OpenText Capture Center

Classifying and extracting data from documents: OCR, ICR, and IDR

The content of scanned documents (or faxes) is readable only by humans. For computer applications, they are merely a collection of meaningless pixels. In order to facilitate the automatic storage of documents within a repository and provide correct attributes for those documents, the type of document and relevant metadata have to be known. In order to automatically trigger business processes and populate business transactions with information, relevant data from the documents must be available.

OpenText Capture Center (OCC) extracts information from bitmap documents by using the most advanced Optical Character Recognition (OCR), Intelligent Character Recognition (ICR), and Intelligent Document Recognition (IDR).

Using OCC, your organization saves money from reduced manual keying and paper handling, speeds up business processes by using digital workflow right from the start, improves data quality by capturing all relevant data from documents, and reduces compliance risks by keeping track of document-related activities.

Using OCC

OCC is used in many areas, including the following:

Digital mailroom

The digital mailroom automatically captures and classifies all information entering an organization and then routes it to the appropriate person, department, or backend Enterprise Resource Planning, Customer Relationship Management, Enterprise Content Management (ECM), or workflow solution. It also provides tracking and auditing of that correspondence.

Traditional mailroom processes are slow and inefficient, dominated by paper documents. By moving to a digital mailroom, companies can reduce operational costs, streamline and accelerate business processes, and deliver improved customer service.

Transaction and process management

OCC captures images and processes data for certain business processes. The purpose of scanning is to input data into a business process. Ideally, all business data for a transaction should be extracted and validated. As a result, the process may be fully automated. The scanned document will be attributed by some metadata, stored in an archive, and associated with the business transaction.

WHAT IS ALL THIS: OCR, ICR, IDR?

A scanned document is just a bunch of pixels. Optical Character Recognition (OCR) yields the character code for each written character on the document.

Intelligent Character Recognition (ICR) extends OCR to contextual algorithms. ICR also allows for the recognition of hand-printed characters.

Intelligent Document Recognition (IDR) analyzes a document and finds specific information from an unknown document layout, like an order number from purchase order. IDR typically builds on top of OCR/ICR results.

BENEFITS

- *Reduce operating costs:* Automate manual tasks and deploy a single input management platform.
- *Improve information quality:* Classify, extract, and verify information and leverage a common set of business rules.
- *Accelerate business processes:* Reduce exception processing and enhance customer relationships.
- *Reduce compliance risks:* Control the flow of each incoming document and connect each document with its business transaction.

Typical documents that may be processed are order entries, application forms, insurance claims, or business reply mail. Invoice processing is also a dominant application for this business-use case, and OpenText Invoice Capture Center is a preconfigured application of OCC dedicated to this application. Automation with OCC reduces costs, increases productivity, increases efficiency of business processes, and improves information quality.

Scanning documents into electronic files

Documents have to be put into a repository or classified and routed to a centralized point as quickly as possible.

Typically, the process is batch oriented. The documents are scanned and indexed automatically, sometimes with keyentry to enter or correct indexes. Presorting is eliminated through automated classification. The process is managed centrally, and components of the process may remain local within the scanning system or distributed remotely.

Documents will then be available in the content management and workflow systems for fast access, archiving, and knowledge management. Apart from this, main costs associated with managing paper (manual handling, storage, retrieval, loss, compliance) will be reduced dramatically.

Backfile conversion

Backfile conversion replaces a paper archive with a digital archive.

This business-use case is very similar to batch capture, as the main task is the conversion of high-volume paper to electronic images with index and metadata. However, it is important to conserve the original file structure of the paper archive. This is done with various kinds of separators to reflect the hierarchical structure of the filing repository.

Conversion of the paper archive will leverage all the benefits related to a digital archive.

Ad hoc capturing

Ad hoc capturing is characterized by low-volume, on-demand document capture. For example, an office worker who wants to convert paper documents into usable electronic documents does ad hoc capturing.

The devices used are low-speed scanners or networked office Multi-Function Peripherals (MFPs). Systems are sometimes linked into shared repositories, or documents may be moved into centralized repositories, converted, and used as faxes, email attachments, etc.

With ubiquitous availability of network scanners and MFPs, ad hoc ensures that paper is eliminated as early as possible and documents can be shared for collaborative processes.

Example applications

Customer Relationship Management: In business-to-consumer businesses, it is critical to maintain customer records accurately and completely. Customers often send paper documents (e.g., application forms, claim forms, complaint letters, salary certificates, etc.) that must be filed in the right customer record. Useful metadata are customer name, customer number, document type, case identification, contract number, and so on. Typical organizations with a high volume of incoming paper documents are utility service companies, city councils, assurance companies, banks, mail order businesses, and service companies.

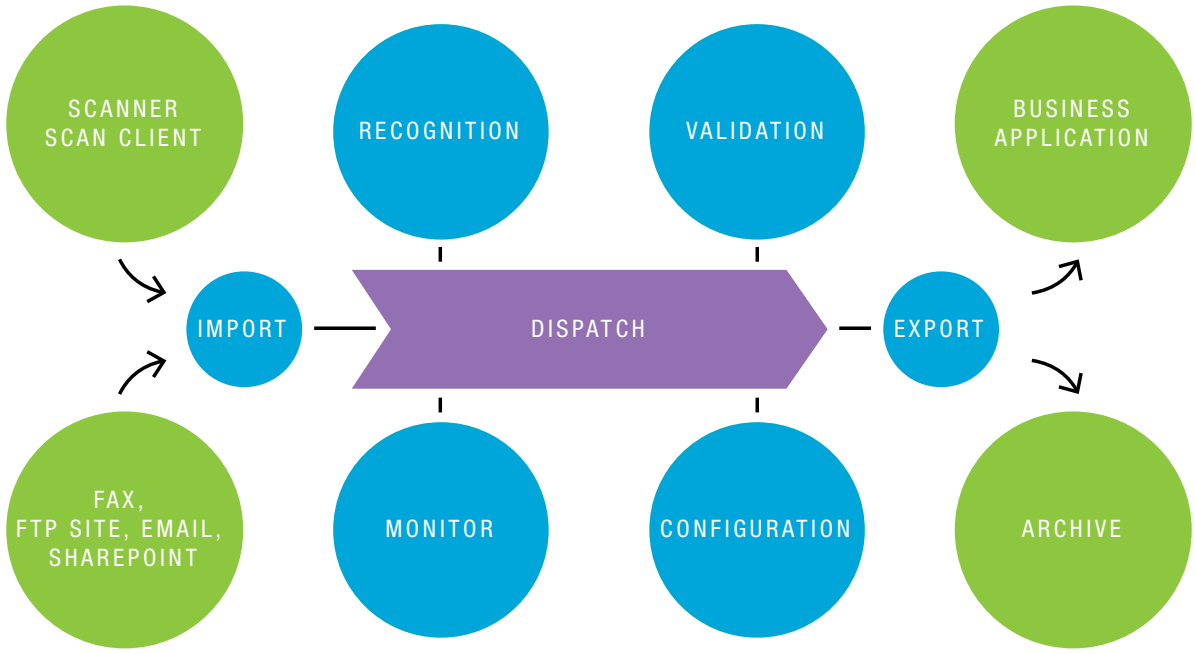
Human Resources (HR): Using electronic personnel files is the most efficient way to reduce chaos in the HR department. This requires an initial conversion of the existing files. From then on, all incoming documents will be stored in electronic format. The management of HR documentation can be very complex, mainly due to the large number of different document types that can exist for each employee. Documents often exist in several formats and locations with different processes being carried out for each document type.

Using OpenText Capture Center, your organization saves money, speeds up business processes, improves data quality, and reduces compliance risks.

Insurance claims: Insurance customers report claims either by using forms or by submitting free formatted letters. The customer needs to be identified in the database along with claim details like injured person, date of incident, or sum of damages. Prepopulating the claim processing mask lets the insurance specialist focus on the subject itself instead of capturing data.

Travel expenses: Expense processing is a time-consuming, manual task with data entry, paper storage, and distribution problems. A capture solution streamlines and automates this process; expense claims are routed automatically to the appropriate individuals for approval, and the resulting expense data can be automatically posted for payment once approved.

eDiscovery: Electronic discovery refers to any process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case. Scanned documents and faxes must provide all of their textual contents in order to be found in case. With OCC, thousands of file types can be classified as records.



Functional description

Extracting data from scanned documents is a multi-step process:

Document acquisition: OCC is closely connected to OpenText Imaging Enterprise Scan. Using the OCC scan profile, documents are automatically transferred to the recognition process when scanning is finished. Documents can also be picked up from a variety of other sources: file system folders, File Transfer Protocol (FTP) sites, email servers, or Microsoft® SharePoint® servers.

Document recognition: Document recognition is a two-step process. First, a document is classified; that is, its document type is determined—whether it is a purchase order, a contract, or an application form. In the second step, a defined extraction profile will be used for each document class. This allows the user to extract the relevant business data from any specific type of document, e.g., a Purchase Order (PO) number for a PO or a contract number for a contract. If only one type of document is processed, the classification step is omitted.

Document validation: OCR, ICR, and IDR do not always extract all required data. Due to dirt, document damages, irregular fonts, or unusual document layout, some data will not be identified with a sufficient level of confidence. For these cases, manual data entry is supported by a powerful data entry client that is designed according to the highest ergonomic standards. Keyboard usage for advanced data keying personnel is supported, as well as mouse-based data capture using OpenText Desktop Capture.

Document release: Release modules for OpenText Content Server and Microsoft SharePoint automatically transfer the document into the required folder or library and fill in metadata. For other systems, the document image and data is stored in the file system and can be picked up from there. A programming interface allows for the development of custom release modules.For managing the following steps, OCC supplies modules for customizing and administration.

Configuration: OCC classifies each document, i.e., determines its document classes. A document class defines the set of fields (also known as metadata or index fields) and how OCC is expected to locate and extract these. Classification determines the document class without manual presorting. All of this customizing occurs with an intuitive user interface. For all basic extraction tasks, design and test tools allow for recognition rate control, verification, and optimization.

Application Programming Interface (API): Using the API via programming (.NET™) or scripting (JavaScript), advanced adaptation to project-specific requirements are possible. Customizing code can be injected at almost any step of the recognition process.

Production monitoring: To control the production process, the administrator can look into the current state of each of the batches that are in the system. By selecting a certain subset, personnel can easily spot production problems like missing resources or failures of connected components. Statistical reports help to allocate resources or to distribute costs in a shared service environment.

Technical administration: In case one of the modules runs into trouble, (e.g., a release module cannot connect to the target system) the administrator uses the technical administration tools to identify and fix the issue.

Automated recognition: the heart of OCC

The tasks that can be automated are document separation, document classification, and data extraction. For most of these tasks, OCC offers several automation methods that can be configured according to the nature of the document. Some types of documents include the following:

Structured documents

Typically these are forms with fixed locations for each piece of data.

Semi-structured documents

Typical business documents are semi-structured. POs or delivery notes follow some general layout patterns so that rules can be defined concerning where to look for certain pieces of information. However, unlike forms, there is no defined geometric region for each piece of data.

Unstructured documents

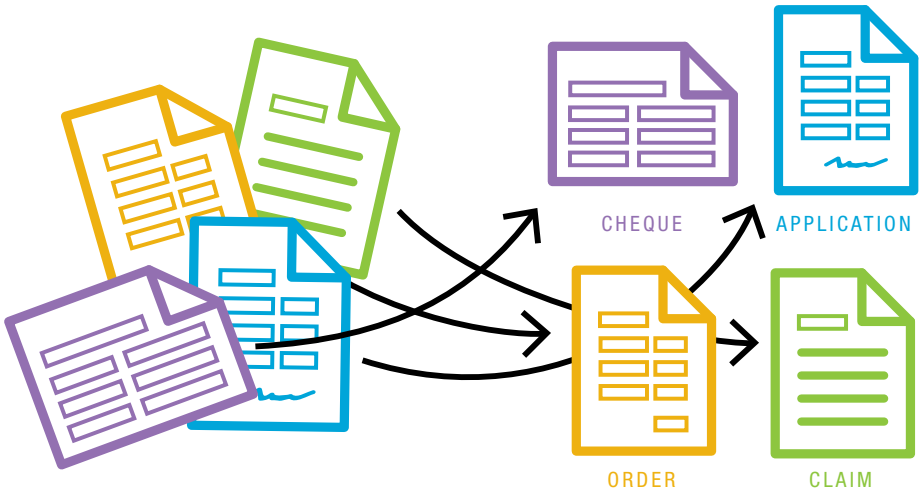
In business-to-consumer environments, correspondence follows no regular pattern. This is a typical example for a case of unstructured documents. Only the syntax of the information and semantic pattern can guide the search for information.

Document separation

OCC can assemble a batch of joined images into documents. The cutting points in the separation process are either defined by the content of extracted data fields, e.g., barcode or patch code, or by a defined number of pages.

Document classification

The document class is an attribute of the document that is typically used to determine the relevant business process or the folder into which it should be stored. Within OCC, the document class is used to control which kind of metadata have to form the attributes of a specific type of document. Each document class may have a different set of metadata.



OCC offers several options to determine the document class. These options can be combined.

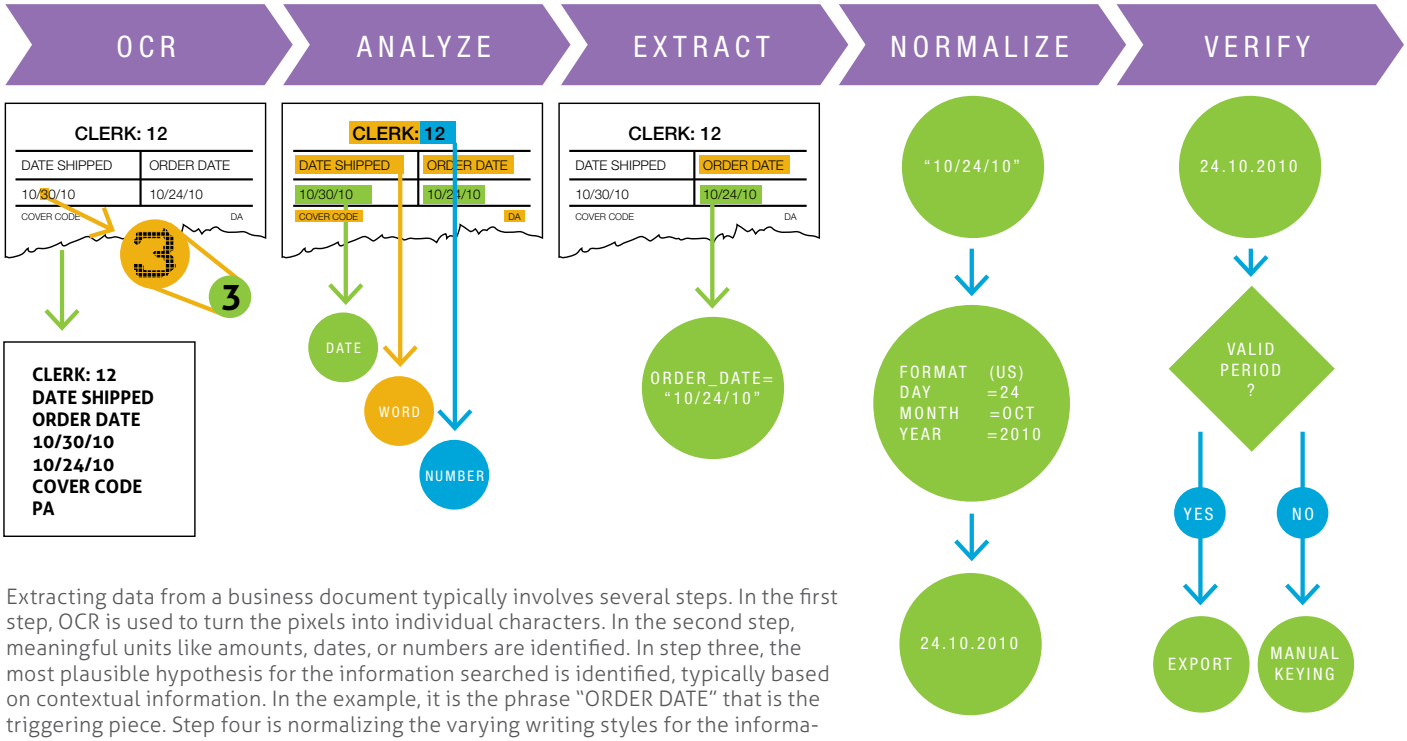
- Adaptive Classification Technology (ACT) is a learning algorithm. ACT uses several samples from each document class, e.g., several orders, claims, or applications, and extracts the characteristic features of these documents (like specific keywords or phrases) based on OCR content. Each document that is to be classified is compared against these features and classified accordingly. This method is well-suited for unstructured documents.
- Rule-based classification uses a man-made set of classification rules. These rules typically use graphical objects, phrases, and combinations of keywords together with their geometric relation. This approach is best used for semi-structured documents as well as forms.
- Forms often contain certain elements for identification, like a form ID number. These can be extracted and used for classification.
- Preset values are used when the scan operator scans just a single batch of documents from the same class. Imaging Enterprise Scan allows for this data to be imported at scan time.

Data extraction

Data extracted from documents is either used as metadata in a repository for structured storage and retrieval or to automate transaction processing in an enterprise application. The set of extraction methods is always the same for both usages. All of the following methods can be applied to a single document:

- Barcode, patch code
- Optical mark recognition
- Forms reading (fixed, anchored location, hand print, machine print)
- Free forms recognition (rule-based extraction)
- Adaptive Reading Technology (learning through validation operator)
- Database-driven recognition (match a record in a database with the document)

OpenText is known for its exceptionally advanced recognition technology all the way through the technology stack. More than 35 years of experience in the field with large-scale operations (like the US census or the German tax authorities) are the foundation of the product. Thousands of users as well as industry partners rely on OpenText's engines for OCR, ICR, and IDR. The power of these components is unleashed by OCC to allow for the highest automation rate in document recognition.



Extracting data from a business document typically involves several steps. In the first step, OCR is used to turn the pixels into individual characters. In the second step, meaningful units like amounts, dates, or numbers are identified. In step three, the most plausible hypothesis for the information searched is identified, typically based on contextual information. In the example, it is the phrase "ORDER DATE" that is the triggering piece. Step four is normalizing the varying writing styles for the information, and the last step is the logical validation. Although depicted as a sequence, the steps really run in cycles, following many hypothesizes in parallel.

Licensing options

OCC is licensed by volume, i.e., number of processed pages per year. Three licenses are available depending on the number of fields for which automation is used: Unlimited, 1-5 and 0. The latter is mainly used for manual indexing and/or searchable PDF creation. Also available are optional modules and one-time licenses. The number of validation clients is not limited by either of the options. ■

www.opentext.com

NORTH AMERICA +800 499 6544 ■ UNITED STATES +1 847 267 9330 ■ GERMANY +49 89 4629
UNITED KINGDOM +44 0 1189 848 000 ■ AUSTRALIA +61 2 9026 3400