

Document & Data Capture

An independently prepared White Paper
Headway Technology Group

Document & Data Capture

Contents

Background	1
The Capture Process	2
Document Preparation	3
Image Capture	3
Data Extraction	5
Manually Entering Information	6
OCR/OMR/ICR and Barcode	7
Forms	9
Imported Sources	10
Exporting Data	11
Systems Management and Reporting	12
Reliability and Scalability	12
Audit and Reporting	13
Conclusions	14

Background

It is inescapable that every company generates and processes information stored on paper based documents. About 95%⁽ⁱ⁾ of a typical company record inventory consists of hard copy documents.

For over a decade, the Industry has provided a variety of solutions for electronically managing information from hard copy. Traditional solutions include:

Workflow Systems

Automatically route images of business documents, such as letters and order forms, throughout an organization.

Imaging Systems

Manage stored images, which provides more sophisticated classification and retrieval functionality than traditional paper storage and retrieval systems.

Document Management Systems

Save every revision of a document, enabling an organization to track data and text as multiple changes are made (by multiple people) to a document.

Storage and Retrieval Systems

Store documents with index tags, such as a customer name, ID, and phone number, enabling operators to retrieve quickly information from a database using one or more index tags.

Vertical Business Applications

Meet the specific data processing requirements of vertical markets, such as insurance and health organizations, by providing customized applications coupled with selective imaging hardware.

These technologies are frequently referred to as document management systems, document-imaging, content management, electronic content management, workflow solutions or knowledge management systems. Whatever the chosen name, they all provide a mechanism to manage large collections of documents with an efficient way to:

- Save documents in substantially less physical space while preserving a copy of the original image in electronic format indefinitely.
- Access the information stored in documents quickly, easily, and simultaneously.
- Retrieve information from multiple internal or external computer networks.

Clearly, document-imaging solutions offer excellent business benefits. Unfortunately, however, these solutions often have limited functionality with respect to volume capture; they concentrate on the management, storage and delivery of documents once within their control, but do not necessarily provide the best scalable or enterprise specific configuration tailored to a company's unique capture needs. It is for this reason that Document Capture has evolved in to a business in its own right.

The Capture Process

Information Capture is defined as the process of converting information stored on paper documents, in fax directories, and other electronic formats into digital data that can then be processed by a variety of technologies and stored for future retrieval.

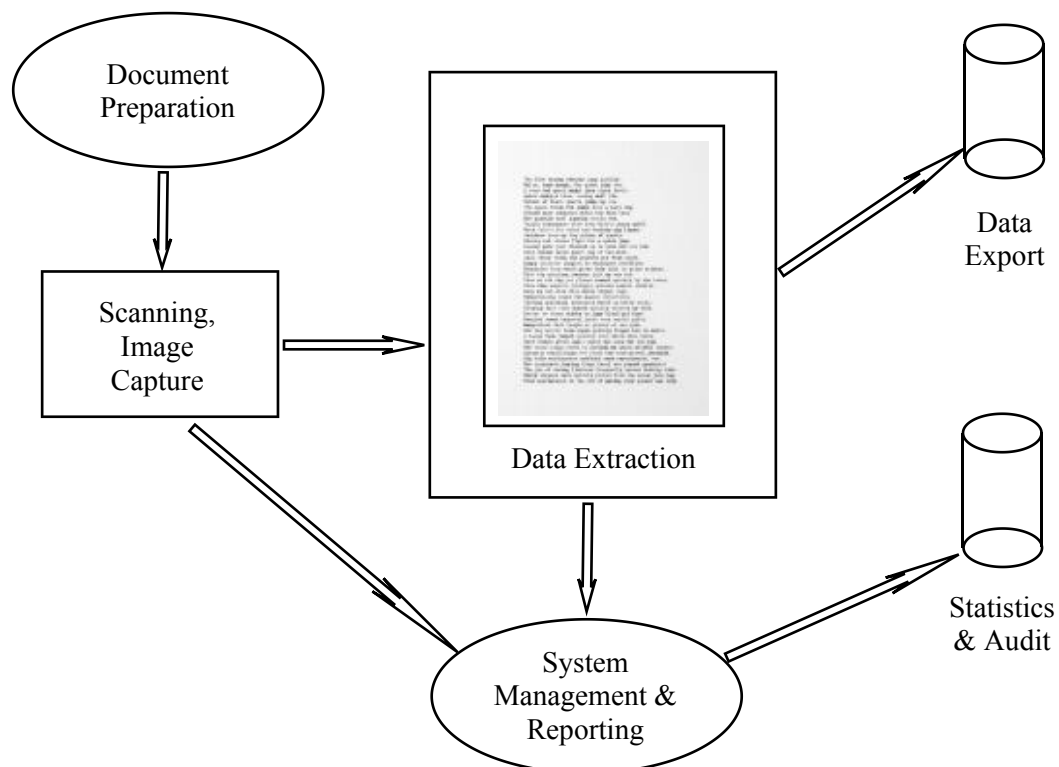
Before a document's information becomes useable "data", several different tasks must be performed. The tasks applied to each document and the order in which those tasks are run, comprise the flow of the capture process, or Capture Flow. The capture flow is a critical concept as not every document undergoes the same set of tasks, neither does every business process require that the same information be retrieved from documents.

This is not to be confused with workflow; the stages data undergoes during its useful life within the main line of business application.

A typical computer system that performs information capture has the software and hardware components that execute the following functions:

- Document Preparation
- Image Capture
- Data Extraction & validation
- Data Export
- System Management and Reporting

Overall Process



Document Preparation

The professional scanning bureaus frequently dedicate more man-hours to preparing the documents than they do to actually scanning them. This initial physical stage requires staff to examine and prepare the documents for the capture process. It includes removing staples or other physical items that are unacceptable, while ensuring individual pages are free of tears and folded corners. In most cases, it is vital to pre-sort or create logical groups of pages prior to the scanning process; almost all capture applications process discrete batches of work rather than accepting pages on an ad hoc basis. For example; separating the claim documents from the warranty forms enables a more streamlined and ultimately, a more efficient environment. This also provides an initial stage of accountability and audit; vital in many business models.

Image Capture

Scanning a document produces a raster (graphical) image that can then be stored on a computer as a digital representation of the original. When choosing a scanner, there is a range of criteria that should be given due thought: The size, volume and quality of paper as well as overall running costs should all be considered before selecting a scanner. The ability to use a wide range of scanners is one of the defining characteristics of a good imaging system; there are in excess of 250 scanners designed for volume-based scanning that are in common use.⁽ⁱⁱ⁾

It is worth considering the advantages an Automatic Document Feeder (ADF) brings. This device allows a stack of paper to be placed into a tray or hopper and automatically fed one page at a time into the scanner, speeding up the scanning process significantly. The majority of scanners without an ADF are designed for imaging graphics and are unsuitable for document capture. However, certain documents; badly damaged or torn, books and pages with additional notes physically attached, etc. may well require scanning on a more traditional flatbed scanner.

Scanners can handle a variety of paper sizes, from business cards to engineering drawings. Most offices only need to scan documents up to A3 although for organizations or departments that use plans or architectural drawings, there are larger-format scanners that support up to A0 documents – albeit slowly.

The speed or throughput of the document scanner is worthy of consideration. Typically, document imaging scanners handle between 6 – 200 pages per minute in simplex or duplex mode. Duplex scanning allows both sides of a page to be scanned in a single pass. Obviously, higher speeds and duplex scanning increase the price of the scanner. In some instances, two 20-page-per-minute scanners rather than one 40-page-per-minute scanner offer significant advantages for software or operational reasons. Be aware that not all document capture systems support multiple scanners while some have licensing or performance issues.

Advances in technology have allowed scanner manufacturers to produce devices that are able to capture pages in full colour, frequently in addition to the traditional black and white image. This can offer significant advantages to the viewer – a colour image often contains useful information that is lost in black and white reproduction – although due consideration should be given to the increase in file size and processing time such images demand. Again, some capture software is able to accept both the colour and black & white renditions, routing both through the data extraction process and delivering multiple images, each optimised for Internet or traditional thick-client use.

Most vendors can provide an indication of the recommended duty cycle for their products and this should be taken into consideration; a higher duty cycle machine uses more robust materials and undergoes more stringent testing at the design stage and this is usually reflected in the initial purchase price. However, the payback of far less unscheduled downtime is often far more valuable for many operations.

It is important to have the ability to enhance scanned images by applying certain image clean-up techniques such as increasing contrast, cropping borders, rebuilding broken characters and removing noise. It is becoming increasingly common to find some of these technologies built-in to scanners and the fully-automated variants relieve the operator of tedious adjustments to the scanners settings and should be given due consideration. It is worth noting that subsequent modification of an image may render it inadmissible in a court of law unless a secure audit can prove what was done, by whom and when – this is particularly relevant in the case of blank page deletion (where the software automatically destroys any pages it considers ‘empty’). To cater for this, some capture systems permit the export of the original image and or subsequent modified variants.

Also considered an intrinsic part of the capture of images is quality assurance. Implementing checks for image rotation, order and optical quality good enough for accurate data capture should all be considered.

It is imperative that the capture software is able to offer a rescan facility. Images with poor quality, incorrect rotation or other rectifiable problems should be reprocessed without interrupting any current tasks or adding avoidable delays to the overall process. Good document preparation makes a significant impact to the failure rate; incorrectly rotated pages, damaged or badly skewed documents can be avoided in most cases by better preparation.

There are two schools of thought regarding rescanning: On-demand and off-line.

On-demand correction uses software that corrects the image at the time of scanning. Because this interrupts the scanner, it is more often used in low volume sites where scanner throughput is not a prime concern.

Off-line rescanning is almost always preferred for a number of reasons:

1. The investment made in high performance scanners is better realised because the machine is working more of the time.
2. A different scanner may be more suited to the task; using a flatbed rather than an automatic feeder, for example.
3. Scanners have a huge range of user-adjustable settings that affect the resulting image and a dedicated rescan operator will learn which options are appropriate for each image, thus minimising the time spent on this task.

Rescan Flow Diagram

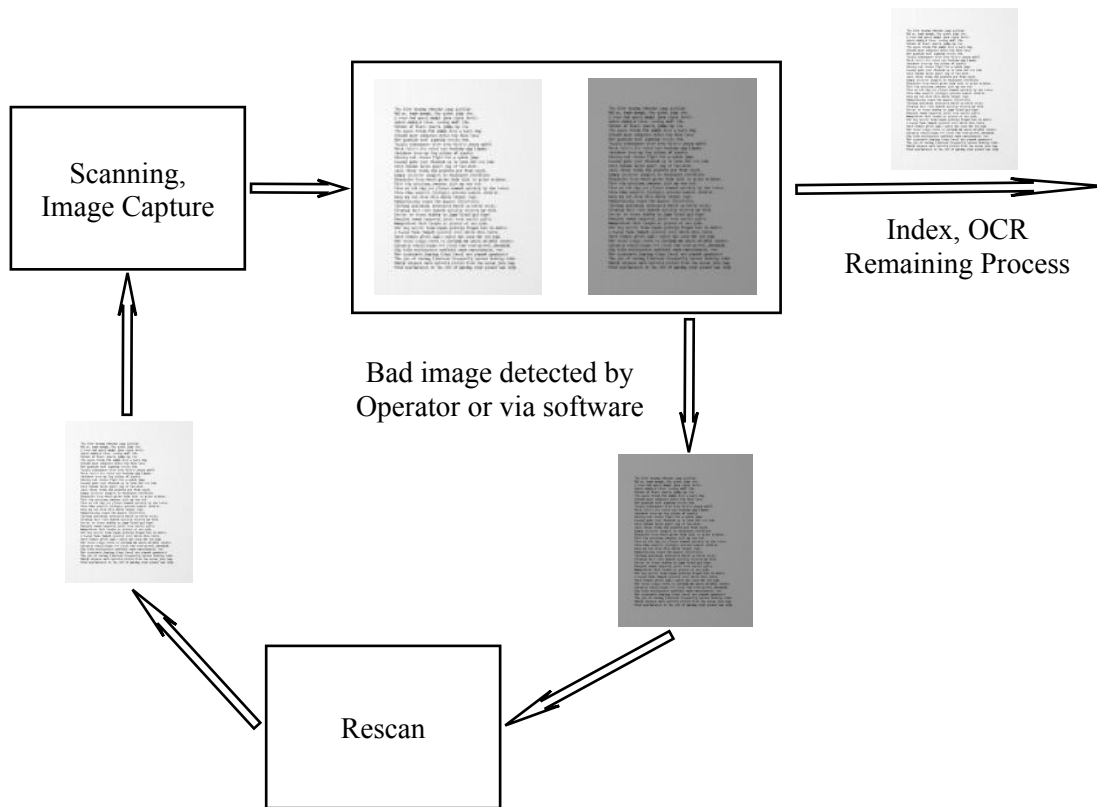


Image Capture may include an element of importing electronic documents (e.g. faxes, emails, word processing documents and the like) and although the requirement for a scanner is negated, many of the Document Capture software vendors integrate it at the scanning stage of their process to take advantage of the data extraction and validation rules in the subsequent stages.

Data Extraction

When paper documents are received in an office, they must be organized to be useful and are sorted, labelled, stapled, placed in folders and filed in a cabinet. Without these steps, nothing could be found in a busy workplace and electronic documents are no different. A document imaging system must have a comprehensive indexing system that organizes documents for future use and, as such, fast, accurate extraction of useful data is the cornerstone of all good capture solutions; there is simply no point extracting erroneous information no matter how quickly it can be done, while taking an inordinate amount of time to ensure the data is accurate will also negate the investment in technology.

There are a number of ways to associate information with an image:

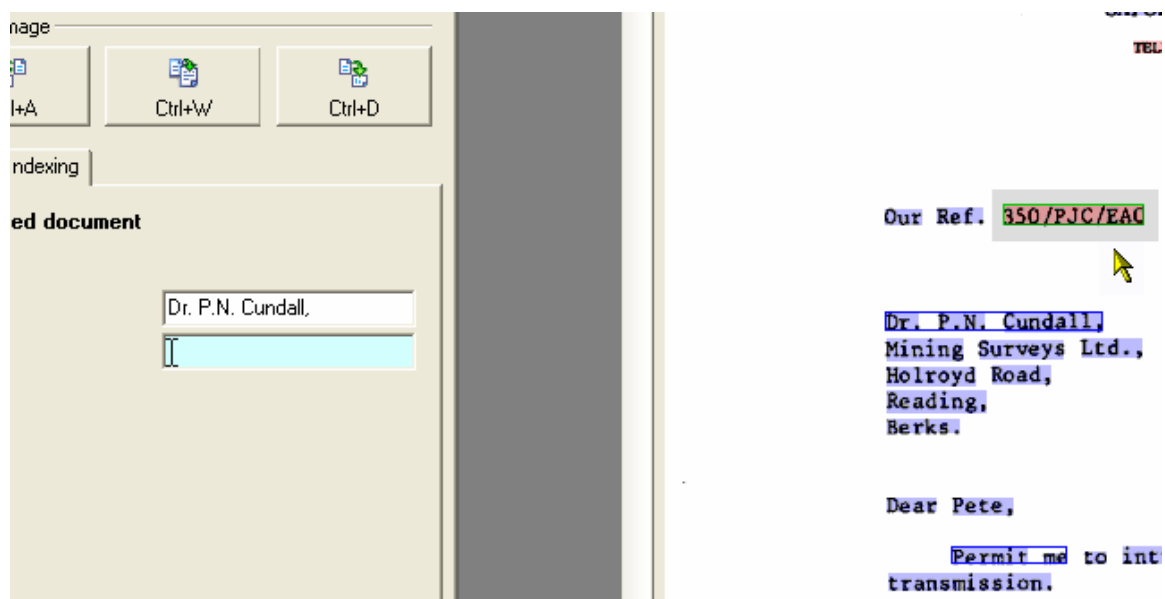
- Manually entering information that identifies captured images and files
- Applying optical or intelligent character recognition (OCR/ICR) and barcode recognition technologies to images to extract alphanumeric data. This could also include recognising tick boxes or multiple choice options.
- Applying forms processing technologies, such as form identification, to differentiate between different documents.

Manually Entering Information

Traditional data entry, or Key from Image, can be laborious and expensive but has the advantage of being highly accurate. The number of fields and their average length form the basis for calculating the overall data entry costs and research indicates that operators are able to input between 8,000 and 11,000 characters per hour for alphanumeric input.⁽ⁱⁱⁱ⁾ Key from Image operators – indeed any data entry – are subject to a natural error rate and this must be factored into the solution.

What price data errors? An industry average for single-operator errors is 2.2%,^(iv) although double-key entry (where a second operator also keys the data and the system compares the two values) drops this significantly. Clearly, for system critical fields, double-key entry should be considered mandatory. Additionally, the guidance of the BSI document *BIP0008*, should be taken. Paraphrasing this: the operator that scans the document should not be the same as either of the clerks that index data if the system is to comply with current standards.^(v) However, a general rule of thumb is to have three fields per document to minimise retrieval failures (the odds of all three fields on one document being mis-keyed is statistically very low indeed).

Notwithstanding, Key from Image is ideal for disparate documents. By their very nature, these are difficult to automate and so require human intelligence, thus requiring keying operators that are conversant with the terminology of the documents. As such, this frequently dictates the use of a knowledge worker rather than a more cost-effective data entry clerk.



Some tools exist that speed the keying process and one very powerful example of this is OCR-assisted index. In this example, the software OCRs the document and presents the image and the underlying text to the keying operator who uses the mouse to click or drag over the text they wish to apply as index data. Using this type of technology keying rates can exceed ten times that of industry averages, minimise training costs and so reduce the total cost of ownership of the capture system.

Whatever the base technology, one significant advantage of Key from Image is interactive lookups, whereby data from one or more fields can be used to query a remote source (e.g. database, file, LoB system), thereby supplying reference data, additional information and suchlike. By permitting interaction with a trained operator, lookups can simplify data capture and increase accuracy.

OCR/OMR/ICR and Barcode

Machine-print Optical Character Recognition (OCR) and barcode reading are established forms of automation. Accuracy rates are subject to variables outside the scope of this document such as scanner design and source document quality, but tests indicate that 0.04%^(vi) of OCR and 0.00002%^(vii) of barcodes will be misread even under ideal conditions. Practical UK experience suggests that raw, uncorrected, OCR will offer somewhere around 10-30% rejects while barcodes are successfully read providing the image quality is good enough.

It is important to note the difference between substitution and rejected errors with respect to OCR failure rates: The rejected error rate is often the one quoted as it reflects the number of characters misread and recognised by the software as a misread; a substitution error is classified as a misread that is interpreted incorrectly and not flagged as such. An example of this would be if the software were to process a 'c' and deduce, wrongly, that it represents 'e' and the mistake is simply never identified.

In addition to this, there is a substantial difference between the failure rate for characters and fields. Consider a document that has 10 fields, each with 10 characters and the OCR engine claims a character failure rate of 2%; i.e. any single character has a 1 in 50 chance of being wrong and so giving a potential field-level error rate of 1 in 5 or 20% for this example.

MR
DAVID SUTHERLAND
124 HAMPTON ROAD
PERSHORE
MERCIA
WS1 1AA
0121 999451
0121 999400
03121957
NJ 78 56 34 V

Space missed

**100 Characters with 2 misread.
A 20% field-level failure**

Read 'U' not 'V'

This is clearly unacceptable in many environments and additional checks, cross-references or external validation is often recommended. Zone-based OCR is frequently used to associate different index fields with unique areas of a document. For example: If a reference number is always printed in the top left of a document, it is entirely possible to configure many systems such that the region is OCR processed and the resulting data applied to the relevant field. Clearly, it is critical that each document satisfies some basic criteria for this technique to work, namely: All the documents must have the data in the same place and use a readable font, the image should not exhibit any skew or positional inaccuracy and ideally, each data item should be verifiable with a secondary field or against a defined rule. It is entirely possible that manual keying, even with all its additional overheads, is a justifiable approach.

A logical extension to zonal-based OCR is to read the entire page, typically used for Full-Text Portable Document Files (the format created by Adobe) and retrieval systems that offer a full-text searching option: It can be far more productive to search for a word or phrase within a large report than rely solely on the title, for example.

Optical Mark Recognition (OMR) technology detects the absence or presence of a mark and can be considered over 99% accurate^(viii) once configured – the notes above regarding positional accuracy apply to OMR even more stringently as checkboxes tend to be placed in close proximity.

Each capture technology provider will furnish their own specifications for the bounding box, but in general, it is considered good practice for the ‘walls’ of the tick-box to be two pixels wide. Assuming 200 d.p.i. (dots per inch), this would indicate a wall thickness of 0.01” (roughly 0.25 mm) as this gives the recognition software a reasonable chance of recognising the borders on all sides and so differentiate customer data from the background. Another solution to this is to use drop-out inks to make the form’s static information fade out upon scanning, leaving only pen marks. This can have legal ramifications as the stored image is no longer a reasonable representation of the paper copy and additionally, may require a redesign of the form.

Intelligent Character Recognition (ICR) is a logical extension of the aforementioned recognition technologies, further developed to read handwritten data. Although greatly affected by the quality and clarity of the original text and image, better rates of recognition are possible by using contextual information. Recognizing entire words from a dictionary is easier than trying to parse individual characters, while reading the ‘Total’ line of an invoice (where the data is always numeric) is an example of a smaller dictionary where accuracy rates can be increased greatly. Knowledge of the grammar of the language can also help; it is possible to use tri-grams; various three letter groups that occur within a language. For example, in English, “ion” is used far more frequently than “dle” and various deductions can be made that improve the overall accuracy.^(ix) Indeed, these techniques have been proven with traditional OCR. It is considered good practice to constrain the writing with combs or boxes, thereby encouraging the writer to space letters out and print at a reasonable size and so maximise the successful recognition rate. Again, accurate page registration is a pre-requisite for any ICR application.

Currently, cursive handwriting is very difficult to recognise with any reasonable accuracy and although there are various organisations examining this sector of the market, we should not expect to see any significant developments for a number of years.

Forms

Forms recognition grew out of zonal (field-based) OCR and the early products simply matched text against expected results and so differentiated one page from another. The technology has developed significantly and most applications now perform page recognition by multiple means:

General overview of the page (Dynamic ID): The technique usually involves building a black and white histogram of the page and comparing it to a library of pre-compiled templates. If the whole image fails to provide a match, many applications will then break the page into quadrants and compare each section, yielding better accuracy. Although the approach is fast – histograms are very small files and quick to create – it is possible to get false positives, so applications tend to provide some form of confidence indicator. Pages falling below a pre-set level are then submitted to a secondary identification process.

Content matching or Keyword matching: A ‘brute force’ approach that OCR’s the page and compares the resulting text with the master document(s). This is obviously more processor intensive, but the success rate is very high. It is frequently the only way to differentiate between pages that have a large number of similarities. Again, applications should return a confidence indicator to isolate the borderline documents.

The image shows a scanned document titled "Application Form". At the top, there are several input fields: "Branch/Agency", "Branch/Agency Code", and "Seller Code". Below these are checkboxes for "Home Loan", "Investment Home Loan", and "Pension Home Loan", followed by a label "Please tick ✓ type of loan required" and a numerical field containing "11819". A section header "1. Personal Details" is followed by two columns: "Principal Applicant" and "Joint Applicant". Each column has checkboxes for "Mr", "Mrs", "Miss", and "Ms". Below these are fields for "First Name", "Surname", "Customer Number", and "Address". At the bottom, there are fields for "Time of the month", "months", "years", and "months". A vertical label "point pen" is visible on the left side of the form.

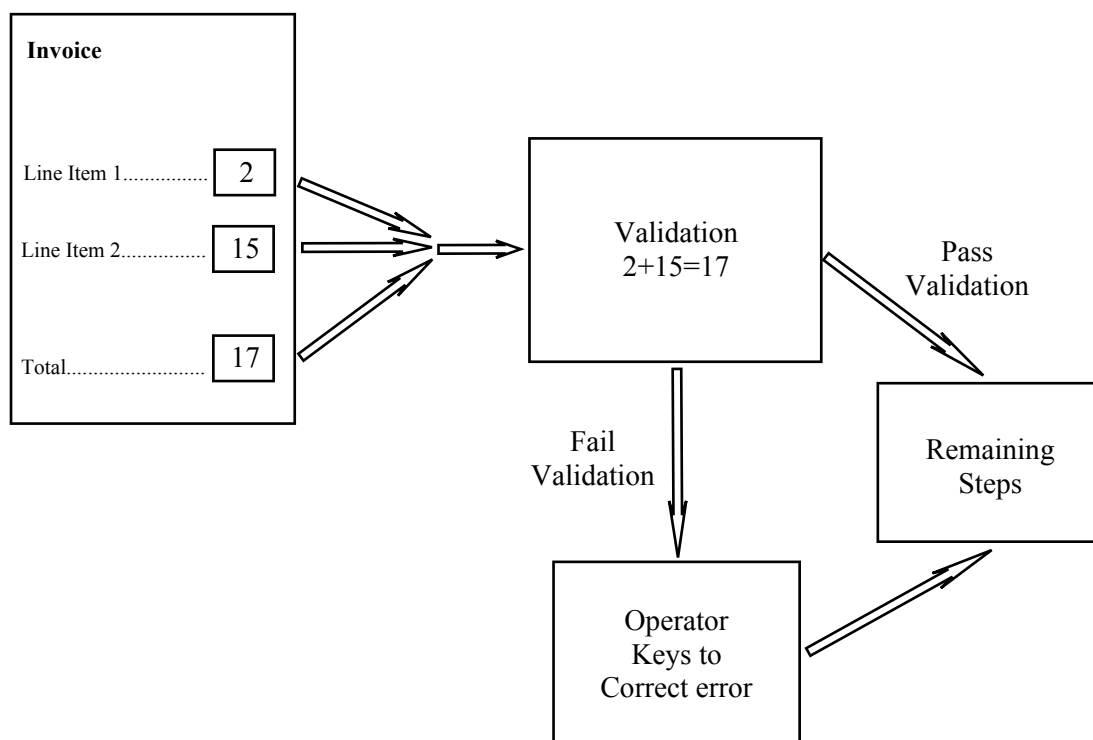
Whatever the technology, it is normal to furnish the process with a manual identification interface to allow an operator to hand-select on the odd occasion that the software is unable to make an automated selection. Some applications will ‘learn’ from the user. That is; a new page that has not yet been seen will obviously end up being presented to an operator and it is possible to then add the page to the library and subsequent pages of the same layout would then be automatically recognised. Other capture applications require the library to be manually rebuilt occasionally to add in new templates.

A rapidly emerging technology is that of Free-Form recognition. This is most often seen in the invoice processing arena, where documents will all contain the same fields (amount, date, invoice number for example), but the layout is substantially different. In many cases, it is simply not viable to create a library of all the variants one might see and so the Free-Form approach is gaining ground. This utilises OCR again and combines it with a set of rules that indicate that, when the word ‘Total’ is seen in the text, the next set of numbers should be read and stored in the ‘Total_amount’ field. There are a number of variations on this basic idea; while some products will only use

the tree of rules, others mix this with partial dynamic ID and create a library of commonly seen pages, thus making it faster to recognise frequently seen forms as the system is used.

Whatever the technique, any automated forms recognition will fail to correctly identify some forms and a manual 'escape route' should always be available. In addition, all OCR engines will fail to read documents that have been badly scanned or are very poor source material; one only has to think of delivery manifests and the physical damage that is very likely to occur to these pages to realise that it is not possible to read the whole document in every case. As such, stringent validation should be considered *de rigueur* in any environments where data integrity is important.

Escape Route



Imported Sources

By far the majority of capture systems today are dealing with paper documents, but almost all of them are able to take previously scanned documents in an industry-standard format and process them as if they had been scanned. Tracking the rise in electronic communications, it is becoming more important to accept standard office documents and process them in the same manner. Consider a purchase order: It could come in on FAX, paper or as an email, yet the processing that it undergoes should be identical. As such, the industry is moving towards a 'global document' approach whereby any file type can be accepted. Referred to as the Digital Mailroom, this carries significant benefits and streamlining for the larger office or department. In practice, to recognise and process documents, it uses all the technologies mentioned above in addition to a complex set of business rules and exception handling; it is the design and configuration of these rules that usually form the bulk of the implementation cost, often far exceeding the cost of the software components themselves.

It is worth considering what future requirements your company may place on a capture system; is it conceivable that voice data might be managed at some point, for example.

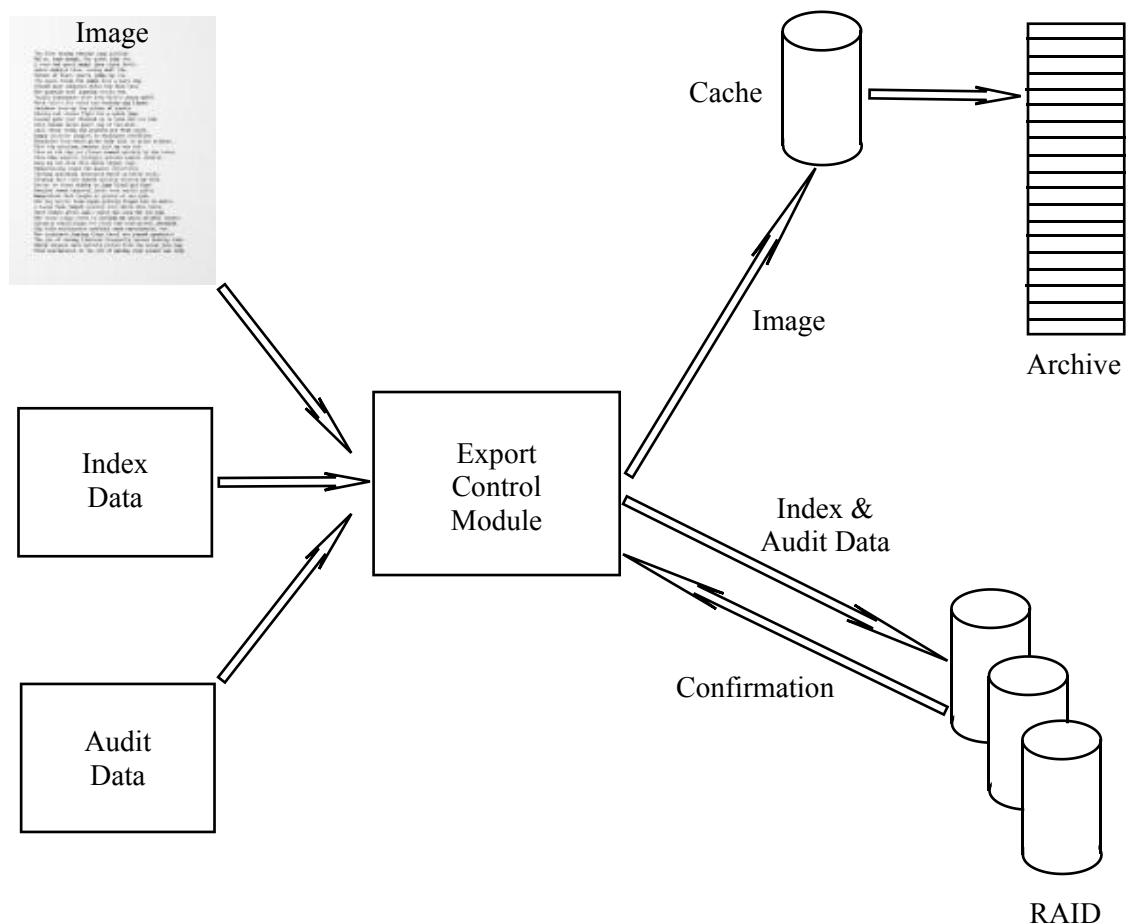
Exporting Data

Ultimately, the role of all capture solutions is to provide data in a useable format.

In a typical office, paper files are found by looking in a particular folder in a particular drawer in a particular file cabinet. It is very easy to replicate this structure using folders or directories on a computer and this has the added benefits of a low cost of entry and simple transition for office staff. As such, storing documents in computer folders is often a very successful solution but it also lacks the sophistication of more powerful solutions; there is no easy way to search multiple folders for a mislaid document, for example. The format of the exported image should be considered; a Group 4 Tagged Image File format (TIF) is universally accepted and recognised as an open standard, it represents the original image dot for dot; that is, it does not lose detail during the compression process and it can be read on a huge number of systems.

Adobe's Portable Document Format (PDF) is also gaining ground as an acceptable format, especially useful for Internet distribution because the format viewer is free and acts as a plug-in to most common web browsers.

Export Flow



Given that, by design, capture systems create a significant volume of information, it is often desirable to have some advanced form of data management system in place that can accept, process and store this traffic. The range of such systems is beyond the scope of this document however, it is imperative that the capture system is able to communicate effectively with whatever software is used; images (if required) and the data from a document must be transferred – ideally without human intervention – quickly and efficiently. While all capture solutions support a number of common data-exchange formats such as comma separated values (csv), Open Database Connectivity (ODBC) or extensible Markup Language (XML), what these common formats gain in apparent convenience, they lack in security, traceability and reliability. By definition, they export data to common folders or directories and this is not acceptable for a number of reasons. As such, most systems have the ability to interact directly with a number of popular repositories or workflow solutions, pushing documents and data into the document management software into more precise locations and offering better control of errors or events that may occur during the migration.

System Management and Reporting

Integrating and managing the capture processes, ensuring reliability, and reporting statistics and results are vital secondary tasks for any capture flow.

Reliability and Scalability

The capture software must be stable under all operating conditions. Will an increase in scanned volume add any impact to the system that cannot be accounted and allowed for? An increase in volume requires additional processing stations and only a few solutions offer a truly scalable product; most packages process a batch of documents at any one time and, as such, adding a second index client would not necessarily halve the time to index if only one batch of work is available. A more scalable product breaks work down in to individual pages or documents; only at this level will multiple index stations be able to share a single batch.

Depending on the environment, it might be worth rolling-out multiple, small scanning stations and linking them up to export into a centralised database. In these cases, it is worth investigating what options the capture software allows: Does indexing have to be done at a particular location or by a certain user? What flexibility is there in the configuration and will the software work in a way that is acceptable to both the business process and the users of the system?

It is also important to ascertain how the solution manages to recover from disaster and understand what impact any failure (software or hardware related) will have on the process as a whole. Naturally, stand-alone systems will cease processing images when there's a system failure but this should not be the case for production environments with multiple client stations. Naturally, client / server solutions leverage an architecture that is better able to handle network or hardware failures and should be considered mandatory in larger implementations; recovery from such faults is a standard feature of these systems but is usually not mentioned (or well handled) in workgroup solutions.

Licensing capture software is usually a function of the number of images that are scanned or imported; a double-sided page is typically counted as two images. Unless the business has a very steady and predictable monthly volume, it would be wise to consider an annual license model as this caters for variations in the workload without forcing the purchaser to buy a license large enough to cater for occasional peak month. Some applications (typically client / server types) will use concurrent licensing rather than a machine-specific route. This would allow a more flexible working environment as multiple users could share one access license if their working patterns permitted it. This approach has another distinct advantage for the larger site: it becomes possible to create a 'standard build' computer and so reducing the level of specialist MIS skills and so saving money and time.

Audit and Reporting

In order to meet current compliancy guidelines, it is imperative that the capture software is able to track documents as they move through the system: Who scanned the documents; who indexed them and how long did it take; what, if any, further operations took place and finally, where was the resulting data sent. One should carefully consider if it is acceptable to modify an image at all; as an example, in certain environments, deskewing a document is considered a modification and the original image is also retained. This could be an issue with blank page deletion; under most guidelines an audit of events like these is essential; in practice, after image compression, a blank page takes up so little disk space it is not worth the overhead involved in deletion.

Ensure meaningful audit information can be extracted at varying points during the process and not simply at the end of the day. This type of log does not usually contain the granularity of information that one requires and may be better thought of as status reports and suggest the software was designed for low-volume workgroup processing.

Typical Audit Data Report

Batch	Doc	Scan Op	Index Op	No. Of Keys	Index Time	Export Time
184252						
	1	Alan H.	Peter B	48	21/07/2004 11:24:15	21/07/2004 11:24:15
	2	Tony L	Peter B	42	21/07/2004 11:25:01	21/07/2004 11:29:10
	3	Anne T	Peter B	15	21/07/2004 11:25:31	21/07/2004 11:30:18
184253						
	1	Anne T	Geoff B	10	21/07/2004 13:10:48	21/07/04 13:11:55

Conclusions

In an information capture system, several different tasks must be run before a document's information becomes useable "data," which can then be released for use by a backend application.

Within a particular Capture Flow, several capture tasks are performed. As we have seen, these could include: manual data entry; data validation and quality checking; automatic data gathering, such as recognition of barcodes and form fields; image enhancement; and export. As each document can require its own unique set of capture tasks, it is important to ensure the capture processes be fully customisable in order to meet current and foreseeable requirements. It is important to consider what standards must be adhered to and choose a platform that is able to meet or exceed these obligations.

The following table summarizes the most common problems encountered by companies that purchase information capture systems today:

Requirements	Important considerations
New functionality	Ensure the capture system manufacturer frequently incorporates new technologies and products from a range of suppliers
Functionality not currently available or need bespoke development	Ensure that the information capture system provides customisation tools, such as Module Development Kits, in the language or languages you use
Process a high volume of pages	Choose a scalable information capture system that has a significant and demonstrable number of large volume sites
Support new scanner(s)	Ensure the capture system has made a commitment to support industry-standard scanner drivers
Process different types of document capture jobs	Focus on information capture systems that allow you or your solution provider to establish necessary process flows, rather than rely on the software vendor
Support for a particular document management system	Ensure the capture system is from a vendor who can readily demonstrate suitable software connectivity and has an active partner arrangement
Complex business rules	Flexible Capture Flow that can be modified by you or your IT partner
Compliance with, for example, Sarbanes-Oxley. ^(x)	Platform that is statistics- and audit-rich that can be exported with ease

- i. Captiva Software Corporation, Feb 1999
- ii. Pixel Translations July 2004
- iii. Various sources
- iv. *Census 2000 Testing, Experimentation, and Evaluation Program* July 22, 2003. Titan Corporation Kevin A. Shaw, Project Manager Planning, Research, and Evaluation Division
- v. *A Code of Practice for Legal Admissibility and Evidential Weight of Information Stored Electronically*. 2004. BSi Alan Shipman.
- vi. *Census 2000 Testing, Experimentation, and Evaluation Program* July 22, 2003. Titan Corporation Kevin A. Shaw, Project Manager Planning, Research, and Evaluation Division
- vii. Researched by Ohio University: Centre for Automatic Identification on Code 39. Date unknown.
- viii. *Research in Optical Mark Recognition (OMR)* April 2004. U.S. Census Bureau Acquisition Division.
- ix. Recognition & keying methodologies technical white paper. 2002. Neurascript Ltd
- x. The One hundred and seventh congress of the United States of America 23rd January 2002 passed the Sarbanes-Oxley act to protect investors by improving the accuracy and reliability of corporate disclosures made pursuant to the securities laws, and for other purposes.

Special thanks for the contribution made by Dave Evans

Headway Technology Group Ltd
Headway House, Crosby Way, Farnham, Surrey, GU9 7XG
Tel: 01252 717071 Fax: 01252 741223
www.headway.co.uk