WHITE PAPER

Cognitive Document Automation

It's Not Just About OCR

KOFAX

Contents

The Evolution of RPA

Robotic process automation (RPA) is quickly moving from simply automating repetitive tasks and augmenting the work of employees to a richer set of capabilities that address broader business requirements. Yes, RPA interacts with websites, business and desktop applications, databases and people to execute repetitive work, but organizations are now viewing RPA as an essential component of a larger "intelligent automation" requirement to digitally transform business.

Intelligent automation involves the use of RPA, cognitive document automation (CDA) and business process orchestration to achieve the efficiency, compliance, revenue and customer goals of your organization. This white paper will assist the RPA-knowledgeable reader in understanding CDA concepts, including CDA capabilities and technology basics, and what aspects to consider to ensure a successful CDA deployment.

At its core, RPA solves problems associated with data-centric manual tasks. But document-based tasks are equally if not more time-consuming and costly. Organizations with document-intensive processes struggle to tame high volumes of paper and electronic documents, multiple disparate communication channels, manual repetitive processing steps, and hard-to-integrate backend systems. These challenges reduce information visibility and enterprise agility, swell operational costs, and slow customer response and resolution times. Businesses also struggle to keep up with government and industry regulations due to lack of control and visibility over each step of their business processes.

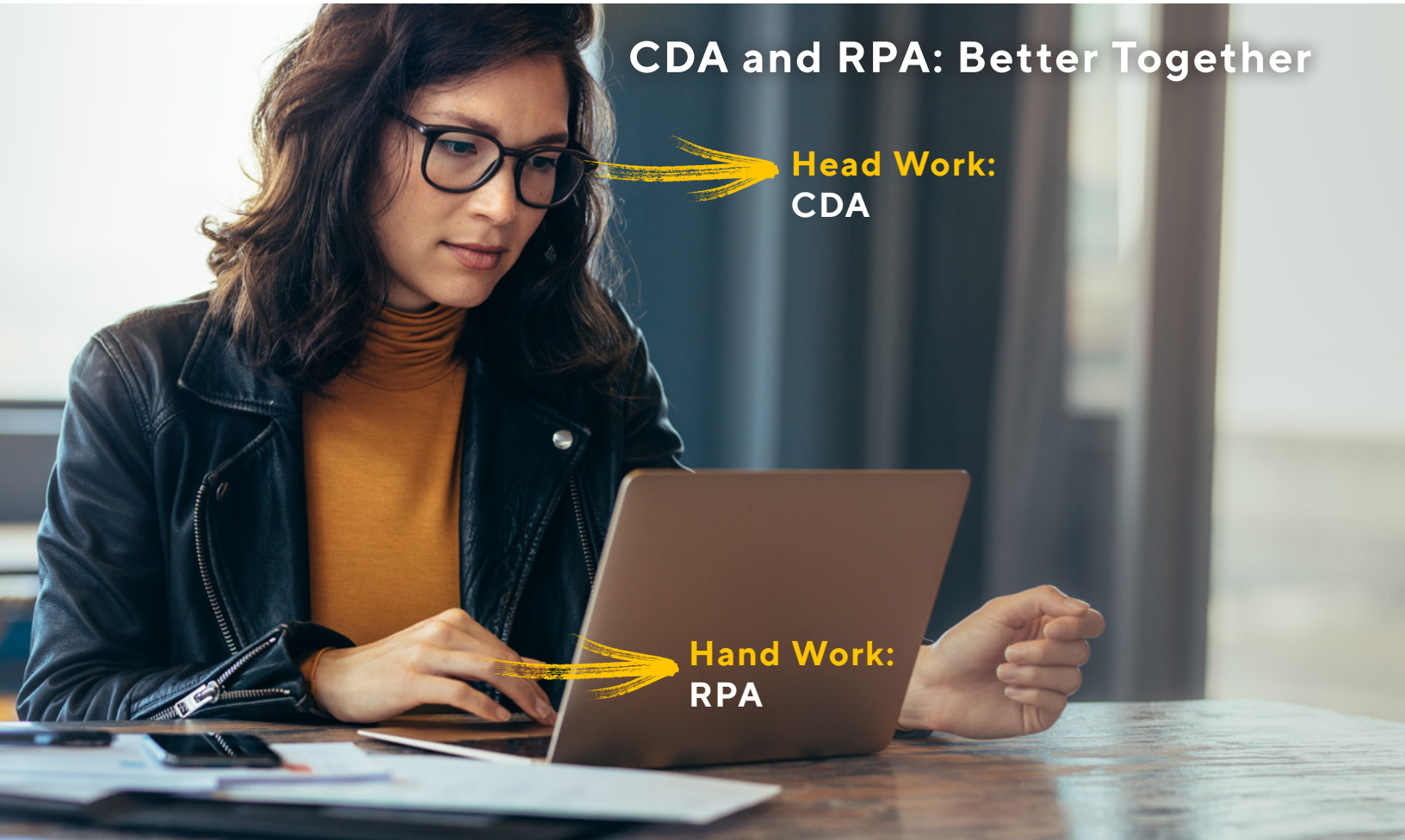
Organizations want to make these document-based processes more efficient by automating the acquisition, understanding and integration of the documents and information contained in them. Cognitive document automation software uses artificial intelligence (AI) to make this a reality.

RPA and CDA: Better together

RPA and CDA work well together by using AI and software robots to automate different kinds of manual processes:

- RPA automates repetitive manual tasks that interact with websites and applications, trigger responses and communicate with systems—this is the repetitive "hand work" of processing electronic data.
- CDA automates the processing of unstructured data contained in documents and emails—this is the intelligent "head work" of understanding what the document or email is about, what information it contains and what to do with it.

Done right, CDA improves information visibility, reduces document processing costs, increases productivity, accelerates processes, increases data quality for fewer errors, ensures compliance, and improves customer engagement and responsiveness. By the time you finish this paper, you will know better how to "do CDA right."

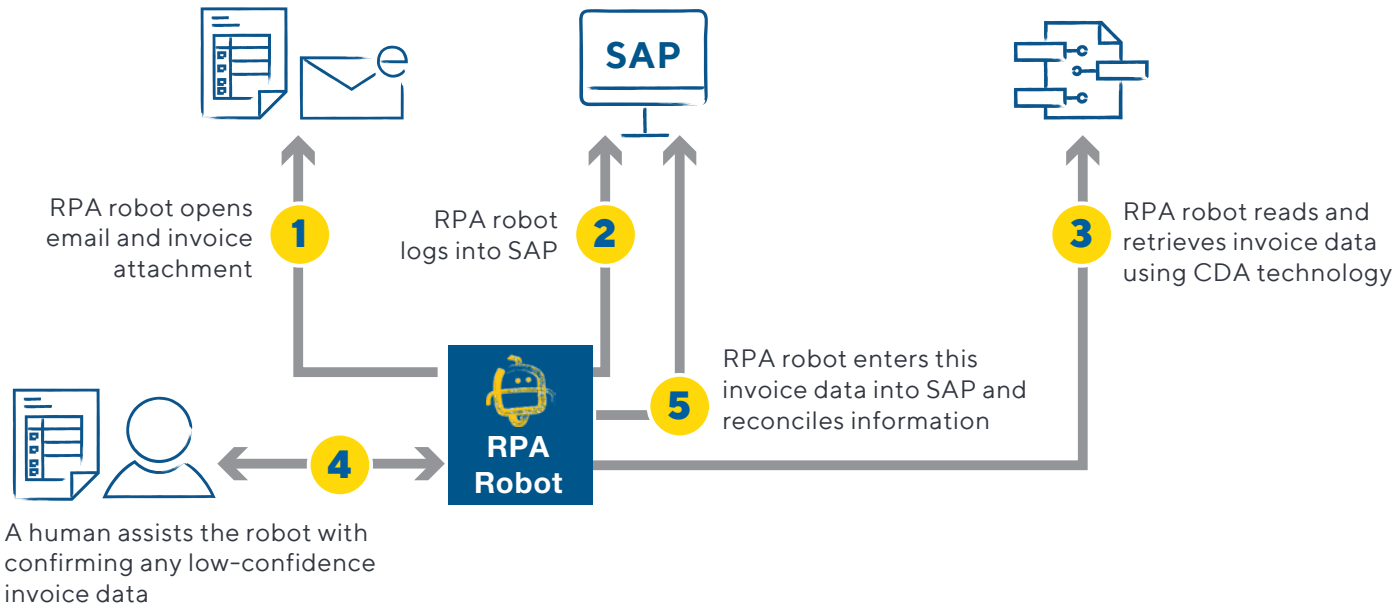


Ask yourself:

What documents are presenting challenges? Maybe it's new customer application forms, tax forms, insurance claim forms, mortgage packages, onboarding documents, supporting documents, invoices, sales quotes, sales orders, purchase orders, contracts or shipping documents. These are areas where both RPA and CDA can help.

Here is a simple step-by-step example of RPA and CDA working together to automate invoice processing:

1. RPA robot opens email and invoice attachment
2. RPA robot logs into ERP
3. RPA robot reads and retrieves invoice data using CDA technology
4. A human assists the robot with confirming any low-confidence invoice data
5. RPA robot enters this invoice data into ERP and reconciles information



"Digital platforms turn a complicated process into an experience that acquires documents and information from any channel, intelligently extracts relevant data, then integrates with a wide variety of back-end services to enrich that information—ultimately eliminating customer and company friction points."

- 451 Research

The Intelligence of CDA

RPA and CDA both encompass three process stages: Acquire, Understand and Integrate. RPA and CDA can acquire documents and electronic data from numerous sources; extract, aggregate and transform this data; and deliver the transformed data to the systems and processes that require it. Let's take a closer look at these three stages for CDA specifically.

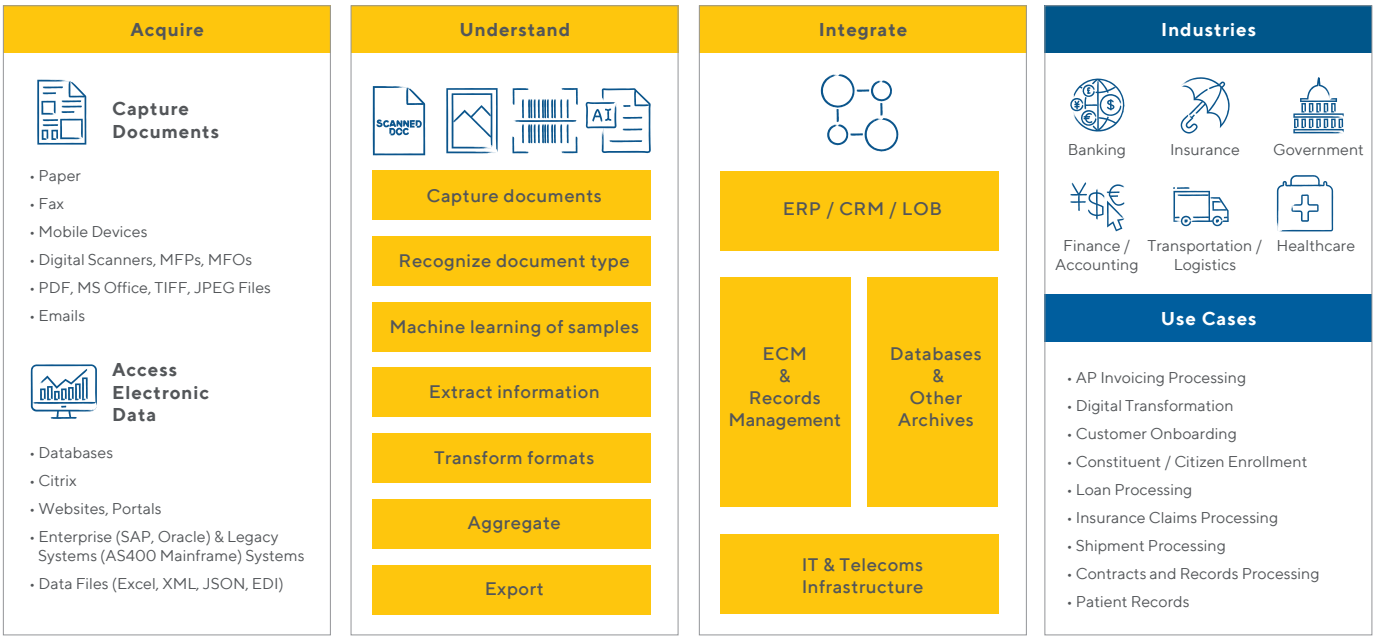
Acquire

This step involves acquiring documents from common channels, formats and devices including email, fax, folder, PDF and Office files, website uploads, MFPs, scanners and, especially, mobile devices. Customers expect the document submission process to be smart enough to allow use of different channels at different times during the same process—without having to adjust to different instructions, re-submit or start over. Mobile apps and capture-enabled websites should include embedded document capture capabilities that enable real-time capture and data display and the ability for users to correct data before submitting.

Understand

Once a document is acquired, CDA answers the following questions: What is this document or email about? What information does it contain? What should be done with the document and the information? This step involves the cognitive transforming of documents and document data into intelligent, business-consumable content required by downstream processes and systems (BPM, CRM, ECM, ERP, etc.). Typically, the steps required to understand a document are:

1. Document classification
2. Separation between documents
3. Data extraction
4. Data validation (human or automated)



Integrate

In the Integrate step, CDA integrates with downstream processes or systems of record through either pre-configured system-specific connectors or API or standards based connectors. CDA can also leverage RPA robots to integrate with systems where these connectors are unavailable. In this case, RPA employs built-in integration capabilities that easily map data between source and destination systems, without the need for exposed APIs or web services and without writing integration code.

Must-Have CDA Capabilities

Leading CDA solutions offer much more than just OCR (optical character recognition). When evaluating available solutions for CDA, keep these must-have capabilities in mind:

Distributed capture (field/branch offices)

Support for both centralized back-office document capture using production scanners, as well as distributed use cases where field and branch offices must perform the document capture, is crucial. To minimize TCO, ensure the CDA solution offers central administration, licensing, reporting and scanner profile management.

Multi-channel capture

Comprehensive multi-channel capture to accommodate all customer preferences is essential, including mobile, email, web, fax, desktop scanner, folder and MFP front-panel integration. Ensure the mobile SDK allows developers to integrate a full suite of mobile capture capabilities including image capture, compression, perfection, classification, recognition, extraction and data validation into their own websites and apps. The more these capabilities can run on-device versus on-server, the better the customer experience. Check for any mobile extraction solutions pre-configured for common document types (such as ID documents and bank checks).

Mailroom capture (multiple departments)

To avoid the wasted cost and effort of developing and maintaining disparate solutions from multiple vendors, the CDA solution must deliver benefits across multiple departments (and their associated documents). This includes support for the acquisition and understanding of any document type (forms, invoices, shipping documents, mortgage documents, onboarding documents, medical documents, emails, letters, contracts, etc.) across any department for a fully comprehensive digital mailroom.

Document classification

Human workers should not be needed to apply barcode stickers or insert cover sheets: automated document classification capabilities for any document type are key. Documents should be automatically classifiable without human intervention using multiple methods, including document layout, document content and regular expression-based rules, and these methods should "machine learn" via document samples (see machine learning later in this document).

Document separation

The CDA solution must provide automatic document separation capabilities within a batch (stack of documents) or document package (customer-specific) without human intervention to insert separator sheets. Document separation should "machine learn" based on samples given to it (manual configuration should not be required).

Data extraction

The CDA solution must support extraction of data from any document, in any language and in any format:

- **Structured forms:** The desired data on the document is known (e.g., account number) and is always in the same location on the page.
- **Semi-structured documents: T**he desired data on the document is known (e.g., invoice number), but its location could vary on the page, as is true for invoices, which are vendor-specific.
- **Unstructured documents:** The desired data on the document may not be known (e.g., is the account number even present in the letter or contract?), and its location could vary on the page and/or be buried within a paragraph of text.

The CDA solution must extract all types of fields, including machine print (any font), handprint, cursive, barcodes, bubbles and checkboxes. For maximum automation, ensure the solution can vote between multiple OCR engine results. Just obtaining the raw OCR data merely scratches the surface of what leading CDA solutions can deliver.

Data validation / validation rules / database matching

An intuitive and keyboard-friendly data validation interface is essential for enabling humans to quickly locate and correct any unconfidently extracted characters and fields. Validation rules should be supported (e.g., Field1 + Field2 = Field3), as should database lookup shortcuts. "Fuzzy" database matching for extraction and validation (for vendor and PO lookups, for example) is also important and should scale to very large databases (>1M records). Without "fuzzy" matching's ability to extract correct data from imperfect OCR, humans spend more time manually correcting misrecognized characters.

Machine learning of documents and data

Strong CDA solutions leverage AI algorithms such as machine learning, classification with machine learning, natural language processing and unsupervised learning to perform document clustering (organization), classification and separation, OCR, data extraction and human language understanding. Machine learning is especially critical: learning from samples to train the system and continuing to learn from production user input increases the system's document classification and data extraction intelligence over time—without the cost of maintaining rules.

Natural language processing of emails and other unstructured content

Natural language processing is a key component of AI for CDA; it helps drive better understanding of the content and sentiment of unstructured documents (like emails, letters and contracts) so humans don't have to intervene. The CDA solution must either include this type of technology natively, or call out to third-party cloud providers like Microsoft and Google via REST services.

Document and data exports and integration to systems of record

The CDA solution must support the export of documents and data to common ECM and ERP systems—without the need to write and maintain integration code. Ideally, the CDA solution includes pre-built export connectors to common destination systems and no-code ways to integrate with unsupported systems or systems lacking exposed APIs. RPA is ideal for integrating with hard-to-reach systems that lack exposed APIs.

Process intelligence

Process discovery and analytics are essential capabilities in helping the business identify automation opportunities and track performance of the CDA solution. Analytics should include document sources, classification and extraction automation rates, user productivity and costs per document and per channel, at a minimum.

Project customization

No two projects are the same: a CDA solution should make it easy to perform common functions and also enable (via scripting) more unique, application-specific projects. The ability to add script to CDA projects and easily debug scripts is paramount in making the system do exactly what is required by the business.

Integration with RPA and BPM/DCM

As mentioned earlier, RPA applications and requirements are expanding to include CDA and process automation (BPM and dynamic case management). Process automation is necessary to handle business rules, user forms and exception handling capabilities, at the very least. Any CDA solution should be a part of a broader platform that delivers these robotic and process capabilities to minimize complexities in procurement, licensing, operation and maintenance, and to ensure consistent strategic direction from the vendor offering these components.

The objective of any CDA project is to realize the expected benefits of greater visibility, lower costs, faster processes, fewer errors and improved customer engagement.

User productivity =
OCR accuracy + user efficiency

How to Measure CDA Success

The objective of any CDA project is to realize the expected benefits of greater visibility, lower costs, faster processes, fewer errors and improved customer engagement. The question is: how do we measure CDA success?

Simply put, the largest single indicator of success is user productivity - the degree to which people become more productive because CDA helps get work done better and faster.

User productivity is made up of two components:

- OCR accuracy
- User efficiency

OCR accuracy

Perhaps the most common question after watching a CDA technology demonstration is: What level of OCR accuracy can I expect to achieve?

The short answer is, it depends and varies widely across use cases. OCR accuracy, and more generally classification and extraction accuracy, depends on multiple factors, including:

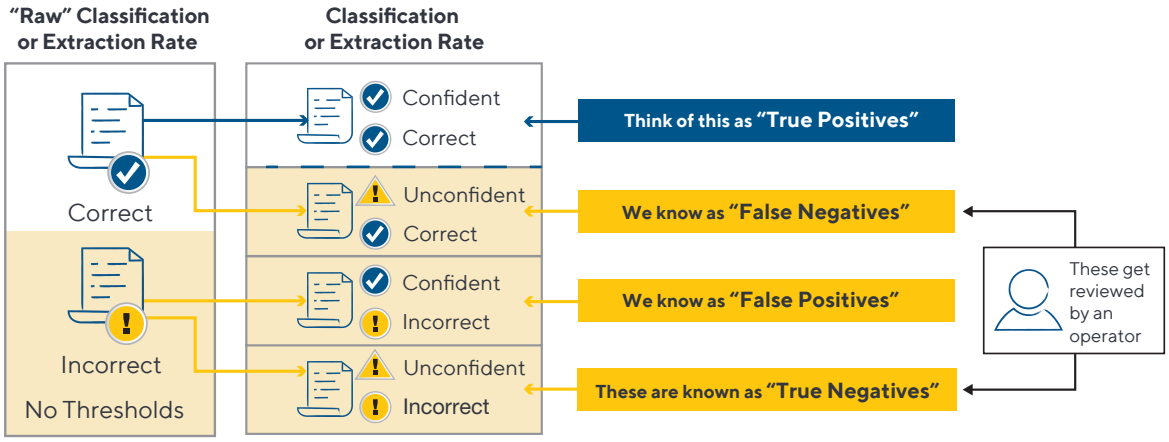
- Scanner hardware
- Scan resolution
- Image quality
- Document type (form, invoice, letter)
- Machine-printed/hand-printed/cursive
- Document language
- Font type and character spacing
- Field boxes/shadings
- Ability to database match or check for checksums and other rules

Some of these factors will be discussed later in the paper, but for now, remember this: the higher the accuracy, the more classification and extraction automation; the lower the accuracy, the less automation and more manual labor.

Because accuracy varies so much, it is best to perform benchmark testing for

classification and extraction accuracy rates on the business' actual real-world samples. Use these results to optimize project settings for each document type and field, and thereby increase accuracy and automation. Benchmark testing should record, for each document type and field, its confidence (yes/no) and correctness (yes/no), with the goal to maximize true positives, minimize false negatives and true negatives, and completely eradicate any false positives from being exported to downstream people, processes and systems.

A related term, "straight through processing" (STP), is also used as a metric to describe CDA results. This is the measure of the percentage of documents run through the CDA "acquire, understand, integrate" process untouched by a human. The STP rate will never be higher than the lowest OCR accuracy field on the document. To maximize the STP rate, focus on the lowest-accuracy fields being extracted on the document, and adjust settings for those fields.



Benchmark testing goes beyond if a field is correct or incorrect.

User efficiency

OCR accuracy is just one side of the coin of user productivity. The other side is user efficiency for exceptions. Documents and fields that don't pass through untouched (known as "low-confidence") must be reviewed by a human to ensure they are classified and extracted correctly. User efficiency is all about how quickly a user can review a low-confidence document or field, make a decision on what needs to be corrected/confirmed and then execute that decision. So the human validation interface must be designed for the most efficient use of eyes and hands during the document classification review and data validation process.

Here are some examples of user efficiency features in leading CDA solutions:

- Jumping to the field that needs to be validated, skipping over confident fields
- Highlighting that field on the actual image for context
- Displaying an image snippet of the field in question next to the data entry area
- Custom positioning of panels to each user's liking
- Correcting a single character in the field rather than re-entering the entire field
- Hitting a hotkey to call a database lookup for a field
- Auto-complete of the field based on the document type list or full page OCR
- Completing a table's worth of data by simply highlighting it

The effort spent on user efficiency and user experience will produce ten times the user productivity compared to the same effort spent on improving field OCR accuracy. That is why it is best to maximize a human's work speed processing these data exceptions via effective user engagement and minimal keystrokes and mouse movements.

User productivity

This brings us to user productivity, which is the combination of OCR accuracy and user efficiency, and represents the single most important metric for a CDA project's success. User productivity can be defined as: the number of documents per hour/day/week/month each staff member can process with an acceptable level of data quality.

For example, consider a mortgage application form. Some form fields will be more important than others, so an "acceptable level of data quality" will vary depending on the field. Benchmarking the CDA project to understand per-field OCR accuracy is necessary to optimize extraction rates for high-priority fields such as social security number and annual income.

When configured effectively based on the success metric of maximizing user productivity, CDA solutions will deliver an attractive ROI and payback, frequently between 6 and 18 months from system launch.

For a more detailed look into how to measure CDA success, including which metrics work and which don't; how to measure cost and labor savings before and after automation; and how to create a compelling CDA business case, see the white paper "Cognitive Document Automation Success Metrics: The Truth About OCR Accuracy."

User productivity, which is the combination of OCR accuracy and user efficiency, represents the single most important metric for a CDA project's success.

Tips for Navigating CDA Challenges to Maximize User Productivity

RPA customers implementing CDA solutions will not be without challenges in attempting to maximize user productivity (i.e., accuracy and efficiency). CDA solutions can still exhibit pitfalls and limitations. Below are some of these challenges, along with advice on how to successfully navigate them before embarking on your CDA journey.

Image source

The image source will affect the quality of the image and, therefore, the level of classification and extraction accuracy. Faxes will inherently have lower image quality than an emailed, born-digital PDF, for example. Scanner hardware delivers different levels of quality depending on the vendor and model.

Image file type and resolution

Some image file types have better inherent quality than others. 300 dpi gifs are most common, but frequently companies cannot control the file type received from external sources. Lower-resolution images will have lower levels of classification and extraction accuracy—300 dpi is considered ideal.

Image quality

The saying "garbage in, garbage out" also applies to CDA. Images faxed multiple times; mobile images with skew, tilt, blur, similar background or bad lighting; monochrome scans; documents with stamps, scribbles and stains… all of these can affect classification and extraction accuracy. All images acquired by CDA solutions should be image-processed and perfected before applying automated classification and extraction to ensure maximum possible accuracy.

Document collection

The number of samples and their similarity to the real world also impacts accuracy. Generally speaking, the more samples that are "machine learned" by the CDA solution, the better. The number of samples required ranges from a few to hundreds, depending on the type of document. These samples should reflect as closely as possible what will be seen in the "real world" during production processing.

Structured forms

Structured forms generally exhibit the highest level of classification and extraction accuracy, and require the fewest number of trained samples. Nonetheless, the form design will have a significant impact on accuracy—from proximity of fields to each other, to field boxes vs letter boxes, to field shading (if any). If the organization has control over the form design, they can design the form for maximum automation potential.

Semi-structured documents

Semi-structured documents (such as invoices, purchase orders, sales orders and bills of lading) generally show lower accuracy than structured forms. Different CDA solutions have different approaches for locating the desired data, and some are more reliable than others at finding the data and extracting successfully. These documents also tend to have embedded tables (e.g., invoice line items), multiple tables, or tables within tables that may have lower extraction accuracy rates than regular fields.

Unstructured documents

Unstructured documents such as emails (body), letters and contracts are the most challenging to classify and extract automatically. AI-based technologies such as natural language processing have improved extraction accuracy rates for these types of documents in recent years.

Print type

The type of print on the document also affects extraction accuracy rates. Generally, machine-printed fields have the highest accuracy rates, followed by hand-printed fields and then by cursive fields. For machine print, font type and character spacing also impact accuracy rates. Document language can also impact accuracy rates, as OCR engines used by CDA solutions exhibit varying OCR accuracy depending on the language, with Latin languages typically claiming the highest accuracy rates.

Barcodes and checkboxes

Barcode and checkbox fields typically show the highest extraction accuracy on a document. It is not uncommon for CDA solutions to boast in the high 90s percent accuracy in extracting barcode values and checkbox/bubble values. However, there are dozens of barcodes in use, including 1D, 2D and now 3D barcodes (2D with color), so make sure the CDA solution supports the most frequently encountered.

Signatures

One of the primary reasons paper is still in use by many organizations is the requirement for a signature, and the paper signature must be captured, classified and extracted. Moving to electronic signatures can remove the need for paper scanning, thereby improving the productivity and capacity of your CDA users. Consider whether you simply need signature presence detection, or signature verification and fraud detection as well.

Databases

A CDA solution's classification and extraction accuracy rates can significantly improve through the use of databases. By matching to similar content in databases, minor OCR errors can be ignored. This results in less human involvement to confirm/correct low-confidence OCR results. Database content can include customer names, account numbers, ERP data such as PO number or vendor name, word dictionaries specific to industries or languages, etc.

Rules

Rules can also be used to increase the extraction accuracy of a field. For example, checking that subtotal plus tax equals total is a simple rule that can flag any errors, even after a human corrects one of those field's values. Formatting rules are also a simple way to ensure high field accuracy (e.g., a social security number should always have the format xxx-xx-xxxx, where x is a number between 0 and 9). Checking for field values' checksums also increases field extraction accuracy.

Destination systems

CDA solutions are not complete without an easy way to send the documents and data to the systems, processes and people who need them. User productivity decreases immensely if users must manually move document images and data from one system to another. Remember that an RPA robot can automate the process of moving and aggregating data between systems if an out-of-the-box connector for the destination system is not available.

Why Kofax for CDA

Kofax is the only single-source vendor to offer both market-leading CDA and RPA for a complete intelligent automation solution. The company's patented AI-based approaches have been used successfully for over a decade to deliver CDA for the broadest range of document types and formats. No other supplier has the depth and extent of proven AI expertise for document understanding.

Kofax is also the only single-source vendor to offer CDA and RPA within a powerful end-to-end process automation solution that orchestrates processes, business rules, errors and exceptions, and human-based decision making, and delivers comprehensive process analytics. Kofax automates documents (CDA), tasks (RPA), and processes (workflow) in a single, unified platform for complete digital transformation of business.

Kofax is the widely acknowledged enterprise document capture/cognitive document automation market leader, with multiple CDA products successfully deployed by thousands of satisfied customers over two decades. Kofax has received market leadership recognition by leading industry analysts including Forrester, Gartner and IDC.

For more information on digitally transforming your business with intelligent automation contact your Kofax sales representative or authorized reseller.

Work Like Tomorrow.™

KOFAX

kofax.com