error in the data being captured. Determining that point is key to getting the most value out of an automated data capture application.

Mobile and touchscreen capture

Like any technology-driven application, automated forms processing is subject to changes and evolution in the IT market at large. And, the increasing adoption of mobile computing devices and touchscreens is affecting the market. One trend is that more and more users are looking at utilizing their phones in lieu of document scanners.

There are already several popular apps for capturing data from business cards utilizing the camera function in smartphones. There are also apps for capturing receipts for expense reporting. Intuit even offers an app that enables end users to capture an image of a W-2 form with a mobile device. The information on the image is used to automatically populate the correct fields in an online tax form. Once an end-user validates that the fields are correct, they can e-file by hitting the "send" button on their mobile computer. In these applications, OCR is typically done on a server or in the cloud and users perform the QA on a smartphone interface before submitting their data. More mobile-driven document capture applications are on the way.

One advantage of working with touchscreens is that they enable users to capture data by highlighting information with their fingers. OCR has to be applied to the image in the background, but after it is, utilizing the touchscreen, a user can populate data fields by touching relevant words and numbers on a document image. For example, a user can take an image of an invoice, and, when prompted, touch all relevant fields, like "company name," "total amount," "line-item detail," etc. – any word appearing on the document. This can be thought of as entry-level data capture. It's less expensive than automated data extraction, but still saves time in scenarios where knowledge workers need to easily get documents and relevant data into a back-end system.

Time is right to adopt automated data capture

The bottom line is that automated data capture is a maturing area of application. Once thought to be unreliable "black magic," more users than ever are now relying on technologies like OCR, IDR and automated classification to reduce their costs and improve their document-driven processes. In fact, in the 2009 Benchmarking Study conducted by TAWPI/IAPP, it was reported that while in 2006 only 68% of document processing sites were using some form of automated recognition technology, in 2009, 88% were.

So, adoption is clearly on the rise. As we noted, the technology is not perfect, but if expectations are set correctly and best practices followed, there's a good chance that users can achieve a measurable ROI on their automated data capture projects within 18 months, which – in a world where less than 20% of IT projects are described "successful" and estimated ROIs often run out over several years – seems like a pretty solid bet.

Also, while automated data capture may not quite bring the returns and the efficiency that implementing entirely electronic transactions processes might, it's clearly the next best step. It's a way to transition toward electronic transactions without forcing your customers and partners to change their avenues of conducting business with you. They are happy because they can continue to use paper like they always have, and you are happy because your automated data capture system enables you to more efficiently process their paper.

ABBYY Document and Data Capture Solutions

ABBYY develops technology that transforms information and enables greater productivity for business and personal computing applications. From data capture, to OCR to linguistics solutions, ABBYY solutions streamline with unmatched accuracy – leading the way in the global information revolution.

Established in 1989, ABBYY operates 14 global offices and employs more than 1,200 people. Used by more than 30 million people, ABBYY products have won hundreds of worldwide awards. Many leading hardware and software vendors incorporate ABBYY's leading-edge technologies. Every year consumers, business and government authorities process more than 9.3 billion pages with the help of ABBYY products. Annually, ABBYY saves people 970 million man-hours or $4.8 billion.

ABBYY offers the following solutions to meet virtually any document and data capture needs:

FlexiCapture
A powerful data capture software that works with precision accuracy to convert paper and image documents into business-ready data. ABBYY FlexiCapture automates resource-consuming tasks such as data entry, document separation and classification to significantly reduce the time it takes to deliver the data to business processes. ABBYY FlexiCapture's state-of-the-art architecture scales from cost-effective standalone installations for small businesses and departments to distributed client-server systems for enterprise and government projects. In addition, ABBYY FlexiCapture can be integrated with back-end systems and specific business processes to further improve efficiency and reduce operating costs for organizations of all sizes.

FlexiCapture Engine
An SDK for integrating data and document capture technologies into server, desktop applications as well as mobile and cloud infrastructure, the FlexiCapture Engine delivers comprehensive data capture functionality. It combines technologies and tools for processing forms as well as semi-structured and unstructured documents. In a single developer environment, it delivers data verification, document classification, mobile data capture and export to backend systems (ERP, CRM) and archiving (ECM, ERM). No matter how complex documents may be, FlexiCapture layouts are specially developed to analyze and identify data even when your documents are only partly structured. The flexibility of the FlexiCapture Engine allows for stable and reliable multi-page processing and intelligent handling of long tables.

Mobile Data Capture Solution

ABBYY Mobile Data Capture Solution turns camera-equipped mobile devices into a powerful entry point for data input. Corporations save time and money by enabling customers to input and verify data, then launch processes, via their mobile phones. Mobile Data Capture provides the ability to capture machine-printed text (OCR), hand-printed text (ICR), checkmarks (OMR) and barcodes (OBR) from the snap shots of IDs, driver licenses, receipts, bills, W-2 forms and other structured and semi-structured documents. The architecture of this solution includes a mobile front-end with advanced photo imaging technologies and a server- or cloud-based backend for accurate resource-intensive document recognition, classification and data extraction.

Recognition Server

A server-based software for automatic conversion of paper documents and document images into fully searchable electronic text suitable for archiving, e-discovery and enterprise search. Robust and powerful yet simple in deployment, ABBYY Recognition Server is ideal for organizations that are looking for an efficient document capture solution. It provides a complete set of tools for scanning, recognition, conversion and delivering the electronic versions of documents along with metadata to business workflows and ECM systems.

Learn More

For more information or to request a demo, go to **www.abbyy.com** or email **sales@abbyyusa.com**.

About the Author

Ralph Gammon is the editor and publisher of the Document Imaging Report, a semi-monthly newsletter covering business trends on converting paper processes into electronic ones, as well as the popular industry blog Document Imaging Talk. He has been covering and analyzing document capture and automated recognition technologies and implementations since the late 1990s. Ralph has seen automated recognition evolve from a technology in search of a solution to an enabler of a multitude of business process improvements in several vertical and horizontal markets.

1 From Handprint Data Capture in Forms Processing: A Systems Approach by K. Bradley Paxton, Fossil Press, Rochester, NY, 2011.

ABBYY USA
880 North McCarthy Blvd., Suite 220
Milpitas, CA 95035, USA
Tel +1 866.463.7689
Fax +1 408.457.9778
sales@abbyyusa.com

www.ABBYY.com