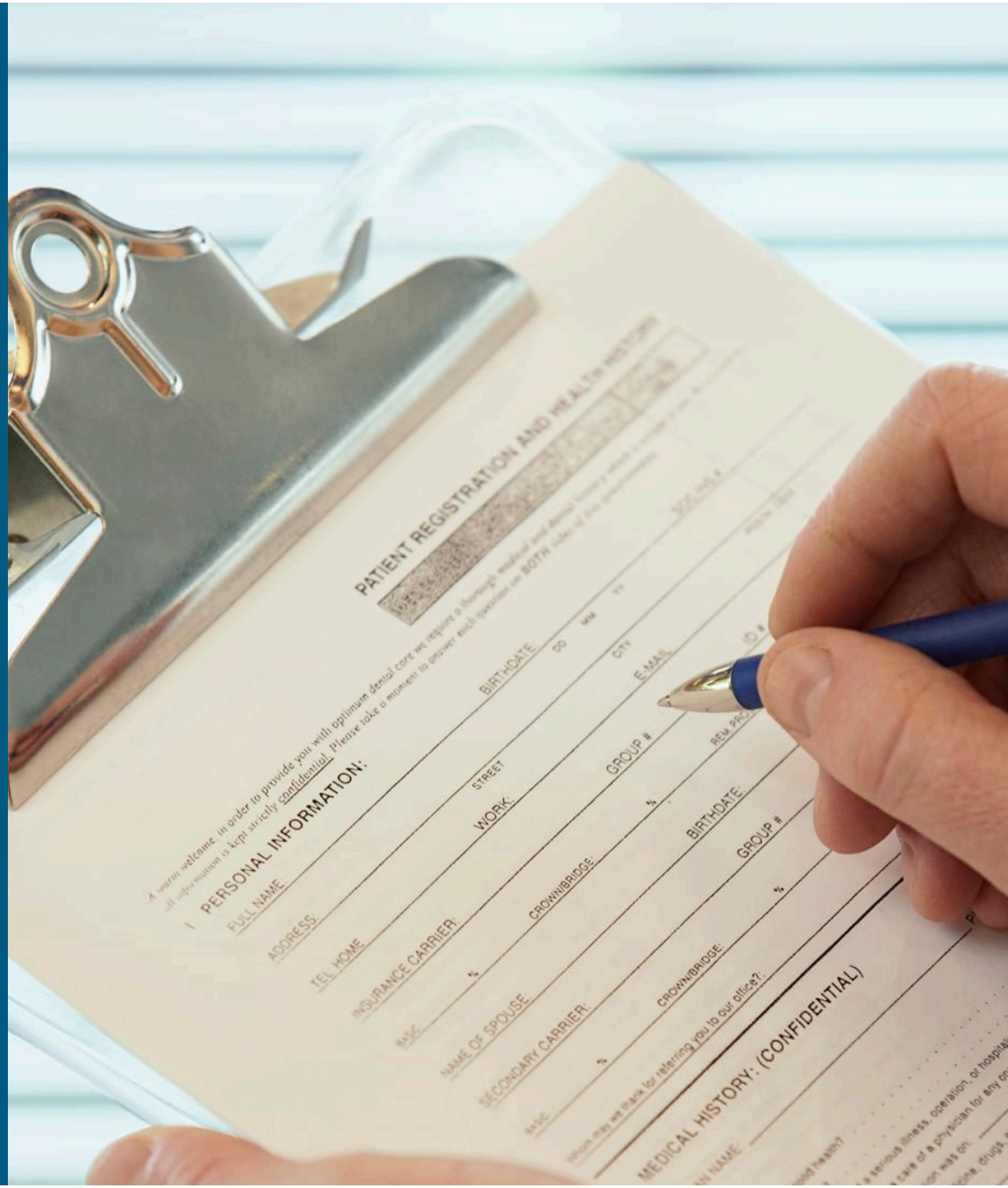
Intelligent Document Processing (IDP)

A Buyer's Guide

PARASCRIPT®

Table of Contents

Contents



Intelligent Document Processing (IDP) Projects

Intelligent Document Processing (IDP) Projects

Determining the objectives of an IDP project is difficult enough without adding the complexity of identifying and understanding vendor solutions. Rather than include these capabilities, this buying guide focuses on the key capabilities of modern IDP solutions and how they map to your business requirements.

While some capabilities—such as availability of APIs or methods of document capture can be easily compared. Other capabilities, such as document classification and data extraction, can only be compared by analyzing actual results to verify the accuracy of the output. For these areas, a checklist cannot replace actual system testing.

We will identify specific areas that require real testing.

Before delving into the specific capabilities, it is important to first cover the primary differences between OCR SDKs and solutions focused on forms and IDP.

The Difference Between OCR & IDP

The Difference Between OCR & IDP

When defining solution scope and evaluating document processing automation, the most common challenge is comparing full-page OCR technology with forms and other document automation technology.

OCR vs. IDP

Optical Character Recognition (OCR) is a technology designed to identify and extract text from digital images. By converting image elements into usable data, OCR facilitates the integration of paper-based text into various software applications and electronic tools.

It is commonly used to recognize text in scanned documents and images. OCR software can be used to capture text on a physical paper document or image and turning it into an accessible electronic version.

This technology forms the foundation for Intelligent Document Processing (IDP).

The difference between OCR and data extraction solutions like IDP is the amount of additional programming that goes beyond the basic OCR. OCR software typically performs a full-page conversion of a document image and can deliver data such as characters located and their X/Y coordinates.

Additionally, table data can be identified by locating the presence of rows and columns. The output is still a literal transcription, minus any character recognition errors. If a project simply requires a transformation of document images into machine-readable text, then OCR technologies can be an adequate solution.

Locating specific data within documents

Intelligent document processing extends beyond basic text capture to contextual interpretation. If the project scope requires the identification of particular information within a document, additional development becomes essential. This involves transforming the literal content of the document into actionable and usable data.

For example, when an invoice is captured by a basic OCR solution, it can generate machine-readable text, but it doesn't distinguish between invoice number, date, or total amount without further processing. It is the same for document classification, routing, or validation of information.

Comparing full-page OCR technology with forms and IDP

Programming Requirements

OCR requires additional programming or manual handling to achieve the level of same data results as IDP. For example, after using OCR, programming or manual tasks can be employed to extract the invoice number or total amount from the invoice data and integrate it into the business system. *This functionality is available standard in IDP solutions.*

IDP

IDP solutions take the OCR output and convert it into meaningful information. This can be categorizing, pinpointing, or extracting precise data from a page.

Some solutions provide the necessary technology and algorithms to process OCR output without the need for additional development, but accuracy can be a concern.

Other solutions provide user interfaces that allow the creation of rules without requiring staff with programming skills.

Field-Based Processing

While OCR solutions can provide image clean-up, the result is still text output. For documents using pre-printed data or document structure like tables, any data overlap might compromise the capture. This leads to data loss.

Some IDP solutions offer field-based processing that can clean up and even remove document structures. This leaves only relevant information for extraction and improves preservation of data.

Example of overlap in form requiring handwritten information. The left example will process correctly with a basic OCR solution, but the example on the right might generate an error.

Workflows

Most OCR solutions stop at output data and have no built-in ways to review and correct errors.

IDP solutions are equipped with the necessary workflows and software to shepherd an entire IDP process from input to final output. Additionally, there are solutions that integrate seamlessly into broader business processes or systems.

Image Processing

Image Processing Capabilities

Image processing involves preparing an document image to be efficiently classified, separated, and data-extracted.

1. Detect DPI (Dot Per Inch)
2. Scale, flip, or rotate
3. Remove introduced noise
4. Remove field-level form structure and pre-printed text
5. Reshape distorted images
6. Adjust contrast and brightness
7. Remove backgrounds and watermarks

Improperly scanned documents are common, and IDP solutions offer basic pre-processing to mitigate issues. Common errors include documents scanned at an angle or upside down and introduced noise (e.g. specks, streaks, or borders) caused by the device used to digitize the original paper document.

Modern systems account for today's multi-channel document scanning needs that include fax, portable scanner, and smartphone imaging.

The quality of the image is fundamental to accurate IDP.

While systems can handle a wide variety of image quality issues, process improvements like incorporating high-quality scanning should be considered in project scope.

The goal of image processing is to get the image to a condition that will optimize classification, separation, and data extraction.

Be sure to select examples of poor-quality documents and compare the system's ability to rectify the problems.

Document Identification, Separation, and Classification

Document Identification, Separation & Classification

Identification & Separation

The capability of a capture system to identify document types and distinguish the start and end of each document is crucial for large-scale processing. Although using document separators such as pages or barcodes is common, modern solutions offer features that eliminate the need for 'batch preparation'—the task of organizing documents before scanning.

Automated Classification

Advanced systems can automate document classification. A system capable of automatically distinguishing between invoices or receipts reduces manual preparation. Employing machine learning that develops and improves document classification eliminates the need for a person or team to organize and prepare documents for the next processing step.

Automated classification empowers a business analyst to establish document types effortlessly by submitting samples of a specific document class to the system. This process teaches the system to recognize future ingested documents.

The software analyzes the document and identifies key characteristics important for determining the type during production. Additionally, systems can then create individual documents automatically from the stream of scanned images.

This enable individuals to import a batch of documents while removing the necessity of sorting or inserting separator pages. Additionally, the system identifies attributes that specify particular pages, such as the first or last page, while treating all other pages as those in-between.

Data Extraction & System Configuration

IDP System Deployment

Configuration & Deployment

While you can compare the solution capabilities for data extraction in terms of the supported techniques, systems will differ in performance even when both use, for instance, keyword/value pairs.

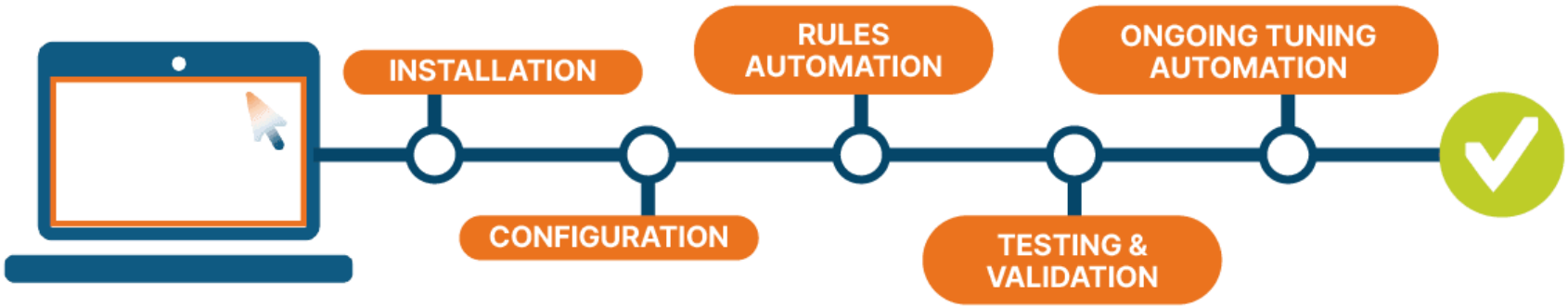
This is because systems use different means of internal validation of field output. This affects both the accuracy of the data itself as well as the confidence scores associated with the data.

When comparing systems, the evaluation of the amount of output and amount of accurate output is insufficient. If your organization wishes to achieve optimal straight through processing, you will need to evaluate *the amount of accurate answers along with the reliability of confidence scores for both accurate and erroneous data.*

System Configuration

Traditionally, configuring systems required labor-intensive manual rule creation and coding using text parsing algorithms. However, contemporary solutions have revolutionized this process, simplifying it so non-programmers can build comprehensive classification and data extraction workflows. This is valuable when the staff lacks programming skills, both during the initial setup and subsequent adjustments in a production environment.

Streamlined systems for classification and extraction, achieved through the reduction of necessary steps and the provision of flexible



Capture System

Classification & Extraction Configuration

Some systems require the use of more than one application to configure rules, so it is critical to understand what is involved in getting your IDP up and running.

The use of simple-to-use User Interfaces (UIs) enables efficient and low-risk configuration. Systems that support a full configuration across different document types (e.g. structured forms vs. unstructured documents) with one application are the easiest to use as it interface reduces the complexity involved in configuring and supporting different classification and data extraction rules.

Configuring automation rules is only part of the process. You will need to consider how those rules perform on production-level documents.

Solutions enabling users to test document classification and data extraction results without executing the entire capture workflow significantly simplify the iteration process for maximizing system accuracy. These solutions allow testing results at the configuration stage rather than requiring the user to completely configure the entire IDP workflow.

Another process improvement is solutions that include the option to automate rule creation. Some solutions take user feedback during production (such as data entry operator or actual machine learning) and use it to improve data extraction and classification. This is called ‘online learning’

Alternatively, offline learning allows a user to teach a system how to classify documents or extract data, all without programming knowledge.

System Precision

Organizations stand to gain efficiencies from adopting solutions offering automated, machine-learning-based document classification and separation. However, the most accurate method of comparing these systems is by evaluating the quality of their output, specifically assessing the system's ability to correctly classify and separate documents within a sample.

To do this, one must gather a representative set of documents, not just a few, run them through the system, and then directly compare the precision of one system to another.

This is critical when documents vary in layout, scale, or when there are differences in image input resolutions (e.g., images from faxes versus scanners). This will ensure the system can process different configurations and variances between documents.

Having a diverse range of location and extraction capabilities ensures that you can optimize the effectiveness of implementation without incurring additional labor costs.

Capture System Deployment

IDP is a complex process often involving many different steps and underlying technologies. Vendors approach their solution offering in a myriad of ways. They range from a modular approach that involves installing and configuring many different applications to databases or workflow engines. Other vendors adopt a more simplified single application strategy that involves fewer 'moving parts'.

Many systems use third-party applications to present a full solution. These applications might be included at no additional cost, or they are prerequisites that must be purchased and configured separately.

Understanding what is part of the standard software package vs. what is not is essential to understanding the total cost and requirements.

In addition to IDP, you have choices regarding the automation of the installation and configuration process itself. Some systems support a very simplified single-computer installation to enable the ability to quickly test the software without incurring a lot of effort.

Integration, Operations, and 3rd-Party Integrations

Integration, Operations & 3rd-Party Integrations

Integration

Traditional systems provide a ‘black box’ solution with limited integration and deployment capabilities. Businesses have to deploy them as stand-alone systems that release data into another system.

Modern systems have incorporated the ability to provide deep integrations to deploy wider capabilities as a platform instead of requiring a stand-alone implementation.

Consider these aspects when selecting a solution if your organization has invested in business process management/ workflow software or uses line-of-business systems.

By integrating classification and extraction capabilities into these applications, user processes see minimal disruption and processes becoming more streamlined.

Operations

When it comes to managing your classification and data extraction processes, the most important aspects involve understanding efficiencies associated with core system accuracy along with issues relating to slow-downs in throughput with a significant impact on successful business processes.

Many solutions offer data on processed document count, processing time, and process status. However, modern solutions go further by collecting and displaying information related to underlying factors influencing throughput and accuracy. These include individual staff efficiency, the ability to redirect work to available or higher-throughput staff, and the ability to reprioritize overall tasks.

When it comes to system accuracy, solutions can report on accuracy at the field level and alert if averages drop below a certain threshold.

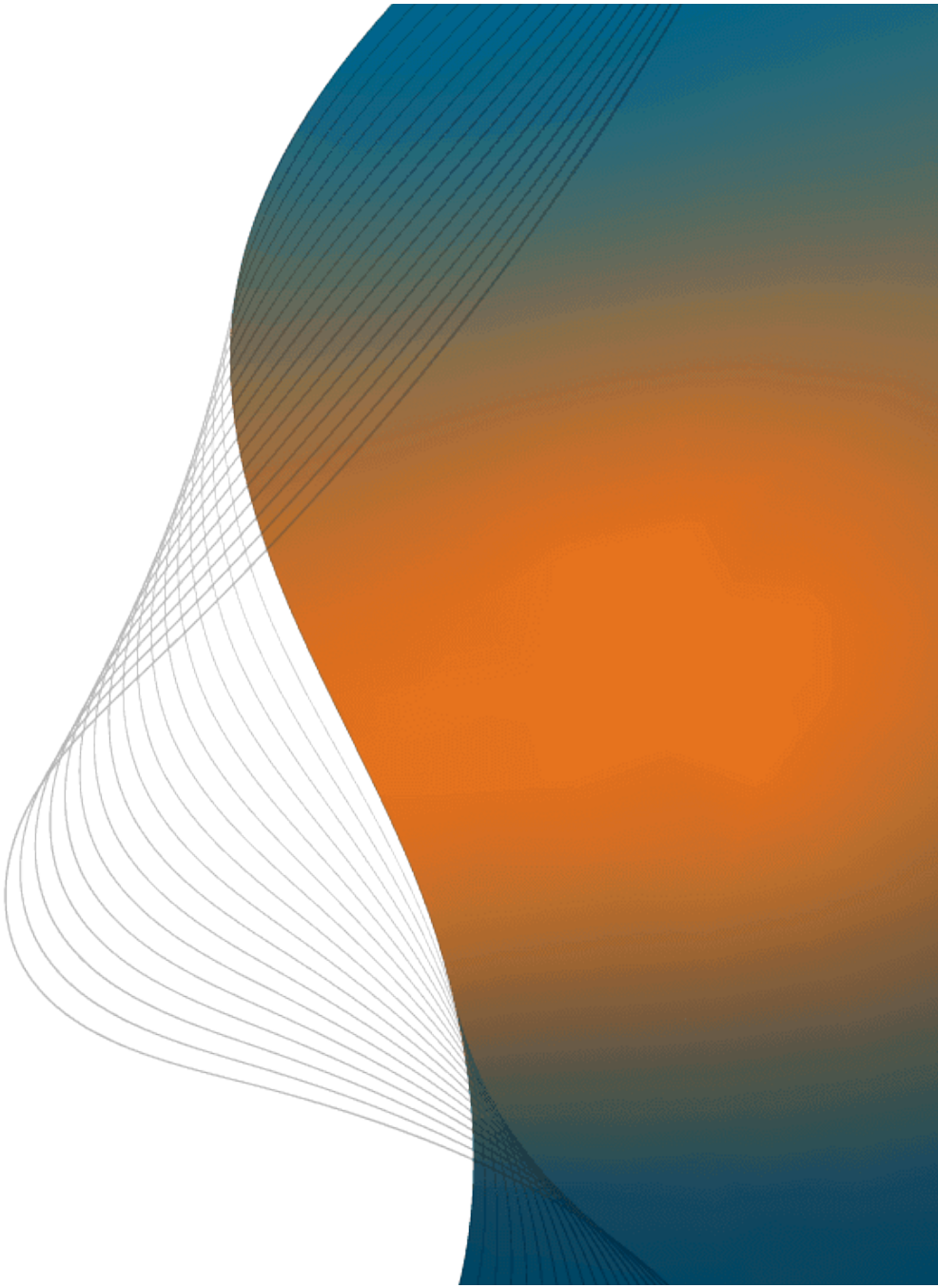
3rd-Party Integrations

No IDP software lives in its own world. It must support other processes. Whether it is an accounts payable system, CRM, or ERP system, there is a need to get documents and data into the proper system for utilization.

Pre-built integrations reduce the dependency on IT or system developers, leading to less time spent on maintenance and quicker implementation.

Interactive poll not supported

View online version



Checklist for IDP Systems & What's Next

IDP Capture System Checklist

See It To Believe It

When determining the capture solution that best fits your organization's needs, the variety of available options can be overwhelming. This comprehensive set of criteria goes beyond the typical capabilities that advanced capture systems offer, and it will help you select the right system for your organization.

IMAGE PROCESSING CAPABILITY

- Detect DPI
- Detect & re-scale
- Field-level form structure removal
- Field-level removal of pre-printed text
- Reshape distorted images
- Remove backgrounds/watermarks
- Remove introduced noise

CLASSIFICATION TYPE

- Visual classifiers (no OCR)
- Content classifier
- Combination

ADVANCED SEPARATOR TYPE

- Rules-based document separation
- Automated first-page separator identifier
- Automated last-page separator identifier

DATA EXTRATION FILES

- Native support for electronic documents (no OCR required): PDF, RTF/Word, spreadsheet, presentation, email, HTML
- Support for full range of document types (structured/semi/unstructured)
- Support for full-range of data types including unconstrained handwriting

CONFIGURATION CAPABILITY

- No-code classification and extraction configuration
- Ability to test configuration easily in designer application
- All range of data and document types using one designer (fixed data, variable data, handwriting, text)
- Pre-built document types mean no configuration is required
- Offline learning system
- Online learning system
- Configuration-level testing and tuning

DEPLOYMENT SUPPORT

- Single computer option
- All-in-one installation & configuration

- No 3rd-party software prerequisites
- No-database required option

INTEGRATION SUPPORT

- Full set of REST services for both SDK and full workflow
- Fully-embeddable granular .NET API for the SDK
- Available as both an SDK and full capture or, using APIs, any range in between
- No-database required option

OPERATIONS CAPABILITY

- Workflow analytics
- Accuracy reporting
- Ability to prioritize batch workload
- Ability to load-balance staff work

PRE-BUILT INTEGRATIONS

- Document Management

What's Next

Once you have prioritized your goals for an IDP system, it will be easier to determine which of the capabilities are critical to accomplish your goals. You can use our list of capabilities as a reference, or survey your team members who will be using the IDP solution to see what they need to accomplish.

From there, determine the pricing model that best fits your organization. If you are still unsure of the available technologies, seek expert guidance where available.

Thank you for reading

Intelligent Document Processing Buyer's Guide

Contact us

sales@parascript.com
www.parascript.com
(888) 225-0169