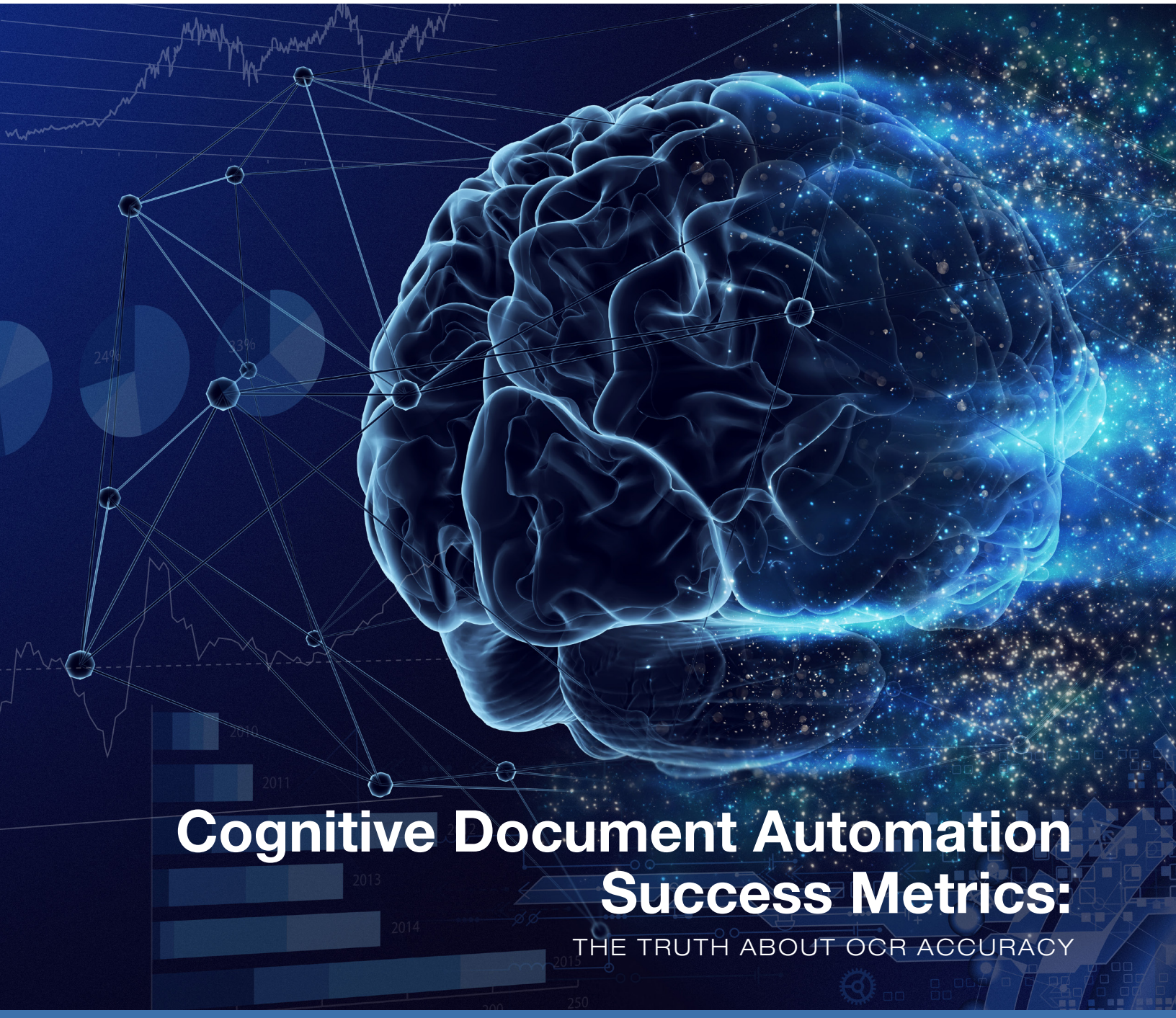
WHITE PAPER



Cognitive Document Automation Success Metrics:

THE TRUTH ABOUT OCR ACCURACY

KOFAX

TABLE OF CONTENTS

Executive Summary

Metrics matter. Today's business climate requires organizations to justify any technology investment—in time, effort and, of course, money. An organization must employ the right best practices and success metrics to maximize their investments. This paper discusses those best practices and success metrics in the context of Cognitive Document Automation (CDA) and how they help determine the return on investment and payback period for a CDA solution.

CDA Metrics - Hard Truths

The objective of any CDA project is to realize the expected benefits of greater visibility, lower costs, faster processes, fewer errors and improved customer engagement. The question is: how do we measure a CDA project's success?

There are different ways to evaluate the success of a CDA deployment; some are effective and some, not so much. Let's review a few CDA metrics truths before we land on a single success measure for CDA.

A good CDA metric gives insight into the concept of correcting automation exceptions. Unfortunately, traditional metrics fail to accomplish this. Below are examples of traditional metrics that have been used by organizations and explanations of why they are generally unhelpful in determining the value of the solution.

CHARACTER OR FIELD LEVEL CONFIDENCE

Take for example a CDA solution that can perform document classification and data extraction on a wide range of document types. The result of these automated tasks is a confidence score. This confidence scoring metric is a numeric value assigned to the document or field by the underlying OCR (optical character recognition) engine(s) on how confident it "thinks" that value is accurate; this is expressed in a range from 1-100, with 100 being the highest confidence score. Unfortunately, this measure is not of significant value because OCR engines score themselves differently, using different algorithms, so it's difficult to know what confidence level is acceptable.

CHARACTER LEVEL ACCURACY

This metric is usually found in data entry solutions where keystrokes-per-hour are measured. This is also a poor metric in CDA because there are different costs to different types of fields. A date field with an OCR error is usually detected when the date is invalid or is too far in the future or past. An amount field with OCR errors is detected because A+B<>C, and in some of these cases, numerical values can be automatically repaired. An error in a field that is less important may be

minor. But an OCR error in fields such as account number or social security number is extremely critical. In addition, a field could have multiple incorrect characters (making correction more time consuming) or just one character incorrect (quickly correctable), so character level accuracy does not help determine how quickly a field can be processed. A character level accuracy of 94% gives us no information about the human effort required to correct the other 6% of errors.

FIELD LEVEL ACCURACY

This measure is an improvement over character level accuracy; for example, social security number accuracy is considered more significant than a description field's accuracy. But there is still the problem of how much effort is required to fix the errors in those fields. Organizations have no insight into how many errors are in a field—often it is one character, sometimes two, and perhaps (rarely) three. Correct or incorrect doesn't tell us the full story, as some fields can be confident when they shouldn't be, while others are not confident when they should be. Field accuracy is part of the metrics equation, but not all of it.

PRECISION AND RECALL

This is a statistical metric that measures relevance. In other words, these measurements evaluate the correct results against a known set of answers. This metric suffers the same challenges as field level accuracy: it does not tell us enough about the problems and the amount of effort required to fix them. It also does not place a high significance on some types of exceptions over others—it simply ignores them.

Now that we've examined how many of the traditional metrics employed by organizations fall short, let's take a look at the key metric that can prove most useful for a CDA solution.

The CDA Success Metric That Truly Delivers

THE USER PRODUCTIVITY METRIC

The goal of a CDA project is to achieve maximum acceptable data accuracy, with the least amount of human intervention, to optimize end-to-end processing productivity. Therefore, it stands to

User productivity = accuracy + user efficiency

reason that productivity is the most useful metric for a CDA solution. Measuring CDA success is ultimately about improving human productivity by reducing effort in document processing and data entry tasks. Even solutions with a very strong emphasis and strict requirements on data accuracy are ultimately projects about productivity because they still need to reduce the human effort required to correct inaccurate data. Making human productivity the number one goal of a CDA project naturally leads to solutions that contribute to efficient workplaces where employees can more easily produce accurate, high-quality data.

In essence, productivity is about how many documents a person can process in an hour/day/week/month with high-quality results. Here, the synergistic feedback efforts of a computer and a human are calculated together. This metric is easily convertible to a measurable ROI and payback analysis for the CDA solution.

User productivity is made up of two components:

- OCR accuracy
- User efficiency

OCR ACCURACY

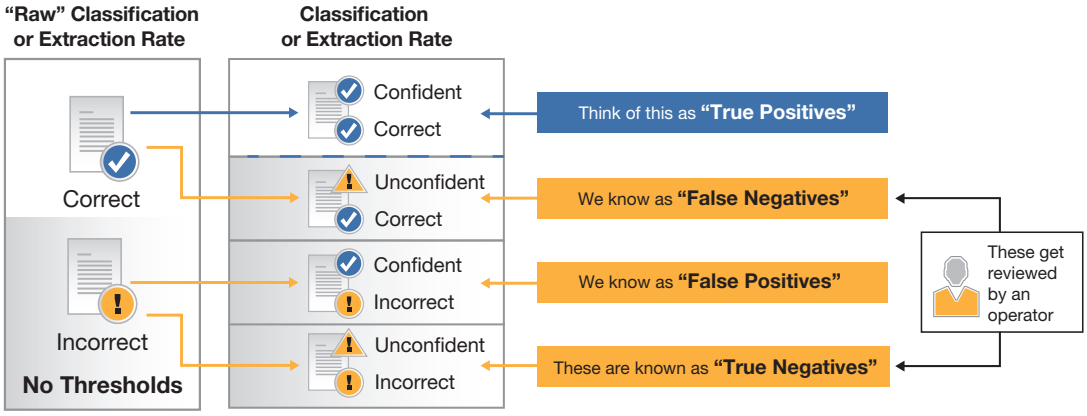
Perhaps the most common question after watching a CDA technology demonstration is: What level of OCR accuracy can I expect to achieve?

The short answer is that it depends and varies widely across use cases. OCR accuracy, and more generally classification and extraction accuracy, depends on multiple factors, including:

- Scanner hardware
- Scan resolution
- Image quality
- Document type (form, invoice, letter, etc.)
- Machine-printed/hand-printed/cursive
- Document language
- Font type and character spacing
- Field boxes/shadings
- Ability to database match or check for checksums and other rules

Why is accuracy important? The higher the accuracy, the more classification and extraction automation; the lower the accuracy, the less automation and more manual labor.

Because accuracy varies so much, it is best to perform benchmark testing for classification and extraction accuracy rates on the business' actual real-world samples. Use these results to optimize project settings for each document type and field, and thereby increase accuracy and automation. For each document type and field, benchmark testing should record both confidence (yes/no) and correctness (yes/no), with the goal to maximize true positives, minimize false negatives and true negatives, and completely eradicate any false positives from being exported to downstream people, processes and systems.



Benchmark testing goes beyond if a field is correct or incorrect.

A related term, "straight through processing" (STP), is also used as a metric to describe CDA results. This is the measure of the percentage of documents run through the CDA "acquire, understand, integrate" process untouched by a human. The STP rate will never be higher than the lowest OCR accuracy field on the document. To maximize the STP rate, focus on the lowest-accuracy fields being extracted on the document and adjust settings for those fields.

USER EFFICIENCY

OCR accuracy is just one side of the coin of user productivity. The other side is user efficiency for exceptions. Documents and fields that don't pass through untouched (known as "low-confidence") must be reviewed by a human to ensure they are classified and extracted correctly. User efficiency is all about how quickly a user can review a low-confidence document or field, make a decision on what needs to be corrected/confirmed and then execute that decision. So the human validation interface must be designed for the most efficient use of eyes and hands during the document classification review and data validation process.

Here are some examples of user-efficiency features in leading CDA solutions:

- Jumping to the field that needs to be validated, skipping over confident fields
- Highlighting that field on the actual image for context
- Displaying an image snippet of the field in question next to the data entry area
- Custom positioning of panels to each user's liking
- Correcting a single character in the field rather than re-entering the entire field
- Hitting a hotkey to call a database lookup for a field
- Auto-complete of the field based on the document type list or full page OCR
- Completing a table's worth of data by simply highlighting the first row

The effort spent on user efficiency and user experience will produce ten times the user productivity compared to the same effort spent on improving field OCR accuracy. That is why it is best to maximize a human's work speed processing these data exceptions via effective user engagement and minimal keystrokes and mouse movements.

MAXIMIZING USER PRODUCTIVITY

As previously mentioned, user productivity—the combination of OCR accuracy and user efficiency—represents the single most important metric for a CDA project's success.

For example, consider a mortgage application form. Some form fields will be more important than others, so an "acceptable level of data quality" will vary depending on the field. Benchmarking the CDA project to understand per-field OCR accuracy is necessary to optimize extraction rates for high-priority fields such as social security number and annual income.

When configured effectively based on the success metric of maximizing user productivity, CDA solutions will deliver an attractive ROI and payback period of 6-18 months from system launch.

Common Business Case Metrics

The most common business case metrics used to justify a CDA investment are return on investment (ROI) and payback period. The use of these two metrics will influence and, in most cases, provide the necessary data required to help justify the investment within an organization. Below is how these two metrics are derived, based on productivity gains/savings provided by a best-in-class CDA solution.

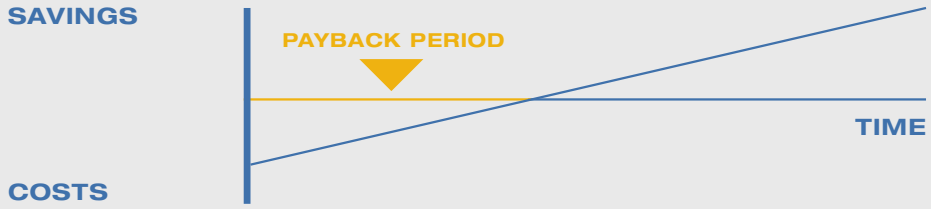
RETURN ON INVESTMENT

ROI is the ratio of the average savings (of the first three years) to the initial software cost.

$$\frac{(\text{SAVINGS YEAR 1} + \text{YEAR 2} + \text{YEAR 3}) / 3}{\text{INITIAL COST}} \times 100$$

PAYBACK PERIOD

A solution's payback period is derived by calculating the time required for net savings to return the initial cost, typically expressed in months.

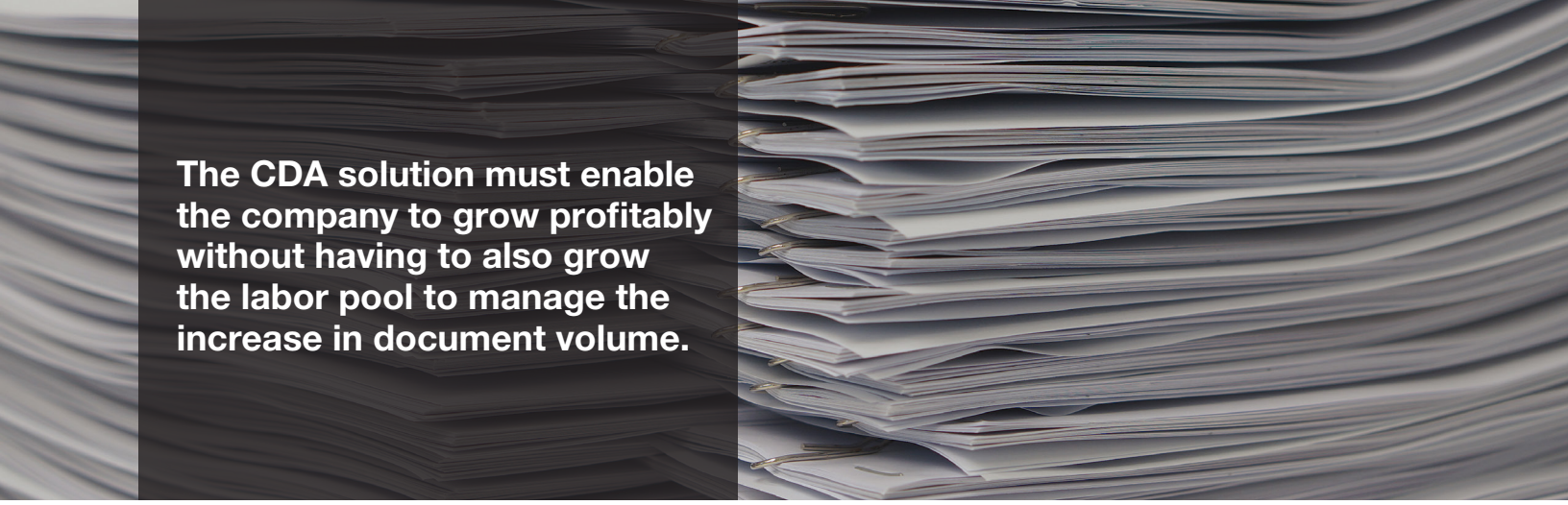


Quantifying the Before and After Picture

To get to ROI and payback period, we first need to analyze the current manual operation (the "before" picture) and the planned automated operation (the "after" picture).

CURRENT OPERATIONAL METRICS

First let's look at an organization's current **operating environment and associated metrics**. In this example, the organization processes 66,700 documents per work day or 16.7M documents per year. Each document averages six pages, for a total of 100 million pages annually. Each document has an average of ten fields that are required to be indexed. There are 85 people indexing these documents with an average salary of $20 per hour. These data points provide the insight required to begin developing a business case.



This is reflected in the table below:

Document Indexing	
Number of document indexing associates	85
Hourly rate	$20.00
% of time dedicated to indexing	100%
Total cost for indexing tasks	**$3,400,000**
Number of documents processed per hour	9,526
Number of documents processed per hour per person	112
Current Productivity Metric (avg in seconds)	32
Estimated cost to index each document	$0.20

With this information, we are able to determine:

1. How much it costs to process each document in the current solution, and
2. The average number of documents processed per employee for the current solution

These two data points are the focus for the ROI and payback evaluation.

VALUE THROUGH AUTOMATION

The goal for most CDA solutions revolves around enabling the document processing operation to become more scalable—effectively handling additional document volume without having to scale up human resources. The CDA solution must enable the company to grow profitably without having to also grow the labor pool to manage the increase in document volume.

In order to calculate productivity post-automation, we must first look at the areas that are impacted by the CDA solution. In this case, we're focused on automating the document identification process and the extraction of business-critical data. As a result of applying CDA automation technologies, exceptions may occur. The CDA solution will provide interfaces for confirming the results of the automation technologies or correcting the exceptions.

EXCEPTION MANAGEMENT FOR AUTOMATED CLASSIFICATION AND SEPARATION

Using the same business example as above, we can **estimate the effort required to manage the exceptions** by estimating how many documents require a human to review and how long it takes for a human to confirm or correct the results.

Classification Metrics	
Estimated classification / separation automation %	85%
Estimated classification effort (in seconds) per document (avg)	5
# of documents to classify or re-classify per day	10,002
# of hours required to classify documents	13.89
Staff required for document classification	**1.98**

In this case, we estimate the CDA solution can automatically identify the documents 85% of the time. Because the CDA solution is designed to only present exceptions to the operations staff, only 15% of the documents require human review. Many of these documents simply fell below a confidence threshold, which requires very little human intervention. For this calculation, we're estimating the average time it takes for a human to simply confirm the results or make a change to the document type designation through an easy-to-use interface. The result is the person hours required to manage this activity over the course of a typical work day, which equates to about two staff workers.

EXCEPTION MANAGEMENT FOR AUTOMATED DATA EXTRACTION

Similar to the classification and separation staffing calculation discussed above, the data extraction portion of the solution also has an exception process that may require human intervention. Because of this, we also need to estimate the person hours required to manage these exceptions. We do this by applying an estimated automation rate (above confidence threshold) for the fields that are extracted. For those fields that are below the configured confidence threshold, we estimate the effort involved in correcting those fields.

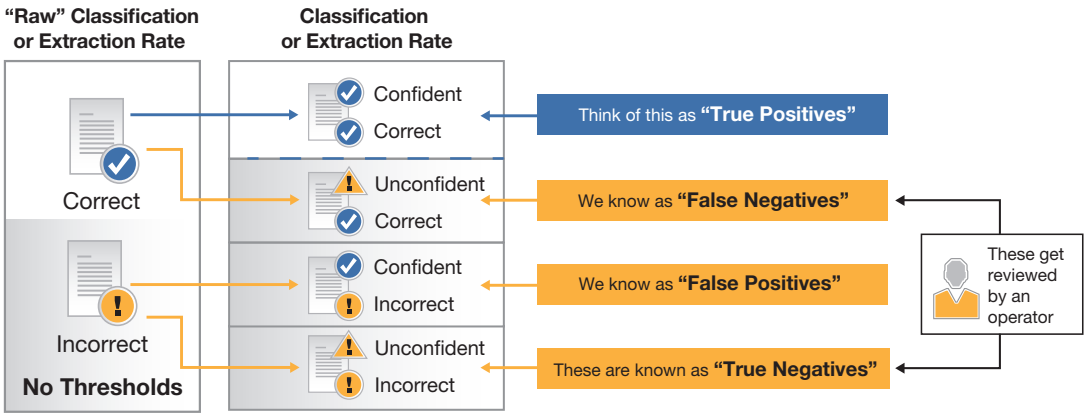
Again, this task often requires a human to simply press the "enter" key. In other cases, the operator is required to enter the correct data into the appropriate field. In either case, the user is drawn to those exceptions automatically by the use of color-coded descriptions to improve application usability and, ultimately, operator productivity.

Using these estimations we can derive the staff required for managing the data extraction exceptions.

Extraction Metrics	
Estimated extraction automation %	75%
Average # of fields extracted per document	10
Estimated effort required to validate extracted fields (seconds)	3
Total number of fields extracted per day	666,800
# of fields falling below threshold	166,700
Total time required per day for validation (seconds)	500,100
# of hours required to validate extracted fields (per day)	139
Productivity measurement (avg in seconds)	7.5
Staff required for field validation	**26.48**

The result of these calculations provide the **estimated productivity** and the person hours required to support the volume of documents given the amount of automation that can be applied at a field level. As you can see in this example, there is an average of 10 fields per document, and by applying data extraction automation technologies, the operators would only be reviewing 25% of the fields. This equates to 139 total person hours in an average day. Given an estimated 7 hours of actual productive time (effective rate) per employee, we can estimate the staff required to manage these exceptions at 27 people.

As mentioned earlier, the automation rates used in these examples are based on a more accurate representation than simply "correct" and "incorrect." Many organizations believe accuracy is black and white or right or wrong. While that remains true, **a successful CDA solution will take these metrics a step further, measuring accuracy by taking the "confidently correct results" (true positives) divided by the total number of results.**



Benchmark testing goes beyond if a field is correct or incorrect.

Creating the CDA Business Case

Now that we've calculated the staff required by applying CDA automation to the existing manual document process, we can compare the current state approach against the future state CDA solution in order to determine the business ROI and payback period of the CDA solution.

PRODUCTIVITY MEASUREMENTS

The current manual process for indexing documents results in a productivity output of 112 documents per hour per employee (volume/days a year/number of people/productive hours a day) across 85 employees. By applying the above automation rates, the organization is able to process the same 66,700 documents per day with only 29 people (2 for classification exceptions and 27 for extraction exceptions); this produces savings of $2.24M annually or 65%.

Viewed differently, the same 85 people can go from processing 112 documents per hour per person to processing 329 documents per hour per person—a 194% productivity improvement. Utilizing CDA automation, growing businesses can forego hiring new employees and process more documents with their existing staff.

This same productivity measurement can be applied to the cost of processing a single document. The annual cost of 85 employees is $3.4M, and the annual document volume that those employees can process has gone from 16.7M documents per year to 48.9M documents per year—almost a threefold increase. The manual process costs $0.20 to identify and index a document. With the improved productivity, the cost to process a document is reduced down to $.07 per document—a 65% cost improvement.

ROI AND PAYBACK PERIOD

As mentioned at the beginning of this paper, ROI and payback period are the two metrics that most organizations use to justify an investment in time, resources and money toward a CDA project. These business metrics are based on the cost savings that the CDA solution delivers.

ROI is represented as a percentage and is typically based on the average savings over a three-year period divided by the cost of the CDA software licenses, maintenance and services. (Note: Some organizations look at ROI over a five-year period.)

A sample ROI for a CDA solution as described above would look like the following:

$$\frac{\text{(YEAR 1 + YEAR 2 + YEAR 3) / 3}}{\text{INITIAL COST}} \times 100 = 84\%$$

The payback period for a solution similar to this would simply be the software costs divided by annual savings times 12 (expressed in months). For example, if $100,000 of software costs gives an annual savings of $50,000, then the payback period is 24 months. In the illustration below, that same $100,000 of software gave an annual savings of $100,000, or a 12 month payback period.

$$\frac{\text{INITIAL INVESTMENT}}{\text{FIRST YEAR SAVINGS}} \times 12 = 12 \text{ MONTHS}$$

Summary

In summary, the **value of a CDA solution is derived by comparing the cost and labor required to manually manage a document-centric process against what tasks can be eliminated** (or substantially minimized) from the operation. This is done by measuring the productivity gains that can be achieved by applying CDA automation capabilities. The payback period for a CDA solution is generally less than 18 months, and often much less.

While “OCR accuracy” is an input to the calculation, OCR accuracy alone is not the measurement that drives a solid business case for CDA. The ultimate goal is for the operation to be more efficient and productive for improved growth and greater profitability. User productivity, defined as OCR field accuracy + user efficiency, is most effectively gained by improving the way in which the user interacts with the system.

To learn more about measuring the success of your cognitive document automation project, contact Kofax today.

kofax.com