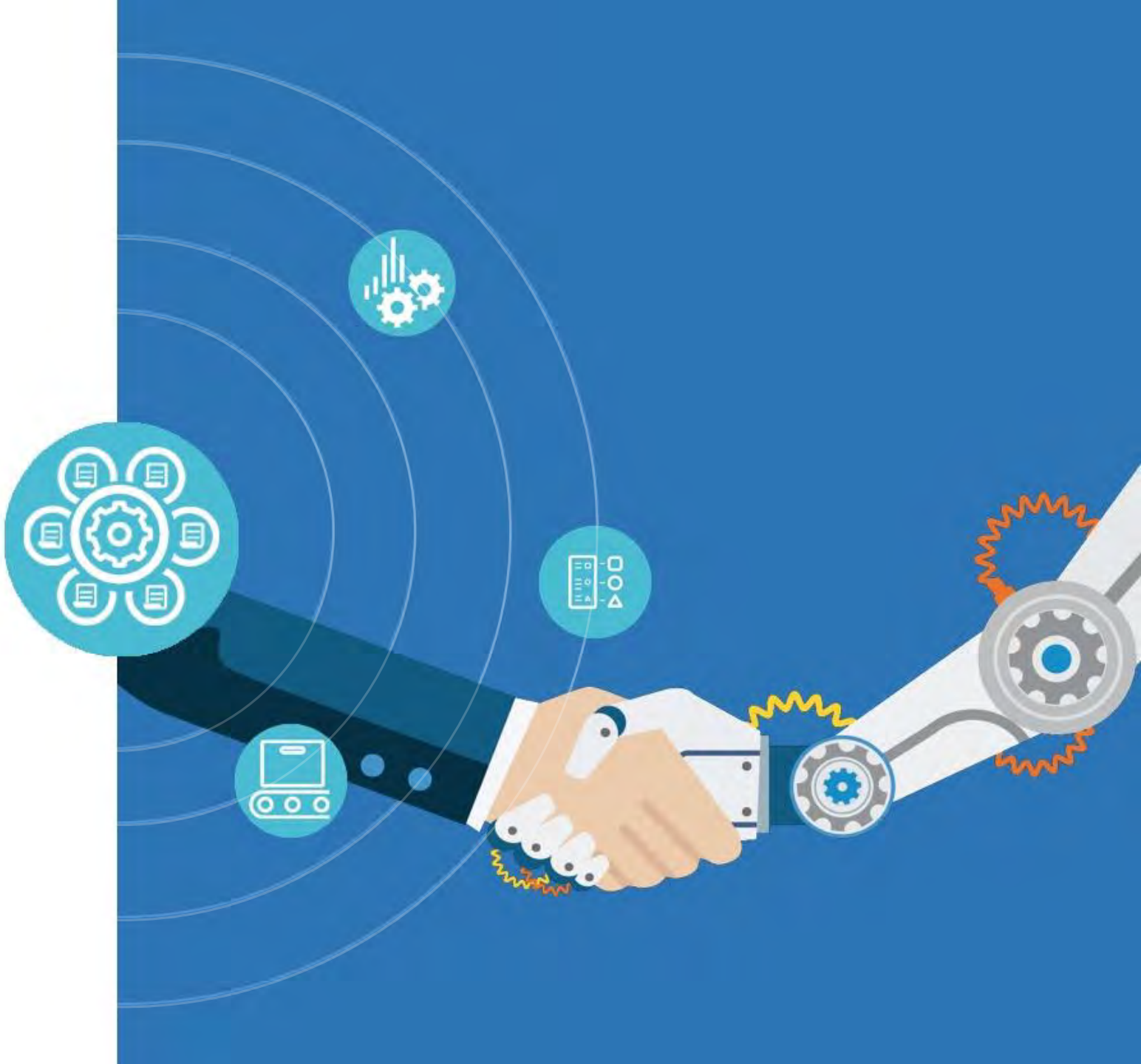
Document Automation

2019 Buyer Guide



Content

Document Processing Automation Projects

Understanding the objectives of a document automation project are difficult enough without adding the complexity of identifying and understanding vendor solutions.

When it comes to document automation, many fundamental capabilities are shared across solutions and do not provide any differentiation. Rather than include these fundamental capabilities, this buying guide focuses on key capabilities of modern document automation solutions that vary across vendors and how those differences map to the business requirements that you might have.

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 form and document automation.

Difference Between OCR & Document Automation

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

Difference between OCR & Document Automation

The most fundamental difference between OCR and data extraction solutions is the amount of additional programming that OCR software requires to provide structured output for document classification, separation and data extraction as an *unattended process*. OCR software typically performs a full-page conversion of a document image that 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) of the text on a page. 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

However, if the project scope requires locating specific information within a document (and to do so as an unattended process with no need for manual review), then additional, potentially significant development is required in order to take the literal output of the document and translate that into specific data.

In the case of an invoice, an OCR engine will output the machine-readable text of the invoice, but it does not provide data on what text the invoice number is compared to, what the invoice date is, or what the total amount is. To obtain this level of “contextual data” from the OCR output, additional development must be performed. It is the same for document classification, if that is required, with additional development required to use text as a way to identify documents.

Difference Between OCR and Document Automation

Comparing full-page OCR technology with forms and document automation.

Programming Requirements

OCR requires extensive additional programming or manual handling to achieve the level of same data results as document processing automation. For example, after using OCR ,it requires programming or manual activities to identify the invoice number from the invoice date or total amount and add it to the business system.

This functionality is available out-of-the-box in document automation solutions.

Document Automation

Document automation solutions focus on taking the OCR output and converting it into meaningful information whether it is classifying or locating and extracting specific data on a page.

Often these solutions provide all of the necessary technology and algorithms to process OCR output without the need for additional development.

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

Data Reliability

While OCR solutions—with additional development—can provide structured data, they output data at a character-level a word-level, instead of at a document-level or data-field leve

This means that they lack the abil to determine accurate data fron data with errors, forcing organizations to review all outpu instead of only what is erroneou

True document automation car statically measure and guarantee specific accuracy levels and determine what data is needed for additional review.

Workflows

No technology automation is 100% perfect leading to the need to have an in-place process in order to manage the exceptions.

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

Document automation solutions come fully-equipped with the necessary workflows and software to safely shepherd an entire document automation process from input to final perfected output.

Image Processing

Image Processing Capabilities

+ Detect DPI
+ Detect and Rescale
+ Remove Field-level Form Structure
+ Remove Field-level Pre-printed Text
+ Reshape Distorted Images
+ Remove backgrounds/watermarks

Image processing involves all the steps necessary to prepare an image of a document to be efficiently classified, separated and the data extracted. The quality of the image is fundamental to high-performing document automation. There are often many problems encountered in production. Almost every system offers basic processing to handle the most common issues such as documents scanned at an angle or upside down, removing borders around documents, and removing any noise (e.g., specks or streaks) that are introduced by a scanner, fax machine or other device.

Modern systems include this and much more to account for today's multi-channel document processing needs that include fax, portable scanner and now smart phones. These systems can deal more dynamically with a variety of image quality issues and can also make high quality scans even better.

Capabilities include auto-detection of DPI and scale, ability to re-scale images, ability to re-align distorted images, increase contrast, remove watermarks or other background imagery that can affect OCR. Some systems even provide the ability to remove the background form structure at a field level.

Comparing features for image processing will not allow you to truly determine effectiveness. The goal of image processing is to get the image to a condition that will optimize classification, separation, and data extraction. Some software cannot handle wide quality variances, so be sure to select examples of poor-quality documents and compare the system's ability to rectify the problems.

Document Identification, Separation and Classification

Identification and Separation

It is common for companies to need to process a variety of documents within a business process. The ability for the capture system to recognize the document type and when one document begins and ends are all essential. While use of document separators, either using pages or barcodes on the first page, are commonplace, modern solutions offer capabilities that completely obviate the need to perform what is called “batch preparation” before scanning documents.

Automated Classification Technologies

Use of automated classification technologies is the current state of the art. These technologies eliminate the need to do batch prep or create manual document identification rules by using machine learning techniques.

Automated classification allows a business analyst to create document types simply by submitting samples of a given document class to the system. The software then analyzes the document in order to identify the key characteristics important for determining the type during production.

When multiple documents are scanned at a time, they must be separated. Modern solutions create individual documents automatically from a stream of scanned images.

These solutions allow a person to simply import a batch of documents without having to sort them, eliminating the need to manually sort and insert separator pages. Automated separation replaces manually created rules that define page boundaries, the system also notes attributes that define specific pages such as the first page and/or the last page. All other pages will be treated as those in-between.

System Precision

How to compare one system to another

While organizations would benefit from the selection of solutions that provide automated, machine-learning based document classification and separation, the only *accurate way* to compare one system from another is to compare them based on the quality of the output in terms of the number of documents within a sample that can be classified and separated *correctly*.

To do this, you 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.

Data Extraction & System Configuration

Data Extraction

Data extraction is the process of locating specific field (or data element) data from a form or other document.

Traditionally, field location was based upon X/Y coordinates, but document automation has evolved to supply a variety of location and extraction techniques including native support for electronic documents.

These new techniques are critical if your documents vary in layout (placement of key data) or vary in terms of scale caused by differing resolutions from faxes or scanners.

Using a traditional approach requires a significant number of different configurations to handle these variances.

Having a variety of location and extraction capacities ensures that you can maximize the effectiveness of your implementation without increasing labor costs.

System Configuration

When it comes to actual configuration of systems, solutions traditionally required a lot of manual rules creation and coding using text parsing algorithms. Modern systems have simplified configuration to the extent that a non-programmer can create full-featured classification and data extraction workflows. This is important if you do not have staff with programming skills either for initial configuration or for making changes to the system once it is in production.

Systems that make it easier to set-up classification and extraction through a reduction of the required steps (or through providing the ability to move back-and-forth between rules) reduces time by allowing a user to view and understand the entire project. Simpler configuration is not only a major factor that reduces both project costs and risks associated with improper configuration, but also enables organizations to evaluate and understand the performance of the system with less effort and cost.

Data Extraction & System Configuration

How to compare one system to another

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.

In order to compare one system with another, evaluation of the amount of output is insufficient. Evaluating the amount of accurate output is also insufficient.

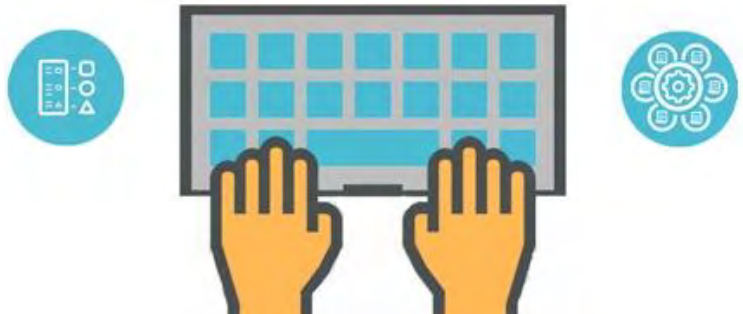
If your organization wishes to achieve optimal straight through processing, you will need to evaluate the amount of accurate answers along with the reliability of confidence scores for both accurate and erroneous data.

Capture System

Classification and Extraction Configuration

Use of a simple-to-use User Interfaces (UIs) are a key enabler of efficient and low-risk configuration. Systems that support a full configuration across different document types (e.g., structured forms vs. unstructured documents) using a single application are the easiest to use. This is because a single user interface reduces the complexity involved in configuring and supporting different classification and data extraction rules.

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



Configuring automation rules are only part of the process. You also have to understand how those rules perform on production-level documents.

Solutions that allow a user to test the results of document classification and data extraction without running a full capture workflow make it much easier to iterate through adjustments to maximize system accuracy. These solutions typically allow testing results at the configuration stage rather than requiring the user to completely configure the entire document automation workflow.

More recently, systems have added the ability to automate rules creation. Machine learning involves the ability to take user feedback during production (sometimes a data entry operator or actual machine learning) and use that to improve data extraction or classification.

Offline learning allows a business analyst to teach a system how to classify documents or find and extract data without programming.

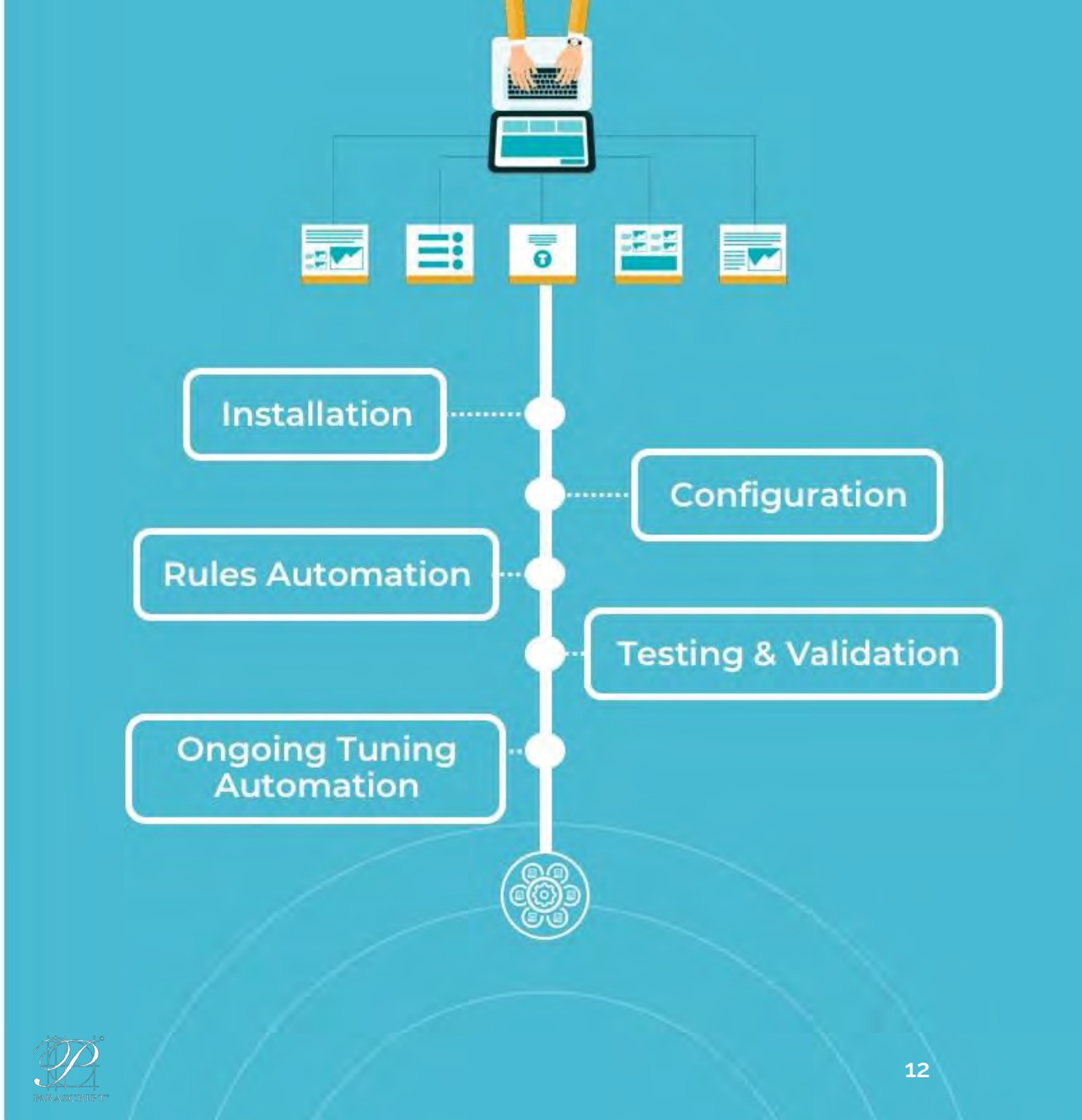
Document Automation
Capture System Deployment

Document automation is a complex process often involving many different steps and underlying technologies. Vendors approach the overall solution packaging in a myriad of ways from offering a modular approach that involves installing and configuring many different applications from databases to workflow engines. Other vendors adopt a more simplified single application strategy that involves few ”moving parts.”

Many systems use third-party applications to present a full solution. These applications might be included with the software at no additional cost or as 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 document automation, 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 & 3rd Party Integrations

Integration

Many 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 and the ability to deploy specific capabilities as a platform instead of requiring a stand-alone implementation.

This is important if your organization already has invested in business process management/workflow software or uses line-of-business systems.

Integrating classification and extraction capabilities into these applications allows for minimal disruption to user processes while making them 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 significant impact to successful business processes.

While many solutions today provide information regarding number of documents processed by time periods and process status, modern solutions collect data and display information associated with underlying aspects that affect throughput and accuracy such as individual staff efficiency, ability to redirect work to available staff or staff that have higher throughput and even reprioritize overall work.

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

3rd Party Integrations

No document automation software lives in its own world. It always supports other processes, and these other areas include software.

Whether it is an accounts payable system, a CRM or ERP system, there is a need to get documents and data into these other areas.

Having pre-built integrations means less reliance on IT or systems developers to create requirements and these integrations themselves, less time maintaining them, and quicker implementation.

Document Automation Capture Systems

The Proof is in the Pudding

When determining the capture solution that best fits your organization's needs, the variety of available options and new terminology can be overwhelming. This comprehensive set of criteria—summarized below—takes us beyond the typical capabilities that all advanced capture systems offer and will help you select the system that is right for your organization.

Checklist

IMAGE PROCESSING CAPABILITY

- ☐ Detect DPI
- ☐ Detect and Rescale
- ☐ Field-level Form Structure Removal
- ☐ Field-level removal of pre-printed text
- ☐ Reshape distorted images
- ☐ Remove backgrounds/watermarks

CLASSIFICATION TYPE

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

ADVANCED SEPARATOR TYPE

- ☐ Rules-based document separation
- ☐ Automated First-page Separation Identifier
- ☐ Automated Last-Page Separation Identifier

DATA EXTRACTION FILES

- ☐ Native support for electronic documents (no OCR required): PDF, RTF/Word, Spreadsheet, Presentation, Email and HTML
- ☐ Support for full-range of document types (structured/semi/unstructured)
- ☐ Support for full-range of data types including unconstrained handwriting
- ☐ Ability to handle complex table data

CONFIGURATION CAPABILITY

- ☐ No-code classification and extraction configuration
- ☐ Ability to test configurations 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 and configuration
- ☐ No third-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 the 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

- ☐ CRM
- ☐ Document Management
- ☐ ERP
- ☐ AP

What's Next?

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

The next step is to determine the pricing model that best fits your organization. And then, if you are still unsure of the available technologies, you can always seek expert guidance.

Document Automation

2019 Buyer Guide

Your solutions power businesses.
Our technology delivers the data.

PARASCRIPT®

www.parascript.com
info@parascript.com
888.225.0169