

# A Practical Guide to Building an AI-Powered Document Processing System

When it comes to manual document processing, the business adage "time is money" couldn't be more apt.

## **Data entry and processing are costly.**

Workflows that involve physical document management, or exchanging PDF or DOC files via email, often result in bottlenecks and substantial data entry costs – as much as \$20USD to manually process a single document.

## **High-value data is inaccessible.**

Even in digitized industries, countless documents still exist only on paper, essentially locking down valuable data. Some estimates put the amount of business data in unstructured, non-digitized paper documents into double-digit percentages.

## **Fraud detection is complex.**

Any organization that offers online services is laser-focused on the consequences of fraud, with losses in the US costing billions annually. Implementing manual document review processes is a step in the right direction, but even that comes with a cost.

This is where automated, AI-driven document processing comes in.

With the ability to automatically ingest documents, perform Optical Character Recognition (OCR), extract relevant text, and understand and classify documents, AI document processing is becoming an essential tool for modern organizations with ambitious goals.

From invoices to purchase orders, mortgage applications to government forms, virtually any organization can use this technology to reduce costs, improve efficiency, protect customers, and unlock data that can power hugely impactful analytics-based insights. All you need to do is build the right foundation.

In this guide, we'll share insights and techniques that you can use to plan a Minimum Viable Product (MVP) of your AI Document Processing System (DPS).

## Planning Your DPS MVP

The intent of an MVP is to (a) produce a baseline capability that can provide immediate quantifiable business value and (b) provide a foundation for iterative development that can extend the scope of the MVP.

To ensure the success of your DPS MVP, the following recommendations should be part of your discovery process. While this list is non-exhaustive, it provides a reasonable foundation for your business to make the right decisions from the start.

→ **Define your goals and use cases.**

Understand the current state and pain points, such as throughput, failure rates, and fraud concerns. Being able to articulate your needs and success criteria will help you create a strategy and action plan for the MVP.

→ **Understand your document types and formats.**

Identify the types of documents you need to process and their formats, such as PDFs, scanned images, or electronic documents. This will inform the OCR and document-understanding technologies you use.

→ **Develop an OCR strategy.**

Determine the best OCR strategy based on your document types, formats, and quality. Depending on the quality of the documents, you may need to use advanced OCR techniques to extract text accurately.

→ **Develop an entity extraction strategy.**

Identify the entities you want to extract and understand from the documents. This can include named entities, dates, addresses, document coordinates for graphical elements, and other information relevant to your needs.

→ **Develop a document understanding strategy.**

Determine how to understand the document content, including the context and relationships between entities, as well as any relevant business logic or rules.

→ **Develop a document classification strategy.**

Identify the different document types that you need to classify, and develop a classification model using machine-learning techniques, such as supervised or unsupervised learning.

- **Determine whether HITL is needed.**  
Depending on the complexity and variability of the documents, you may need to use Human-in-the-Loop (HITL) to verify and correct the output of the automated DPS.
- **Develop a fraud scoring and detection strategy, if applicable.**  
If you need to detect fraud or suspicious activity in the processed documents, develop a fraud scoring and detection model using machine-learning techniques, such as supervised or unsupervised learning. Additionally, you may find HITL a critical part of fraud detection. For example, your DPS may pass on the above data, including a fraud score, to a human adjudicator.
- **Define an architecture.**  
The level of flexibility and leverage you need to build into your architecture will depend on a number of factors, such as: whether your DPS is standalone, whether it's part of a bigger picture technology endeavor, and/or whether you'll need to integrate with other technologies.
- **Test and feedback/refinement plan.**  
Launching an MVP that produces low accuracy out of the gate is a recipe for failure. To improve accuracy and reduce false positive/false negative rates, you must have a test plan and process to adjust AI parameters based on visual confirmation of performance.

## Building on Your Successful MVP

An AI-based DPS MVP is considered a success when you can measure significant improvements in the time and money required to process a single document: from hours to seconds, and from dollars to pennies.

Once you have launched your MVP and measured its success, you can start to explore how you can use the DPS to drive value elsewhere. For many organizations, integrating the DPS into existing digital systems, or building a digital platform around your document processing core, is the next step. Either strategy provides a level of seamlessness that will lead to use-case scalability and economic leverage.