DECODING INTELLIGENT DOCUMENT PROCESSING

Rajesh Agarwal
Senior Vice President, RPA

So let's start with the IDP definition. So as I told you, **IDP stands for intelligent document processing, which means that, you know, it's not like just data capture. It involves right from document classification, then capturing the information from unstructured documents, verifying, validating the information and then formatting it to the desired level and then putting it into the downstream applications.**

So right from classifying, extraction, validation, formatting and putting it into the downstream application, the whole thing is IDP. And not only the data capture.

Some people even confuse it with the OCR or ICR. So OCR or ICR definitely a part of it, but a very small part of it. There are a lot of things which are required to complete IDP, which includes the artificial intelligence NLP, machine learning, fuzzy logic, et cetera. I'll talk about that. So let's start with, you know, the document types. So IDP takes care of different types of documents which are in different format, maybe PDF,

TIFF, word, excel, jpeg and they can come from different sources, maybe like scanned folders, FTP or by email attachments. And the layouts may be very different.

So like to give example, invoices coming from different vendors or import collection document coming from different banks may not have the same layout. **So you definitely need an artificial intelligence layer to extract the information from these unstructured documents. And before extracting the information you need to classify it.**

So in the case of trade finance, you need to know okay this in the letter of credit. This is the bill of lading. This is the commercial invoice copy. This is the certificate of origin. So one layer of the IDP includes the document classification. And then depending on the type of the document, you need to extract certain information. So, for example, in LC, you need to extractthe LC terms and in invoices uou need to extract the amount, currency, supplier, et cetera. So that also, you have to

extract it from different locations.

So IDP consists of an artificial intelligence layer and a fuzzy logic layer which will enable the solution to extract information from anywhere on the document. And these documents, which are coming are sometimes the scanned images or paper based documents, which are not necessarily of a very good quality. So one of the challenges faced by many of the customers is that the document quality is not up to the mark and that's why the extraction is not very accurate. So there needs to be a layer in IDP which takes care of the image quality to enhance the brightness contrast, remove the dirt, remove the noise, remove the background lines which are interfering with the desired information is definitely required in IDP to take care of the image quality.

Then the OCR and ICR may happen to extract the information and post that, there should be an intelligent layer to remove the typical OCR errors like five and S eight and B zero and O, I and

L or one. So all those typical are known to people for many years. But how do we automatically correct them? That is part of the IDP.

So it takes care of some auto correction auto validation to ensure the accuracy and to enhance the efficiency of the system.

So there is a pre-process, there's a post-process and it takes all the challenges of the unstructured content and there should not be any need to create a template. So if you have to train the engine or you have to create the templates for every document layout, then it will be very very time consuming. And then it would reduce the benefits of the automation. So you need to have a layer which can automatically extract the information from these unstructured documents without creating templates. So that is one part of the IDP.

And then it should be able to format the information. Like the amount should be in certain decimals. The date should be in certain format, like the incoming document may have dates in different formats. We have seen almost like 30-40 data formats, but the output required for a system is in the certain format. So how do you

correct that? How do you format it automatically? That is part of the IDP.

And finally, the data has to be flowing into the downstream application, which can be an ERP, which can be a workflow, which can be a core banking system, which can be a CRM or which can be a proprietary application or a web application. So it should connect to the application.

If it does not, then you have to do something outside IDP to take care of it. But if IDP can put the information into the downstream application then it gives us an end to end flavor. So right from ingestion of the document, classifying the document, extracting, validating formatting the information and putting the information into the downstream application will complete the IDP, which is part of the automation and not necessarily the only part.

So if I talk about how it's different from RPA now, so RPA definitely takes care of the movement of data or mimicking the human action, but that does not take care of extracting the information from documents. Whereas IDP takes care of that part. And many of the organizations do RPA but they're lacking in IDP. So their document extraction is not automated.

They have to give a structured input RPA and that takes a lot of time. And the benefit of RPA gets minimized because of which they are not able to scale up the RPA.

So IDP in short complements the RPA because it takes care of the document processing, it handles lot of unstructured content, it takes care of a lot of quality enhancement aspects of the document and the data which makes the process highly automated. So IDP complement RPA. IDP consist of OCR and ICR, but not only OCR and ICR and lot of other things I talked about. And this can be extended to various industries right from banks and insurance, which are very, very document-centric. Up to going to manufacturing, which we have financial documents, logistics, they have the bill of lading or they have the fuel tickets or they have the vehicle logs, they have the license plates. Manufacturing and retail, telecoms, all of these organizations deal with a lot of documents. So besides RPA, if they make great use of IDP, they can make a great success of the automation.