NTT DATA

POINT OF VIEW | DATA INTELLIGENCE & AUTOMATION

How Intelligent Automation Accelerates Document Processing

Augmenting data extraction with AI-enabled optical character recognition

APRIL 2021

NTT DATA

Table of contents

Paper, paper everywhere …

Did you know that enterprises are exceedingly dependent on paper documents despite the high penetration of digitalization? Even today, documents are critical to at least some business operations across industries — remember your passport application? And what about your bank or other financial statements? In the office, an average U.S. employee consumes around 10,000 sheets of paper in a year.[1] That's a massive 7.5 billion documents every year in the U.S. alone.[2]

The tangible nature of a paper document and the sense of security that it engenders make it difficult to let go of paper-based processes and switch to their digital twin for highly sensitive documents — such as invoices, contracts, correspondence, payment receipts, etc. This trail of paper doesn't only slow down an organization's digital transformation initiatives. It also requires substantial effort and manpower to manage, maintain, scan and integrate the information a paper document contains into a digital system and then properly dispose of it. This cost could be as much as 31 times the cost of purchasing the paper.[1]

COVID-19 has further pushed the demand for digital documents because enterprises want safe, contactless options to conduct business.[3] Take, for instance, the sudden surge in medical documents due to the pandemic. This demand has forced several industries, especially government and healthcare agencies, to look for robust document processing and data capture solutions.[4]

Several enterprises are turning to their long-trusted ally — optical character recognition (OCR) — for document processing. Afterall, OCR has not only served them well since the mid-1990s but become very pervasive. This technology is at work when we get our passports scanned at the airport, for example, or while driving on a toll road. It is also the technology of choice for receipt recognition in industries like retail and food delivery.

OCR engines have grown in stature and come a long way from when they were first introduced. Yet, these traditional document processing systems aren't equipped to meet the demands of modern enterprises. In this paper, we'll discuss the challenges of traditional OCR-based systems and how artificial intelligence (AI) can be leveraged to address these woes and make the entire document processing experience more intelligent, scalable and reliable.

Business challenges of document processing with OCR

Traditional OCR-based engines have limited capabilities to extract and process meaningful data from the variety of document types that exist in an enterprise environment. Data residing inside these documents may be structured, semi-structured, unstructured or hand-written; it may also be not-so-perfectly scanned. It could be hidden in text or PDF files, emails, journals, webpages, presentations, social media posts, meeting recordings and transcripts, among other contents. Whatever its location, this data may be crucial to business decisions.

Each document also needs a different approach and level of precision to be processed, because it can have different degrees of predictability based on structure and content. Consider a customer who sends a merged PDF document for invoice processing. That document includes 100 pages, and each page has a scanned invoice from a different vendor/sub-vendor. If you only need to process the invoices for a few select vendors while holding the others, the individual invoices will need to be grouped by vendor before processing begins. It's easy for human agents to create such groups on the fly (based on business needs) using a key visual element — namely the vendor logo on the invoice. However, it's not so easy a task for OCR engines.

Similar document categorization or classification may be required if, for instance, your sales team receives a set document with scanned copies of invoices, custom declarations and insurance information, and you need only the invoice data (which has business value). You'll need to parse the document carefully to avoid picking incorrect information.

Here's another scenario: processing only those invoices with authorized approval signatures, which may not always look the same and may be at different non-standard locations on different documents. Again, it's a simple task for human agents but challenging for document processing engines.

Or, what if there are "Approved" or "Not Approved" stamps on your paper invoices and you need to process only those that are approved? Because they're applied manually, the location and orientation of the stamps are random, making it challenging for an OCR engine to define a location or region to read and classify.

Figure 1: Sample stamp on documents which may have important information (date, revisions, approver details, signature, routing information and much more)

These challenges persist, even when coupled with robotic process automation (RPA).

Today, enterprises have the choice to opt for a document processing system powered by OCR and RPA-based platforms. When an RPA bot mimics human activities to handle repetitive back- and mid-office processes, an OCR engine acts as its eyes and helps search for specific information within documents, extracting text characters in images and scanned documents and then converting that content into a machine-readable format.

However, any real-world documents fed into an RPA platform for processing may not be in great shape. And even if you opt for an OCR engine that claims high accuracy, there could be challenges with individual documents that can make processing and extracting data difficult. Some of the common challenges and factors that may impact output quality and accuracy include:

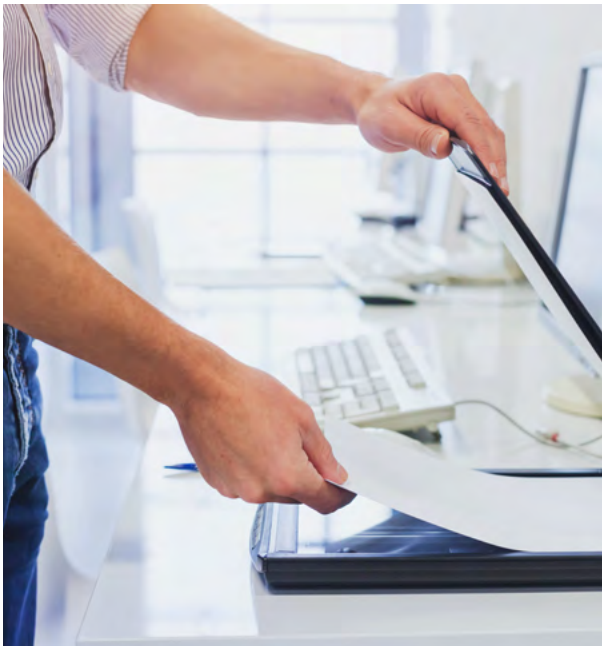
- Digitizing physical documents using a mobile phone camera, which may cast a shadow over the text
- Scanning the image of an already scanned document
- Scanner glitches and watermarks in a document
- Colored paper, background shading or blurry copies
- Skewed scans and carbon smudges
- Fold marks, staple damage and wrinkled pages
- Small text, unusual fonts and handwritten text
- Mathematical formulas

When you leverage RPA for semi-structured business documents like invoices and purchase orders, it either uses a region- and anchor-based text extraction approach or simply extracts the entire document and later recognizes the keywords. Both approaches have their challenges:

- The region and anchor approach fails when the position of the desired text changes from the predefined anchor or region. It's also likely to fail when a new, undefined template is sent for processing.
- The extract-all text approach results in a lot of meaningless text that needs to be dealt with carefully by writing rules to account for all the variables in the document's layout. It's extremely time consuming and may also lead to higher document processing times and costs (by, for example, quickly consuming your OCR subscription).

These traditional approaches, used in many RPA platforms, often lead to failed automation. The task then has to be routed to an often time-consuming manual queue, unless your organization opts for a more intelligent document processing approach.

Because the information an RPA bot requires to complete a business process may reside in either structured or unstructured documents, it's not easy to extract with extremely high accuracy. In addition, processing documents using different OCR engines may return different results, as one engine may extract the desired text while another results in meaningless text.



Document quality and data accuracy

It's pivotal for any organization to not only extract and capture relevant information but also validate its accuracy before using it in day-to-day business workflows.

A mistake in reading data accurately could impact business success. For example, you'll see a loss of $9,990.33 on your statement if your OCR engine reads $10.00033 from an invoice instead of $10,000.33.

The quality of the digital document plays an important role. Scanning documents with at least 300 dots per inch (DPI) is recommended for an optimum balance between quality and file size.[5]

Moving beyond OCR bottlenecks with intelligent document processing

It's time to look beyond the conventional template-based approach to OCR because it slows down the entire process and offers zero flexibility to accommodate variations in document formats. Occasionally, it may force you to bring human agents into the loop to make manual corrections, which ultimately defeats the entire purpose.

It's imperative to enable machines to self-identify what information needs to be extracted from a document and where it resides. Such capabilities will give you the freedom to process documents on the fly without worrying about templates and the need to define coordinates for data extraction.

The answer lies with AI. Machine learning can be trained easily to classify documents into different categories based on different parameters associated with each document. And natural language processing (NLP) can boost an OCR engine's performance to dig deep and harness unstructured data within a document and retrieve meaningful business information at very high speed. Computer vision, on the other hand, can extract data from non-textual information such as stamps, diagrams, images or tables. Today, AI-enabled data extraction solutions can expand the capabilities of OCR beyond templates and execute tasks previously reserved only for human agents — for truly intelligent document processing.

Understanding intelligent document processing engines

An intelligent document processing engine is a packaged set of next-generation tools and components that uses advanced technologies to read and understand a document, and then extract the desired information. It has relatively high accuracy compared to general-purpose OCR engines.

Several standalone products and solutions, such as Google Tesseract, Microsoft Office Document Imaging (MODI), ABBYY, Kofax and Parascript, work well with leading RPA platforms to automate a process using documents. One important aspect of an intelligent document processing engine is that it looks at the input document and uses an advanced mechanism to prep the document for OCR data capture by:

- Identifying and removing noise like watermarks and shadows in the document
- Adjusting the brightness and contrast of the document for better visibility
- Automatically rotating or de-skewing images within the document

AI augmented automation is going beyond the surface level. IDC predicts that around 1.4ZB of information will be touched by cognitive systems by 2025, affecting many businesses and organizations around the world.[6]

Introducing NTT DATA Nucleus Intelligent Document Processing

At NTT DATA, we've rebooted traditional OCR-based document processing for more intelligent data extraction. Powered by the NTT DATA Nucleus Intelligent Enterprise Platform, this advanced cognitive automation engine eliminates the need for a rules-based character matching approach. Not only does Nucleus Intelligent Document Processing dynamically extract relevant text from a document using deep learning and NLP, it also derives meaning and sentiments for valuable business processing.

Nucleus Intelligent Document Processing uses a convolutional neural network (CNN) model to dynamically match patterns in an image-based document and multiple target identifiers to extract the desired text, images, logos, handwritten text and signatures without having to manually define coordinates within a document. It uses NLP to dynamically identify the entities and values in a table, and even analyzes sentiments (how positive, neutral and negative the text is). The AI-based intelligent data extraction engine can recognize named entities (relevant nouns like people, places and organizations) and count the number of words in extracted text. Using CNN and computer vision, it can dynamically analyze and identify patterns like handwritten text or signatures in the file.

Domain-independent, Nucleus Intelligent Document Processing works seamlessly with RPA platforms and includes an option to plug-in third-party OCR/intelligent character recognition (ICR) engines via application programming interfaces (APIs) and dynamic link library (DLL) integrations to extend and enhance the engine's performance. Its self-learning capabilities continuously improve document processing accuracy while minimizing failures. It not only speeds up the entire process but also makes it dynamic and template-free.

One of our clients, a rapidly growing organization and a leader in the U.S. healthcare industry offering innovative healthcare products and insurance services, faced tremendous pressure coping with operational challenges in its provider onboarding processes. The high volume of documentation and contracts associated with its growth also left the company unable to adhere to compliance regulations. Its data-intensive contracts loading process was slow and inefficient and included data quality and compliance process issues as well as inaccuracies, which led to heightened enterprise risk.

We were able to transform the organization's traditional, inefficient, time-consuming and resource-intensive client onboarding processes with Nucleus Intelligent Document Processing. Our cognitive automation engine identified and extracted relevant information from the pool of unstructured documents, including complex contract and legal documents in different file formats, and provided instructions based on the extracted information and an intelligent algorithm used to perform the appropriate actions. It processed over 500,000 contract documents within 13 weeks — a task that traditionally would have taken 100 expert human agents about 8 months — saving the company approximately US$2.5 million and reducing its turnaround time from 45 days to under 24 hours.

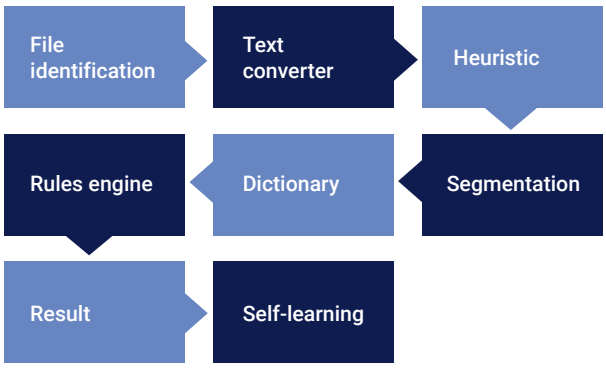


Figure 2: The unique processing sequence of the Robotic Context Processor engine

Solution benefits:

- Frees you to process documents in any format on the fly by easily identifying desired information
- Automatically handles document alignment and common OCR issues
- Logically merges pages in a document based on identifiers (logo or pattern) in the document
- Supports various document types and isn't restricted to processing vendor invoices

Conclusion

Traditional OCR-based document processing systems have served enterprises well for the last several decades. However, it's become clear that these systems are ill-suited to cater to our current and future document processing needs. They struggle with semi-structured or unstructured documents, in particular, because this type of content requires the creation of vendor-specific templates, which is inefficient, tedious, expensive and not scalable.

It's time for enterprises to integrate AI technologies with document extraction and processing workflows. This approach will not only increase the scope of automation but also enable efficient management of all types of documents, including those that are unstructured and low quality. Leveraging an AI-enabled intelligent data extraction engine such as NTT DATA's Nucleus Intelligent Document Processing will be essential. It will provide your organization with a much-needed productivity boost, as well as enable you to analyze, identify and extract important and difficult information, including text, logos, images, handwritten content and values in a table, which is often beyond the realm of traditional OCR-based engines.

About the authors

Dr. Harsh Vinayak, Senior Vice President, NTT DATA Services, Dr. Harsh Vinayak leads Intelligent Automation and Data Solutions and the R&D division. His background in advanced research and development uniquely positions him to provide clients with informed solutions based on extensive data analysis and forecasting.

Dhurai Ganesan, Vice President, Intelligent Automation and R&D, NTT DATA Services, Dhurai Ganesan leads R&D and Intelligent Automation Delivery. His interests include AI and cognitive automation. Dhurai has helped build end-to-end RPA creation, deployment and management ecosystems for identifying automation opportunities through an enterprise innovation program.

Let's get started

NTT DATA Nucleus Intelligent Document Processing combines dictionary data sets, a configurable engine, self-learning and updated libraries, and easy integration with RPA bots via APIs or DLLs. It automatically selects the optimum third-party OCR/ICR engine, based on the document analysis, for text extraction and handwriting recognition and extraction, respectively, providing exceptional quality and accuracy of information recognition.

Let us help you create a template-free environment that fails less and processes documents on the fly, just like a human agent. Contact **rpa@nttdata.com** or visit **nttdataservices.com/rpa** to learn how you can empower the document processing capabilities within your organization.

Contact one of our authors, or visit nttdataservices.com to learn more.

Sources

1. Minnesota Pollution Control Agency. “Reducing office paper use.” https://www.pca.state.mn.us/quick-links/office-paper

2. Bob Loblaw. “The Real Cost of Paper: Facts That Show What Your Business Spends (and Where It Can Save).” QLS Blog. May 29, 2019. http://www.qls.com/blog/the-real-cost-of-paper-facts-that-show-what-your-business-spends-and-where-it-can-save

3. Natasha Lomas. “Scandit raises $80M as COVID-19 drives demand for contactless deliveries.” TechCrunch. May 26, 2020. https://techcrunch.com/2020/05/26/scandit-raises-80m-as-covid-19-drives-demand-for-contactless-deliveries/

4. Chirag Shivalker. “COVID-19 Raises Demand for Document Processing and Data Capture Solutions.” HiTech. April 23, 2020. https://www.hitechbpo.com/blog/covid-19-raises-demand-for-document-processing-and-data-capture-solutions.php

5. Emma Foster. “Top Five Business Processes OCR Can Improve.” CMS Connected. August 5, 2019. https://www.cms-connected.com/News-Archive/August-2019/Top-Five-Business-Processes-OCR-Can-Improve

6. Xinwen Zhang. “The Evolution Of Natural Language Processing And Its Impact On AI.” Forbes. November 6, 2018. https://www.forbes.com/sites/forbestechcouncil/2018/11/06/the-evolution-of-natural-language-processing-and-its-impact-on-ai/#41e9ccbd1119

Visit **nttdataservices.com** to learn more.

NTT DATA Services, a global digital business and IT services leader, is the largest business unit outside Japan of NTT DATA Corporation and part of NTT Group. With our consultative approach, we leverage deep industry expertise and leading-edge technologies powered by AI, automation and cloud to create practical and scalable solutions that contribute to society and help clients worldwide accelerate their digital journeys.

NTT DATA
Trusted Global Innovator