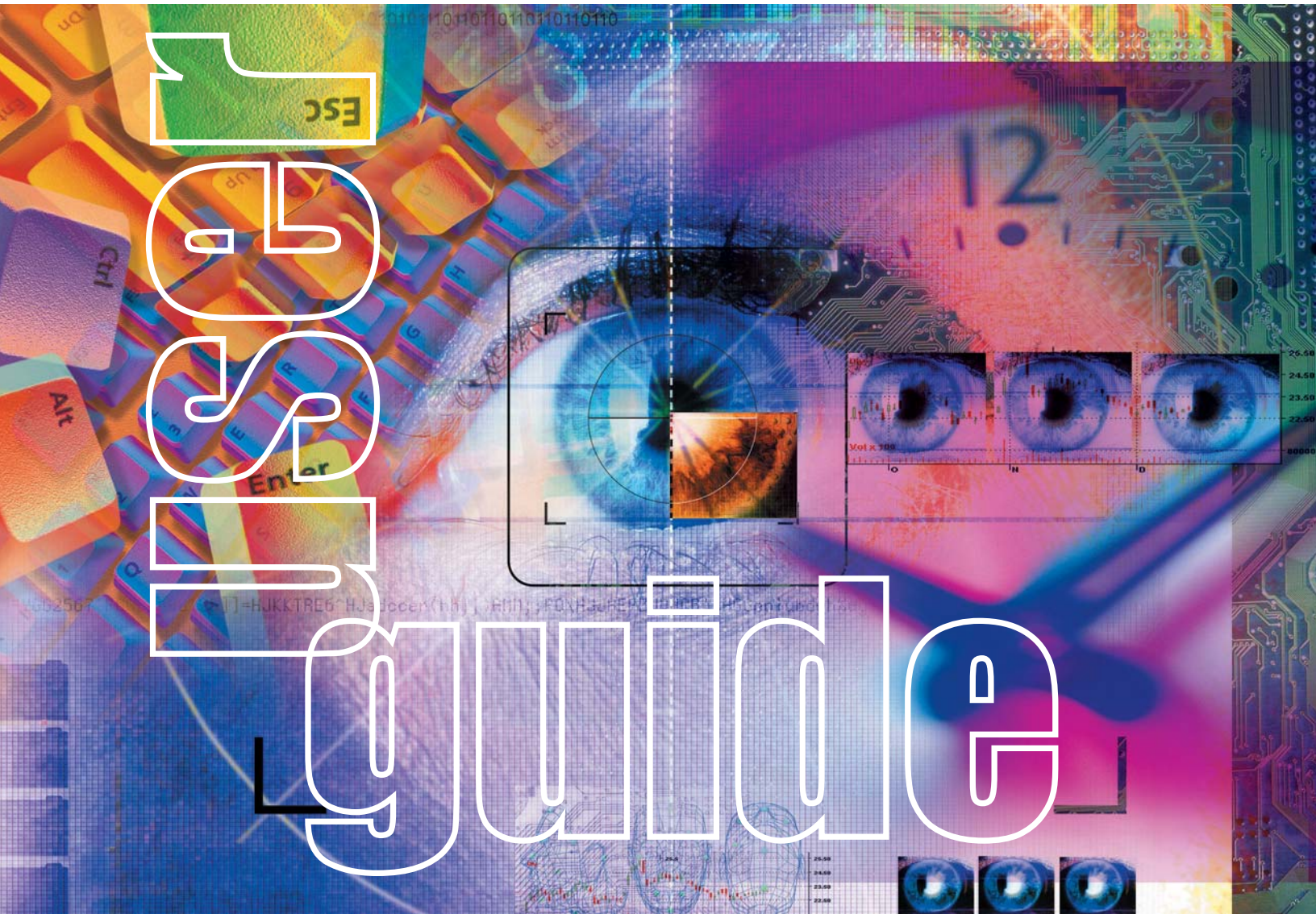
AIIM User Guide

Document and Content Capture



Authored by

STRATEGY PARTNERS

for

AIIM International
The Enterprise Content Management Association

Document And Content Capture

An AIIM User Guide
Authored by Strategy Partners

This is one in a series of User Guides from AIIM International, authored by Strategy Partners. They are intended to educate and inform potential purchasers and users of document and content systems at an initial level, and position the technologies within a business context. They are designed to explain:

- How document and content technologies work
- How they are justified in business terms, and what difference it makes to the bottom line
- How they are used operationally, and what constitutes best practice
- How they relate to, and integrate with other aspects of IT
- The roles of operational users, the IT function, system integrators, and other service providers in the document and content management space

Strategy Partners International Ltd.
Chappell House The Green
Datchet, Berks SL3 9EH, UK
Tel: +44 (0)1753 592787
Fax: +44 (0)1753 592789

Website: www.strategy-partners.com

ISBN 0-89258-396-7

Printed in the United States of America

Published by

AIIM International
The Enterprise Content Management Association

AIIM International Headquarters
1100 Wayne Avenue, Suite 1100
Silver Spring, MD 20910 US
Tel: 301.587.8202 /
800.477.2446
Email: aiim@aiim.org
www.aiim.org

Document and Content Capture
An AIIM User Guide

Introduction to Document and Content Capture

Document and content capture systems are fundamental to Information Technology. They provide a cost-effective, accurate, and operationally simple mechanism to feed content (e.g., news articles, pictures, and files) into Document and Content Management systems, as well as into key business applications such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), eBusiness, and eGovernment.

Without effective capture of documents and/or their content, information-based processes starve and die. eBusiness is still business. How user organizations transfer business transactions and customer information into processes and knowledge bases makes a key contribution to the overall success of the organization.

Document and content capture may be said to have come of age. The core technologies are mature and the understanding as to how to deploy them effectively is well developed. Such systems have developed a strong reputation for delivering pragmatic return on investment, as well as for equipping the organization to face the vision that is the eBusiness age.

This Guide sets out to explain in straightforward terms how document and content capture works, where and how it delivers real benefit to organizations, and the key current and emerging applications and technologies that make it an investment for the future as well as for the present.

Document and Content Capture: What Is It?

Document and content capture is defined as the set of technologies and services required to capture documents and information from documents—paper, microfilm, fax, and electronic—in order to process the content of the documents in a form that meets the need of the repository or business application being served.

As such, it represents a key capability, helping organizations adapt to the eBusiness age, and equipping them to harmonize paper-based processes with electronic approaches. It addresses the key issues of:

- Scanning and capture
- Quality Assurance (QA)
- Recognition (where needed)
- Output to repository or business applications

It does NOT include applications where data is keyed directly from paper, where information about documents (often called attributes, properties, or metadata) is entered into an indexing system for the purpose solely of tracking physical documents, nor does it cover every application with a forms-like user interface.

The input is a document—paper or electronic; the output is an electronic document, metadata, and/or data

from that document.

The processes within document and content capture can be described at three levels: (See Figure 1)

- Capturing incoming content—both paper and digital
- Processing that content to get it into a fit state for onward transmission
- Output or release of documents into awaiting repositories or applications

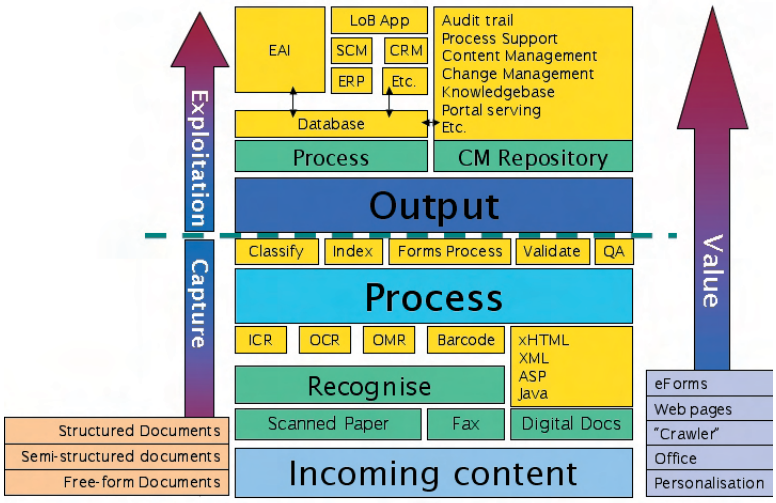


Figure 1: Document and content capture key processes

Note that each level of activity adds value to the document. A captured document on its own, without meta-data or a process to re-use the content, has a low value, while one that has been indexed and validated has a much higher value, depending on the extent and value of the process or repository application being served.

Capture of Incoming Content

Incoming content can be considered in four main ways: paper, microfilm, fax, or electronic. If electronic, the emphasis is first on recognizing the digital format of the document, and understanding how to handle it and capture its metadata without re-keying or recognition. If paper, film, or fax, the document is first scanned to create a digital image, from which one may need to extract part or all of the content using recognition technologies.

Processing and Validation

The capture and validation processes to be followed apply equally to both digital and paper content. Depending on application needs, there will be requirements to classify documents into meaningful groups, index them, process any forms, and conduct validation and quality assurance (QA) processes. For example, incoming post is scanned, Quality Assured, and indexed. An email with a similar content can be printed out and subject to identical processes, or its metadata and content captured digitally and then subject to those processes.

Output from Document Capture

Output can be transferred into a business process or into a repository—or to both (very common).

- Output to a business process application
  If output is into a business process, then decisions need to be made as to just how that is to be achieved—as part of a workflow system, through proprietary links, or via some Enterprise Applications Integration (EAI) middleware link. These need to be capable of passing the documents and/or data to one or more linked applications as required. Applications often require different metadata at different stages in an overall process.
- Output to a repository application
  Where the document is to be used in a CRM or Content Management (CM) application, managed as a business record, or eventually archived, then best practice is to enter it into a repository, in addition to any processes it may be supporting. Examples include financial transaction and "e-transaction" records, regulatory processes, and correspondence management.

Enabling Technologies

Image Capture

Image Capture is about taking paper documents and converting them into electronic images for subsequent re-use, re-processing, and/or storage in a Document Management System (DMS). Such systems provide facilities for:

- Batch management of documents
- Scanner management and control
  Includes the requirements for determining scanning resolution, types of image required, image compression, image processing and enhancement, and others.
- Running different types of scanners, according to need
  Today, scanners are available to fulfill a broad range of needs, from a hand-held text capture device to machines capable of handling 4,000 items per minute; from digital copiers with a scanning capability to high-power microfilm scanners; from credit card scanners to devices for capturing A to E-sized engineering drawings and maps—all in monochrome or full color, single-or double-sided. As such, it is possible to match the scanner—or mix of scanners—closely to the application.
- Determining output formats
  Digital output formats depend on a) how the document is to be stored, used, and viewed; and b) whether it is to be retained in its original, usually revisable format (e.g., Word or Excel), or rendered into a non-revisable format (e.g., Adobe PDF). Such decisions depend upon the application—an invoice processing application expects bitonal images, while a photographic library expects multi-tonal and color formats, for example. An issue arises when these document have to be viewed. Some users and applications require a dedicated viewer, and many can use a standard Web browser, for example.
- Quality Assurance
  Includes functions for scanner monitoring, batch management, and image quality checking.

Recognition

Recognition technologies are used to extract data from documents, for indexing, or for input to a business process. Recognition types include:

- Non-character recognition technologies
  These include Optical Mark Recognition (OMR)—detecting marks in boxes on forms; barcode recognition—used to match incoming completed forms to processes that generated the original forms and also to detect document types, etc.; and Patch Code Recognition—used mainly on document separators when preparing document batches for scanning documents.
- Character recognition technologies
  These include:
  - Optical Character Recognition (OCR)
    OCR is used for recognizing machine printed characters of good quality and across a set range of fonts, point sizes, and type weights. This gives the best results, in many cases, but relies on pre-knowledge of the fonts being used, so is relatively inflexible.
  - Intelligent Character Recognition (ICR)
    ICR is used for recognizing poorer quality machine print, or where the font is not predefined, and constrained handprint (the reading of handwritten block capitals and numbers). ICR also includes a capability to "learn" fonts and handwriting styles in one of two ways. Firstly, by pre-feeding a representative sample of the types of documents to be processed and learning from the way that an operator corrects recognition errors. Secondly, on the fly, progressively learning from the documents that pass through and from the manual correction they require.
  - Magnetic Ink Character Recognition (MICR)
    MICR is used mainly on checks—to read preprinted account, check number, and sort code details.
  - Non-character-based technologies that are emerging
    Includes cursive (or joined up) handwriting recognition—used today mainly for reading addresses on envelopes and the written amount on checks. As the technology develops, it will become possible to recognize handwriting in other contexts.

Key issues to consider when using recognition technologies are centered on recognition rates (how much can you recognize) and the setting of confidence levels (how precise is the data and metadata captured).

Optical Document Recognition

Optical Document Recognition (ODR), also known as Template-based Forms Processing, is used to process paper forms by recognizing and extracting data from them in order to feed it into a business application. It is particularly appropriate for survey processing, through government forms to mail-order purchasing and many others. Any application that involves some kind of structured form that has been completed by machine or by constrained handprint is a good candidate for a forms processing approach.

ODR works by having defined forms templates drawn up. These tell the system how to identify a particular form, which fields need to be recognized, what types of field they are, and what rules can be applied to maximize recognition results.

There are a wide range of techniques employed within this technology, including:

- Image clean-up and validation rules
- Sophisticated use of recognition engines
  Includes using multiple engines to select the "best fit" result and/or passing the image through the recognition engine multiple times at different image settings to optimize the results.

  Once the core data has been recognized, any documents that have failed are routed to reject repair operators, who check the forms and re-key any incorrect data. This is usually done from the image of the document, which is presented to the operator for this purpose.

Key issues that are critical to the success or otherwise of such systems are:

- Ease of template definition and/or forms design
- Quality of recognition
  Avoiding "false positives" tends to be more important than doing additional reject repair but that depends on the application.
- The management capabilities built into the system
  Includes routing captured forms through the process, monitoring the results, reporting on problems, and managing the scalability of the systems to handle multiple scanning points, recognition stations, reject repair routing, and finally, output of the appropriate data stream to the relevant application.

Intelligent Document Recognition

Intelligent Document Recognition (IDR) is an emerging technology, appearing first in Europe, that seeks to identify automatically the document type and its related attributes from the layout and structure of the document. The key benefit of this approach is a significant reduction in the human resources needed to index documents.

IDR is finding applications first in invoice processing and is being deployed for mailroom processing, insurance, and medical applications among others. It works on old image repositories, which can be "mined" either to export the data into knowledge and CRM-type applications or, more prosaically, to re-index the documents so they can be re-used for new purposes brought about as a result of change in the business, through merger, acquisitions, business refocusing, etc.

Why/Where Is Document and Content Capture Important?

Document and content capture is critical in paper or content intensive applications. These include such diverse areas as Customer Relationship Management (CRM), Records Management (RM), Knowledge Management (KM), transaction processing, and other line of business applications. Some examples are described below.

Customer Relationship, Records, and Knowledge Management

Customer relationship, records, and knowledge management are key growth areas for Document Capture. They are document and data-hungry applications areas. For example, OCR can be used to extract data from old paper documents for re-use. A major pharmaceutical company wished to use an existing drug for a new application and was able to re-examine thousands of reports from clinical trial reports carried out ten years

ago. It used OCR techniques to extract data that had not been captured from the documents during the trial to check out the occurrence of data indicating other side effects.

Transaction Processing

Transaction processing in the Document Capture context is defined as the use of capture and recognition techniques to extract relevant data from transaction documents (examples: order, invoices, delivery notes) for entry into the relevant business applications.

Today transaction processing lies at the heart of the ODR market (see above). Transaction processing is also a core future market for IDR, where it will both replace conventional forms processing (ODR) and provide the flexibility of document analysis essential for automatic classification and routing of transaction documents into the appropriate processes.

Drivers include:

- Savings in management time, clerical effort
- Paying bills on time to benefit from discounts
- Management control

The major issue, as raised above, is the increased proportion of transactions that can be undertaken, either digitally across the Web, or manually through interchange of conventional purchase order/delivery advice/invoice/credit note documentation. Hybrid capabilities (see *Key trends*, below) will become crucial in transaction processing, first in B2B, then for many years in the B2C and G2C environments.

An emerging application for IDR technology in particular is mailroom processing for capturing correspondence as it enters an organization. Systems are being installed today in remittance processing "lock-box" applications to capture documents, analyze their content and context, and route them into the appropriate business process(es).

How Document Capture Relates to ERP Applications

Applications such as SAP, PeopleSoft, Baan, JD Edwards, Oracle Financials, and others have explicit interfaces built into them for the integration of document capture and management. A good example is SAP's ArchiveLink interface, which allows captured documents, whether paper or digital, to be associated with particular transactions within the overall SAP R/3 ERP process.

Sector Trends

eGovernment Will Drive Hybrid/Hybrid to Electronic

The governments of North America and Europe are embracing the electronic access to information with fervor, to connect their citizens to government information and services. Some examples:

- In the United States, FirstGov.gov is a portal initiative, designed to deliver Federal and State government services to citizens.
- The UK Government is targeting to make all Government services available electronically by 2005.
- Germany is aiming for 50% of services to be online by 2005.

- Intellectual property/archive capture is a major driver in France, where a project to capture the *Etats Civils* of all citizens back to Napoleonic times is underway.

Hybrid processing (see below) will apply in this sector more than in any other, both in terms of volumes of applications and the potential lifespan of hybrid processing.

Financial, Retail Sectors are Open to CRM—As Is Government

The financial and retail sectors have traditionally been the major players in the document and content capture market. As management focus moves on from the back-office/archive led applications towards those concerned with the front-office, customer-facing environment, so the capture and presentation of customer documents offers a higher value proposition, in two main areas:

- Extracting customer information for profiling and personalization from customer correspondence, both from eForms and using IDR techniques.
- Presenting customers' own documents—correspondence, scanned application forms, claim forms, etc.—alongside internally generated documents in a customer file, either to internal staff in call centers, or to the customers themselves, over the Internet.

Best Practice

If You Are Thinking "Archive," You Are Behind the Game

Strategy Partners research shows that more paper is scanned for use by a business process than for archiving alone. Between 1998 and 2000, the proportion of applications installed for archiving and repository purposes fell from 62% of the overall market to just 40%.

With the development of eBusiness, the developing value proposition for document and content capture lies in the way that it allows organizations to process paper data almost as fast as that entered directly onto a Web page, with high integrity and accuracy in a way that complements the business process being served. As an example, a letter takes time to arrive in an organization, but once it arrives in the organization it is becoming a "must-do" to process it as quickly as, for example, a call-center enquiry.

Implement ODR Today—But Plan for IDR Tomorrow

"Conventional" forms processing (ODR) actually works. It is deployed in tens of thousands of installations in applications ranging from census projects through healthcare claims processing to market research and government forms, among others. As performance continues to improve and as hardware prices in particular continue to fall, so the range of potential applications continues to grow.

As semi-structured document processing (IDR) develops, however, the range of applications that can benefit from automated data extraction from documents will grow incrementally. Today, it is achieving success in the invoice processing market. Tomorrow, new applications in healthcare, CRM, KM, and mailroom and email processing will come to dominate the market for document capture.

Think Tactical Business Advantage, Not Long-Term Strategic Vision

As document and content capture has matured and become safe to buy, so the nature of the decision-making process for its implementation has changed. Systems are now justified on straightforward business metrics

of cost-saving, improved performance, and better customer service. In many sectors, such as insurance claims processing and regulatory compliance, it has become difficult to implement new solutions without using document and content capture techniques and technologies.

Return on Investment/Total Cost of Ownership

The Value of Documents

Any evaluation of the return on investment for document capture systems revolves around the value of documents and the processes they support. Some documents have little intrinsic value in themselves, e.g., a memo on a scrap of paper, but their content can be extremely valuable if it records a decision, event, or transaction. Other documents are structured and perform specific tasks, e.g., checks, which contain numeric and free form text. Others, such as photographs and representations of fine art, engineering drawings, etc., have value embedded in non-character-based content, which is easily assimilated by eye but rarely by machines.

As document capture handles all types of documents, the value can be categorized in three key ways:

- Quantifiable cost savings
  Such as saving costs of city-center office space used for paper filing cabinets, compared to the cost of storing an optical disk.
- Indirect savings by increasing the speed of the business cycle
  Or "business velocity," e.g., giving customer service staff access to customer correspondence on call center screens so that requests can be handled in minutes, while the customer is on the telephone.
- Business survival, in the areas of compliance and customer service in particular
  For example, providing timely evidence concerning safety certificates, tickets used, financial assets transferred, and other mission critical documentation can make the difference between continuing in business, being shut down by the regulator, or, as some are finding out, spending time in the penitentiary.

What Is Changing Now and Over the Next 12-24 Months?

The key developments in Forms Processing and IDR are documented above. Other trends include:

Hybrid Processing

As the importance of eBusiness grows, a greater emphasis is actually placed on getting paper-based interactions into the eBusiness world. Now, eBusiness demands that interactions be processed electronically and it is becoming a strategic "must-have" to be able to process paper in the same way as electronic interactions.

Automatic Classification by Content and Context

Automatic classification is a major driver of document capture, enabling indexing-free scanning and consequent massive cost reductions.

Automatic classification systems go beyond just analyzing content to take into account:

- Context mapping
- Subject, thesauri, etc.
- Implicit and explicit profiling (personalization)
- Relationship mapping—between documents, documents and processes, documents and roles, and documents and people

How Do You Buy It?

Software vs. Solutions vs. Process Outsourcing

The value of document capture lies in the process and repository applications that are populated with the captured information. This means that partners and solutions channels take on a particular importance. Strategy Partners' research indicates that suppliers for document capture split into five major types:

Software Vendors

Many document and content capture vendors sell directly to end-users, as well as through a combination of delivery channels (see below). When selling direct, they offer professional services and take responsibility for the final system, including associated hardware, software, and support. Users go to vendors directly with complex requirements that are essentially capture-centric—forms processing for censuses is a classic example.

Generalist Integrators

Generalist integrators focus on very large project deployments—IBM Global Services, Unisys, ICL/Fujitsu, the Big 5, Siemens Business Services, and others. In general, they lead and prime major government and/or international contracts, and tend to bring in specialists to carry out the capture and recognition parts.

Specialist Integrators/Solutions Providers

Specialist integrators operate in specific applications, vertical market areas, or geographic regions. They seek to deliver whole solutions into their core market. Such solutions tend to have document capture as an important, but not overriding component within the overall solutions offered. Examples include accounts payable applications, or insurance claims processing.

Value Added Resellers

Value-added resellers operate in environments where the user is in control of the overall application and is seeking to add document and/or content capture functionality to its facility. Examples might include inventory control, where goods received notes need to be captured and processed, or Human Resources, where Curriculum Vitaes need to be captured and stored for reference.

Specialist Outsourcers and Application Service Providers

There are three main levels of outsourcers and application service providers:

- Document Capture Bureaus
  These offer functional services, such as scanning, indexing, recognition, and other function-based services. Their output is a document and/or data stream, often on CD and mostly in the archive market. They sell on price and on speed/reliability.
- Managed Services Providers (MSP)
  MSPs set out to deliver documents, data, and process into some part(s) of a business process. As an example, they may take incoming mail, process it, validate it, return anything that needs to go back, handle some customer interactions, and feed the relevant information into their client's business process(es).

  Other examples include document hosting and electronic bill presentment services. It works best where there is an opportunity to implement an annuity or "pay per click" pricing model. Their business model is to find replicable services solutions that drive down the cost of service through economies of scale.
- Business Process Outsourcers (BPO)
  The key difference between BPO providers and MSPs is defined by business outcome vs. technical or functional output. For example: sending out an invoice is a functional output. Handling the accounts receivable process on behalf of a client is a business outcome. The systems approach is largely the same: the difference lies in the value of the outcome to the client, and to the level of understanding the BPO provider has of its clients' business.

  Emerging areas for this level of service are to be found in insurance claims management—with the provider being compensated on reduced costs of processing, rather than numbers of claims processed—and mortgage processing, where the business metric might be based around cost of processing, speed of offer, and/or levels of default.

How to Plan for the Future

Some key guidelines for future planning include:

- In the Forms Processing area, implement ODR, but expect to plan for IDR.
- Don't think archive. Instead, think supporting business process; think value of the document and/or its content; think business survival; think driving business velocity.
- The growth of eBusiness makes hybrid processing a key requirement through to 2004.
- Document capture will be a major factor in the growth of self-service CRM. It is key to capturing customer input in multiple forms to be made available in customer self-service applications. CRM vendors and solutions providers who do not provide such capabilities will find themselves at a competitive disadvantage by 2004.

Summary

Document and content capture is a critical set of technologies and disciplines that:

- Brings your eBusiness and conventional business processes into alignment
- Provides the catalyst for improved customer service and exploitation of knowledge bases
- Delivers explicit measurable bottom line benefits in a wide variety of business cases

Document and content capture is NOT just about document archiving. Today, the technologies and the business environment have reached the point where document and content capture can play a real role in front line mission-critical business processes.

Precisely how users source implementation and fulfillment services depends on their position and organizational culture. The kind of solutions provider depends on the level of the application being addressed; the level of outsourced service depends on whether the requirement is for a document, a reliable service, or a business outcome.

Glossary of Terms

B2B/B2C/G2C
Business to Business / Business to Consumer / Government to Citizen
Segments of Internet-inspired markets.

Backfile conversion
The processes involved in converting paper or electronic documents into a form suitable for incorporation into an Electronic Document Management (EDM) system.

Bar codes
A standard way to mark items for machine-readable identification. Used on documents for automated indexing.

Batch management
Organizing documents into batches, for scanning.

Bitonal scanning
Produces black and white images. Good for character recognition but involves loss of shading and images.

Browser
A program that allows you to receive HTML stream and thus access the Web.

Confidence levels (in recognition)
Most recognition technologies allow the user to specify the "confidence level" at which the system will accept the output from its recognition process. For example, a 90% confidence level means that the system will accept as accurate any recognition that it has judged to have a 90% or greater probability of being correct. Setting the confidence level requirement very high leads to great accuracy of reading, but causes the system to produce a high number of rejects, which then require manual re-keying. Setting it at a lower level reduces operator intervention, but increases the risk of "false positives."

Constrained handprint recognition
The ability to recognize handprint, usually in upper case, where the letters have been written individually, usually into boxes on a form.

CRM
Customer Relationship Management
The processes by which an organization attracts and retains prospective customers, leveraging an initial transaction via knowledge of their requirements into a long-term, ongoing transactional relationship to the financial good of the organization.

Cursive handwriting recognition
The ability to recognize joined-up handwriting on documents—very advanced technology.

DMS
Document management system
Term often used to refer to a document repository system (see below). In this report we are using the more precise term of document repository.

Document repository
Software systems to manage sets of electronic documents with specific functionality to control the check-in and check-out of material from the repository, provide look-up against defined attributes, and control versions of the documents.

EBP(P)
Electronic Bill Presentment (and Payment)
The processes around delivering invoice information to customers in digital format, and providing facilities for electronic payment of the same.

EDM
Electronic Document Management
The set of technologies for electronically managing documents—incorporating document and content capture, workflow, document repositories, COLD/ERM and output systems, and information retrieval systems.

False positives
False positives are when the system passes a character as correct when it is not. This can be critical if not detected, especially when the system is reading numeric data.

Fax Group IV compression
Fax-based compression algorithm. Produces small file sizes for bitonal images, but loses compression power drastically with color or grayscale originals.

GIF
Graphic Interchange Format
A graphics standard from Unisys and CompuServe, optimized for displaying graphical items through a Web browser.

Grayscale scanning
Captures a document in shades of gray. Good for printing and viewing, but less good for character recognition.

HTTP
HyperText Transfer Protocol
The core protocol used to enable Web browsers to communicate with Web servers and to find and extract information, often in HyperText Markup Language (HTML) format.

IDR
Intelligent Document Recognition
Process which uses advanced document analysis to understand a document's content and purpose.

Image processing
This includes such facilities as de-speckling of "dirty" images, skew detection and correction, and automatic image optimization, according to the needs of the application. Some applications benefit from using two different renditions of the same image—one optimized for recognition purposes, which is then discarded, and one placed in the repository for subsequent referral or reuse.

Imaging
electronic imaging, document imaging
An electronic imaging system is defined in this guide as a system that creates, stores, retrieves, and manipulates electronic images. It may include scanning and OCR functions.

JBIG
Joint Binary Imaging Group
Image compression standard. More effective in reducing bitonal image size than Fax Group IV compression.

JPEG
Joint Photographic Expert Group
Color image compression standard with the ability to set the level of compression desired. The higher the compression level, the lower the quality.

Metadata
Data associated with a document that is used to index and/or identify it in the context of a business process or repository application.

MICR
Magnetic Ink Character Recognition
The function of reading and recognizing the magnetic ink line, typically on checks or utility bills.

Multitonal
Documents containing content in color or in shades of gray.

OCR/ICR
Optical Character Recognition / Intelligent Character Recognition
The processes that interpret and convert images of text to produce computer-readable text for word processing, indexing, feeding into business processes, etc.

ODR
Optical Document Recognition
Template, fixed format forms processing.

OMR/MSR
Optical Mark Reading / Mark Sense Reading
The capability of a document capture sub-system to detect the presence or absence of marks in defined areas on a scanned document. This is used for processing questionnaires, exam papers, etc.

Patch Code Recognition
These are printed on document separators when preparing document batches for scanning documents. They are read and interpreted by the scanner.

Scanner management
This includes jam detection and double-feed detection. Some production scanners can also include facilities for document encoding (printing a numeric code or other unique mark on the back of the paper documents scanned) to provide audit trails.

Scanning resolution
Expressed in dots per inch (DPI). The higher the DPI, the higher the quality and larger the file size.

TIFF
Tagged Image File Format
Image document format which supports the full range of compression algorithms available, applying them on a per document basis.

XML
eXtensible Mark-up Language
An established standard, based on the Standard Generalized Mark-up Language (SGML), and designed to facilitate document construction from standard data items. Now being used as a generic data exchange mechanism.

Validation tools
These include look-up tables, the use of spell-checkers, adding up rows of numbers to ensure that they match the apparent total, date format checking, and other rules-based operations used for validating recognition results to detect and avoid false positives.