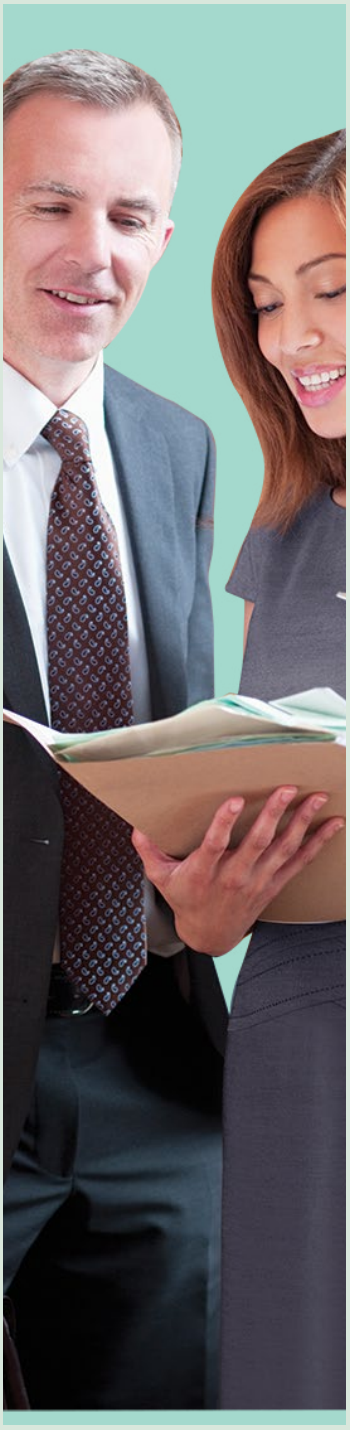
WHITE PAPER



DOCUMENT PROCESSING SOLUTION OPTIONS – STUDY FROM NORTH AMERICAN REGION

Author - Kaustav Datta, Ananth Sarma Brahmandam

Infosys®
Navigate your next

Introduction and Background

Document processing is a key area of many organizations such as Public Services in form of handling of their myriad Forms and Notifications; Manufacturing in form of handling of Purchase Orders, Invoices, Shipping; Financial Services such as Banks in the area of Customer On-boarding, Account Opening – KYC / AML checks and Credit evaluation.

Enterprise Content Management (ECM) combines the capture, search and networking of documents with digital archiving, document management and workflow. ECM developed as the industry matured from electronic document management system (EDMS), typically catering to the need of one department in the areas of imaging, workflow management, document management (indexing for easy search and retrieval). EDMS applications were typically a small-scale application catering to the need of a department to improve a paper intensive process into a paperless one for easier management and better accountability.

According to AIIM (Association for Information and Image Management), ECM is the strategies, methods and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. It's the architecture that glues an enterprises' documents and business content together — making them searchable, explorable, organized, and ultimately meaningful.

The building blocks of a successful ECM consist of the following :

New age technologies such as Cloud, Automation, Machine Learning, IOT etc. is completely re-defining the way content is being created, captured, indexed and retrieved. Disruptions around customer, supplier, partner data being spread around a variety of platforms from traditional On-Prem to Cloud, traditional Relational, NoSQL to Graph databases and a plethora of documents that exists is making enterprise content extremely fluid and agile. Traditional "Monolithic" Architecture for ECM is giving way to new concepts like "Content as a Service", which is a shift for more practical, modular set of solutions that deliver on the fundamentals of ECM such as capture, manage, store, preserve, and deliver.

In this paper we are going to talk about transformations that are currently redefining the Capture process, by looking at various solution options that are currently available in the market to digitize the incoming content (both printed and handwritten) so that it becomes easier to process the information as part of downstream business processes. We will use our experience in Public Service domain to establish a framework for processing these documents and provide a variety of solution options which can be leveraged for one's specific business context.

Problem Statement and Use Cases

As mentioned above, processing an incoming document and extracting meaningful information from the same to automate business processes are applicable in the entire enterprise value chain across various industries. At a minimum this spans across Inbound Logistics, Manufacturing and Operations, Outbound Logistics, Sales & Marketing and Customer Service – the primary activities of any firm and as a result this functionality is applicable for CRM (Customer Relationship Management) , ERP (Enterprise Resource Management) and SCM (Supply Chain Management) systems.

Following are some of the high-level use cases across industries wherein document processing is a key part of the business processes.

Manufacturing	Healthcare	Banking	Retail	Public Services
Purchase Order	Clinical Research	Customer On-Boarding	Customer Service	Visa Processing
Invoice Processing	Prescription Management	Commercial Banking Account Opening	Returns and Credit Management	Customs and Immigration
Customer Service	Claims Processing	Customer Service		Registration & Licensing
				Tax Processing
				Legal Documents Processing

Documents received to start these business processes could be received either through physical mail or email or fax or even submitted online through website or mobile app. These documents could be either handwritten or printed in a structured manner. Typically documents received through physical mail, email or fax are scanned and converted into either PDF, TIFF or JPEG format. Documents received through website, mobile could be already available in these format.

For our illustration of solution options for this paper, we will consider the use case for a Public Service firm operating in the North America region. This program receives over 200,000 reports annually, both in printed and hand-written format. The current process involves manually extracting the information from the documents and feeding it into Oracle Case Management system. Our solution was to look for automating this manual process to being in more process efficiency and increase productivity.

One of the first steps of this program was to capture information from multitude of these incoming documents into electronic format so that these can be further used for tagging, indexing for easier search and retrieval and also used in further downstream business processes such as Case Management – Analysis, Processing and sending out Notifications, Customer Service – responding to Client enquiries etc.

At a high level the initial processes of document capture could be depicted as the following :

One of the major pain-point this Program is currently experiencing is the high manual intervention required in the document scanning and information extraction from the received documents so that this can be used in Oracle Case Management process. The firm is looking for considerable amount of Case automation end to end from Document intake to Case resolution. This becomes even more critical for pandemic situations like Covid-19.

High level Solution Design and Solution Options

At a high level solution to extract information from printed and handwritten documents would have the following building blocks :

The solution would typically make use of an Optical Character Recognition (OCR) or Intelligent Character Recognition (ICR) solution to extract the text from the incoming image file/ document and send over to backend system in suitable format.

The solution would have two principal components - a Robotic Process Automation (RPA) suite and a selection of OCR APIs.

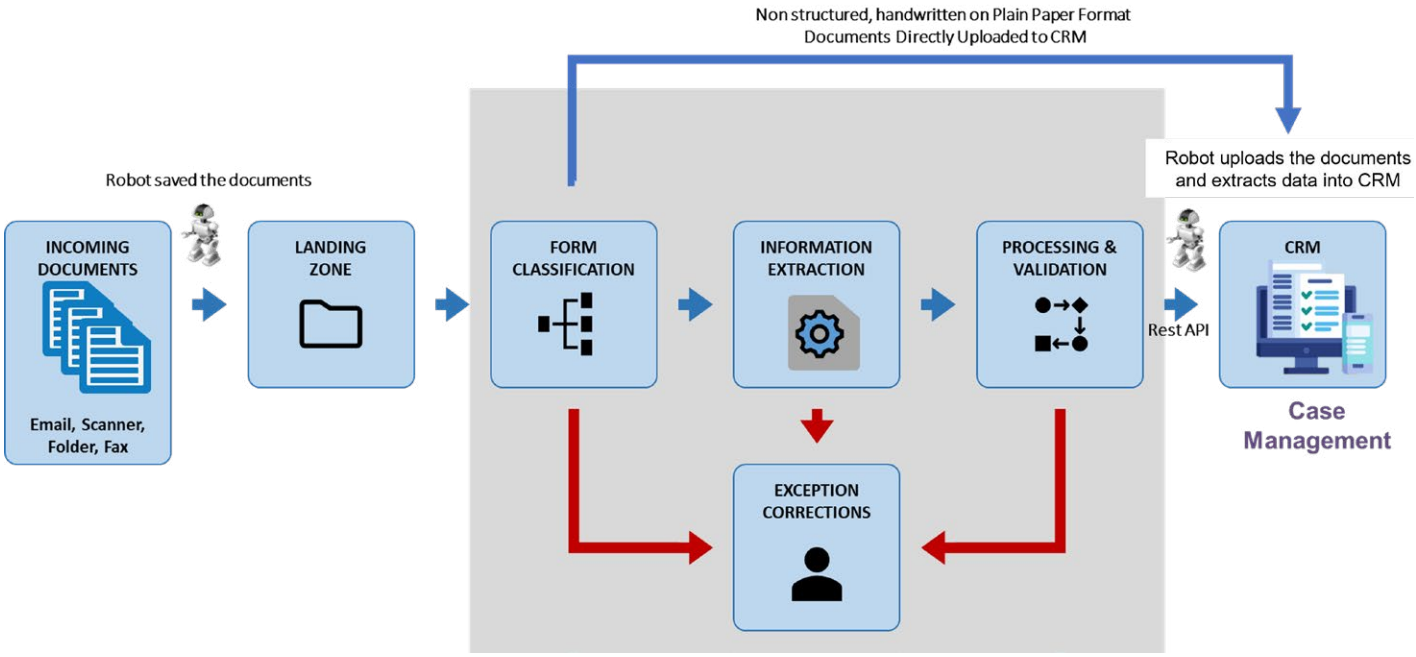
RPA would act as an orchestrator for this entire process. RPA tools such as UiPath, Automation Anywhere, Blue Prism, AssistEdge etc. would integrate with a variety of source systems such as Scanners, Fax, Email and act as an entry point for the documents. As part of the incoming document processing, the RPA tool would also integrate with commercially available OCR APIs such as Azure Computer Vision, Google Cloud Vision, OCRopus, Tesseract etc. to extract information from the printed or handwritten documents. The extracted information would then be presented to the end user for final validation and correction after some amount of automated data correction (post processing). Once validated and approved, the extracted information would then be sent over to backend system in this case Oracle CRM for Case Processing, either in XML or JSON format.

The choice of OCR API is a fundamental and crucial of this entire process. There are quite a few which are commercially available. Most of the OCR APIs work in one of the following ways - Optical Character Recognition (OCR) – one character at a time or one word at a time, Intelligent Character Recognition – works on handwritten text one character or one word at a time leveraging machine learning. For our use case mentioned above, we had evaluated both free and open source options such as OCRopus, Tesseract as well as cloud OCR APIs such as Google Cloud Vision and Microsoft Azure Computer Vision.

The following table compares these various OCR solutions against a variety of parameters –

Parameters	OCRopus	Tesseract	Google Cloud Vision	Microsoft Azure Computer Vision
Cloud / On Premise	On Premise	On Premise	Cloud	Cloud
Printed / Hand Written	Need high resolution images (above 300 dpi)	Not so effective on handwritten input	Helps on face recognition, image labeling and landmark detection also	Best on handwritten content
Configuration	Would need Python and would need to manage dependencies	Written in C/C++. Would need significant installation effort	NA	NA
Output	hOCR	hOCR, pdf, text	JSON	JSON
Open Source	Yes	Yes	No	No
Free / Paid	Free	Free	Chargeable	Chargeable

At a high level the solution for the incoming document processing would look like the following-



Recommendation Solution Option and Conclusion

Based on our findings during this evaluation exercise, we found that while most of the commercially available OCR APIs was able to extract information to a large extent from a printed form, they usually falter on extracting information from a handwritten document. Based on our analysis, we found that Microsoft Azure Computer Vision API was relatively better than other APIs and was able to extract information from handwritten documents.

One thing to note is that every business use case has its own extraction, processing and comprehension need that require a very specific and suitable technology solution. Most available solutions in the market do not address all aspect of end-to-end business need, rather they are usually specialist in specific areas such as document capture, RPA platforms or Cloud based OCR API.

Based on our analysis during the implementation of this solution, we found that some amount of pre-processing and post-processing would be required to get an optimal output. Pre-processing would typically include steps such as de-skewing the scanned document, removing negative spots and smoothing edges, converting an image from color or greyscale to binary, layout analysis or zoning, character isolation and segmentation etc. Post-processing might require usage of lexicon, near-neighbor analysis etc. for contextualized spelling correction, intent extraction, auto correction of date format, special handling of numerals etc.

Image processing, Character recognition is a nascent area of study and is evolving constantly through usage of multitude of AI technologies such as k-nearest neighbor (KNN) algorithm, artificial neural network etc. Our recommendation is to use a framework that combines a cluster of these AI technologies together to provide an end-to-end holistic solution. Some of the key features of this framework could be ~

- Enhanced Pre-Processing
- Document Classification / Indexing
- Handwriting Processing and Signature Detection using Parallel Neural Processing
- Substantial Post Processing
- Continuous Learning and Improvement

About the authors

Kaustav Datta, *Principal Consultant*, having more than 20 years of experience in leading technology engagements in the area of Customer Experience and Digital Transformation in the domain of Financial Services, Public Sector, Automotive.

Ananth Brahmandam, *Senior Principal Architect*, having 25 years of IT experience largely in CRM Transformation Programs across the globe in Telecom, Public Sector, Retail.

References

- https://info.aiim.org/what-is-ecm
- https://en.wikipedia.org/wiki/Enterprise_content_management
- https://source.opennews.org/articles/so-many-ocr-options/
- https://en.wikipedia.org/wiki/Optical_character_recognition

For more information, contact askus@infosys.com

Infosys®
Navigate your next