Digital capture is a key enabling technology in a business world striving to balance the shifting advantages and requirements of paper and digital documentation. This paper takes a closer look at the technology behind data capture, explaining some of the sub-technologies involved and highlighting the key challenges presented by data capture in the enterprise.
Table of Contents

Data Capture Technology ............................................. 3
Methods of Capture ..................................................... 3
Document Analysis ....................................................... 5
Validation ................................................................. 7
Data Capture in the Organization .................................. 8
The Right Technology at the Right Time ....................... 9
The re-emergence of data capture technology

Data Capture Technology

Technology has always held out the promise of “more and better” to businesses. From the early advances of the machine and industrial ages to the rapidly expanding capabilities of the digital/Internet age, technology remains the primary engine driving operational improvement and competitive advantage. Today’s digital solutions—from social media to cloud computing to the mobile workforce—are maintaining that continuity, providing business organizations with their greatest opportunity yet to align technological capabilities with business goals and customer satisfaction.

Digital capture technology is being seen as a key emerging, or better yet, re-emerging technology. Indeed, it is the perfect synergistic complement to a business world striving to balance the shifting advantages and requirements of paper and digital documentation. Despite this, many organizations, business leaders and even IT departments still don’t fully understand these solutions and what they really do. This paper takes a closer look at the technology behind data capture, explaining some of the sub-technologies involved and highlighting the key challenges presented by data capture in the enterprise.

Methods of Capture

Essentially, a capture solution connects a number of input channels and delivers them as documents and data to various destinations such as SAP, Microsoft SharePoint and OpenText Content Suite.

Input channels include:

- High-speed scanners, generally used as a centralized solution.
- Decentralized multifunctional devices and small desktop scanners that deliver documents to central network folders via email.
- Faxed documents (which are effectively scanned at the point from which they’re sent); though they are low in quality, this remains an established practice because of the additional authentication it provides.
- Photos (Smartphone)—most business professionals carry a smartphone with a camera; while this provides a convenient portable scanning device, readability is often poor, images are blurred due to shaking and not well framed, and image size information is not available. In response to these issues, document scanning apps have been developed and are constantly being improved.
- Email, with relevant information either in attachments or body text.
- Document up/download to FTP servers, which allow for fast and direct document exchange.

KEY TERMINOLOGY

- **Capture:** The process of acquiring documents for an ECM (enterprise content management) system or a business process
- **Scanning:** The process of converting a paper document into an electronic document
- **OCR (Optical Character Recognition):** Intelligent technology that converts pixels into coded characters that can be read by computers
- **ICR (Intelligent Character Recognition):** Intelligent technology similar to OCR with the added capability of hand print recognition
- **IDR (Intelligent Document Recognition):** Intelligent technology for characterizing, classifying, and extracting specific data from documents (for example, a case number from a customer letter)
Early Archiving Versus Late Archiving

When documents are scanned as soon as they enter the organization—you achieve maximum solution benefits, such as eliminating the movement of physical paper and enabling parallel access to documents by multiple users. This “early archiving” is in contrast to “late archiving,” where business processes are initially paper–based and relevant documents are scanned and connected later.

Batch Versus Ad hoc Versus Barcode

Capturing documents can be done through batch scanning (many documents at high speed, typically in a central location), ad-hoc scanning (individual documents scanned by employees, ideally from within the ECM or business application, e.g., point-of-sale transactions) and barcode scanning (the document is scanned with a barcode and later connected to the right process or location).

![Diagram of barcode process](image)

Late archiving using barcodes - allowing a user to define which folder to scan a certain document into, and another user to do the actual scanning

Scanned Documents Versus Electronic Documents (Text Versus Image PDF)

There is a big difference between a scanned document and an electronically-generated document. A digitally born document contains all sorts of elements—text, graphics, perhaps pictures—but a scanned page is just pixels. After scanning, however, OCR enriches images with extracted text, making them similar to digitally-generated documents. Neither, however, makes the data needed by business applications readily available. Whether an invoice is received as a fax or as a word file, determining invoice numbers, vendor identity, purchase order numbers or total amounts still requires document analysis.
Document Analysis

What started as OCR in the 1970s is now often called document analysis, a much better term given the level of “understanding” data extraction requires. It is indicative of the constant improvement and fine-tuning that recognition technology continues to undergo.

Progress is measured by how often a piece of information can be correctly identified on a given document. Pushing these numbers up from 50% to 70% to 80% and above 90% is the aim of development teams. Though the task is exponentially more difficult as the percentage rises, benefits can be significant. For example, an increase in the recognition rate from 95% to 96% appears to be an increase of only 1%, but it actually decreases the manual work left to be done by 20%.

Beyond that, major breakthroughs—such as establishing hand print recognition or free forms recognition—have expanded automation capabilities, improved business efficiencies and created new market opportunities.

Typically document analysis consists of four major steps: page separation, document classification, data extraction and interpretation:

**Separation:** Splitting a batch of images into individual (multi-page) documents

**Classification:** Identification of document type

**Extraction:** Searching for basic information on document

**Interpretation:** Enhancing of extracted data with context information

In marketing collateral, you will often find claims of recognition rates over 99.5%. Without a description of the exact extraction tasks, how they were set up, the initial image quality, evaluation methods used and several other variables, such numbers are meaningless and quite possibly misleading.

**Page separation**

When paper is scanned, it results in a stream of pages. Separating these pages into documents is the first recognition task. The use of patch code sheets, blank pages, or barcodes on the first pages of a document makes for a very reliable process. Using the content of a page, such as “Page 1” is possible, but will typically require a project-specific set of rules. Of course, in many situations, such as processing email attachments, this page separation step is not needed.
Document Classification

Document classification is the process of finding out what type of document you have. Some technologies recognize prominent words or phrases similar to search engine functionality. Some look at the arrangement of layout elements on the page. Regardless of the technology, differentiating between rule-based and adaptive recognition is important.

Rules-based recognition means rules are programmed into the software while adaptive recognition means the software itself determines differences by analyzing samples fed in by operators. As the software collects more "training data," fewer documents need manual work. While adaptive recognition may seem more efficient, the best approach is to let the two processes work together since some document classes, such as forms, lend themselves more to rules-based recognition.

Data Extraction

Like document classification, data extraction may utilize different technologies—the main distinction being again between rule-based and adaptive, self-optimizing methods—and three types of documents are involved: forms, semi-structured documents and unstructured documents.

Structured Documents (such as forms)

- Data at fixed positions

Semi-Structured Documents (typically B2B correspondence)

- Data is in logical groups but positions are unknown.

Unstructured Documents (typically C2B correspondence)

- Data may be anywhere in the document.

Forms are designed for data capture and their data is present at pre-defined locations. While this seems straightforward, layout variances due to slightly different printing processes, visual distortions, and variations in hand printing can make forms processing a challenge.

Semi-structured documents like invoices or purchase orders contain information in different locations but retain a logic that can help the software locate it. For example, a “total” amount on an invoice can be identified by certain rules and guidelines that, combined in an algorithm, function with high reliability. These rules may include:

- It’s typically the largest amount
- It’s typically at the bottom of a page
- It has keywords like “total” in its vicinity
- It has a given mathematical relation to other amounts like tax
- It is an amount, meaning a string of digits, currency symbol, and punctuation in a certain sequence

Unstructured documents usually appear in a consumer-to-business situation. The extraction task is generally not as elaborate, with specialized algorithms using database information to search for things like customer name, contract number, insurance number, or letter date.

Interpretation

The final step is conversion and data verification. A date may be converted to the requested format or a vendor name may be converted into a master data record ID. Only correctly formatted data will be accepted by the backend system.
Validation

OCR is not 100% accurate. Whatever the task is, some of the required data needs to be provided in a follow-up step, called validation, correction, data completion, or something similar. This typically involves dedicated software to make the process as smooth and efficient as possible:

- Document images are presented side by side with the data entry form.
- If the OCR program knows the location of the required information, that location is highlighted on the document and the corresponding snippet is shown in an enlarged form close to the data entry element.
- The user is guided from field to field, bypassing those that don’t need attention.
- If the location is not known, the software should be able to highlight all possible candidates on the document.
- The user may be allowed to click on information in an image or grab a section of a document to have it automatically entered into a form, as with the patented OpenText “Single Click Entry” technology.

HUMANS VERSUS MACHINES

OCR is not 100% accurate, but human “data capture” is also not 100% accurate; for example, when you type content into a Word document, you inevitably produce a certain number of errors. Error rates between people and OCR are comparable. It’s true that a certain percentage of errors are inherent in OCR; on the other hand, machines never lose concentration.

When a field requires a date, all occurrences of dates in the document are highlighted to help the user find the correct one.
Data Capture in the Organization

How data capture technology is practically embedded in a given organization’s IT infrastructure depends on the business and which business processes are supported by a capture solution. This can create a number of challenges and considerations.

Enterprise Applications Versus Line-Of-Business Applications

Organizations often implement a capture solution at the enterprise level, so that most business processes are triggered by or affected by incoming documents. Some organizations, however, use capture solutions only for very specific business processes, for example invoice processing in the accounts payable department. Over time, however, the benefits capture solutions deliver often lead organizations to extend their application to other functions, which also extends the solution’s ROI.

Centralized Versus De-centralized Operations

In a centralized organization, scanning and manual verification happen in one physical location and are run by a specialized department or team using one client/server installation. In a decentralized organization, each department will run its own capture operation. The benefit of a centralized operation is of course higher efficiency, driven by better equipment and more highly trained personnel. On the other hand, verification of complex documents may require intimate knowledge of related business processes or may even be enforced by confidentiality concerns, which demands a decentralized approach. For example, the capture solution for HR-related processes will often be operated in the HR department itself.

Data Capture Versus Business Applications

A capture solution does not exist in isolation. It is defined, rather, by how it integrates with business applications. Take, for example, an invoice processing solution with the following steps:

Step 1: Scanning an invoice
Step 2: Data extraction with OCR and IDR
Step 3: Manual verification and data completion
Step 4: Handling price, volume, and other discrepancies
Step 5: Approval
Step 6: Posting
Step 7: Payment

These seven steps may be distributed over three systems: the capture system, a workflow system, and an enterprise resource planning (ERP) solution. While steps 1 and 2 are clearly the domain of the capture system and 6 and 7 the ERP system, determining where the other steps are executed depends on the nature of the document, with the main goal being not to waste efficiencies duplicating the business logic in the capture system. Clearly, solution design choices must be made.
The Right Technology at the Right Time

Change can be intimidating, but for companies who embrace it, change drives opportunity. When it comes to digital change, many new strides are being made in capture technology. Innovative solutions are helping organizations analyze documents in ways that maximize usability across the enterprise while integrating all paper and digital information into a range of enterprise applications and workflows. The advantages are clear, but understanding the technology more thoroughly—something we hope this paper has facilitated—is the first step to leveraging it more effectively.

**OpenText Capture Center** uses advanced OCR, ICR, IDR, adaptive reading, and other technologies to turn documents into machine-readable information. It then captures, interprets and automatically integrates the data content in scanned images and faxes to significantly reduce manual keying and paper handling, accelerate business processing, improve data quality, limit compliance risk and reduce costs. Far from a scan-and-convert point solution, OpenText Capture Center is a complete, end-to-end business process that captures documents and data to archives, workflows and processes as needed.

Document capture technologies, however, can be challenging to implement and customize. Given the diversity of business platforms, capture needs, and IT infrastructures, out-of-the-box implementation is an unlikely scenario for most organizations. Depending on your needs, there are a number of tradeoffs between functionality, performance, and efficiency that need to be carefully considered and that may require third-party consultation to effectively balance.

To learn more about how OpenText capture solutions can help your organization improve efficiencies, compliance, information quality, and business processes, please contact:

Advisors@opentext.com

www.opentext.com