



What is IDP?

The journey from OCR & beyond



What is Intelligent Document Processing?	3
By the Numbers	3
IDP and PDF Documents	4
IDP Use Cases	4
Benefits of Data Extraction	5
Why IDP? The journey from OCR to IDP	6
OCR vs IDP Comparison Chart	7
Key Challenges in PDF Data Extraction	8
How IDP Works	9
Apryse IDP Solutions	10
OCR	10
Form Extraction	10
Barcode	10
Table Extraction	10
Document Structure Recognition	10
Why Apryse IDP Stands Out	11
Data Extraction Buyer's Guide	12
Structured Documents	12
Semi-Structured Documents	13
Unstructured Documents	14
Alternative Solution: Template Extraction	15
Supplier Considerations	16
Real World Use Case Examples	17
Improving Healthcare Software Development with IDP Data Extraction	17
Enhancing AI Model Training with Apryse IDP Data Extraction	18
Next Steps	19

What is Intelligent Document Processing?

Almost 20 years ago, Mathematician Clive Humby coined the phrase,

“data is the new oil.”

Data is valuable, but it must be refined to maximize utility. Fast forward to today, and the utility of data has exploded with the advent of Machine Learning and LLMs which require structured training data, as well as business and workflow data that can be used to enhance the software tools we build and use.

However, collecting and categorizing data becomes a challenge when the data is contained in PDF documents such as forms, plans, notes and invoices, for example. Adopting solutions to unlock and harness this embedded information is essential to meet market demands and drive growth.

In software development, the ability to embed data extraction capabilities powers user functionality such as creating indexable information about documents for search, capturing document information to reduce user data entry, and categorizing documents based on user needs. In addition, data extraction powers growth and revenue generation features such as data analysis and AI model training.

Check out the handy
use cases & buyer's guide
on page 12 of this ebook!



By the Numbers

According to a study conducted by the Journal of Accountancy, human error rates in manual data entry can range from **1% to 5%**.

Incorrect or incomplete data costs businesses approximately
\$3 TRILLION ANNUALLY
in the U.S. alone. (Source: IBM)

Standard invoice processing has an average error rate of **10%**.
Error correction takes **61%** of the invoice processing cost on the labor. (Source: DocAcquire)

IDP and PDF Documents

PDFs are designed to display information to humans, not computers, and the objects under the hood aren't easily parsed by computers. So, OCR and IDP technology are useful for processing PDF documents and gleaming structured data from them. Examples of PDF documents include scans of paper documents, images of text, and hybrid formats.

IDP Use Cases



Accelerating tax
filing and audit



Digitizing records with
smart features such as
searchability & navigation



Extracting data from invoices
and bank statements



Processing
insurance claims



Extracting
data from forms



Improving fraud detection



Automating traceability



Collecting
ML training data



Assessing exams



Automating application
processes

Benefits of Data Extraction



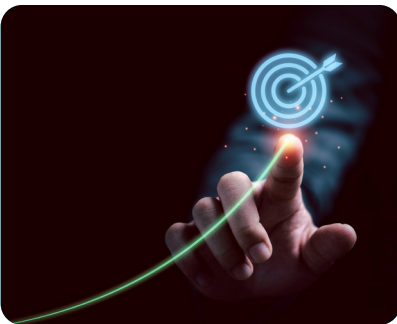
Improved User Experience

With dull data extraction tasks automated and document information prepared in a usable format such as JSON, employees are empowered to create new workflows and generate value faster. Users who submit documents also enjoy a better experience through faster, more streamlined performance, such as faster insurance claim submissions, for example.



Digital Transformation

When document data is effectively gathered and organized digitally, new digital transformation capabilities are revealed. These include new data-driven insights, continuous improvement for work processes, and automation.



Improved Accuracy and Performance

Selecting the right data extraction solution beyond OCR, provides improved performance, accuracy, and reliability while streamlining workflows and reducing manual interventions. For example, using template extraction to simplify processing of forms can be more cost effective than IDP for the same application.

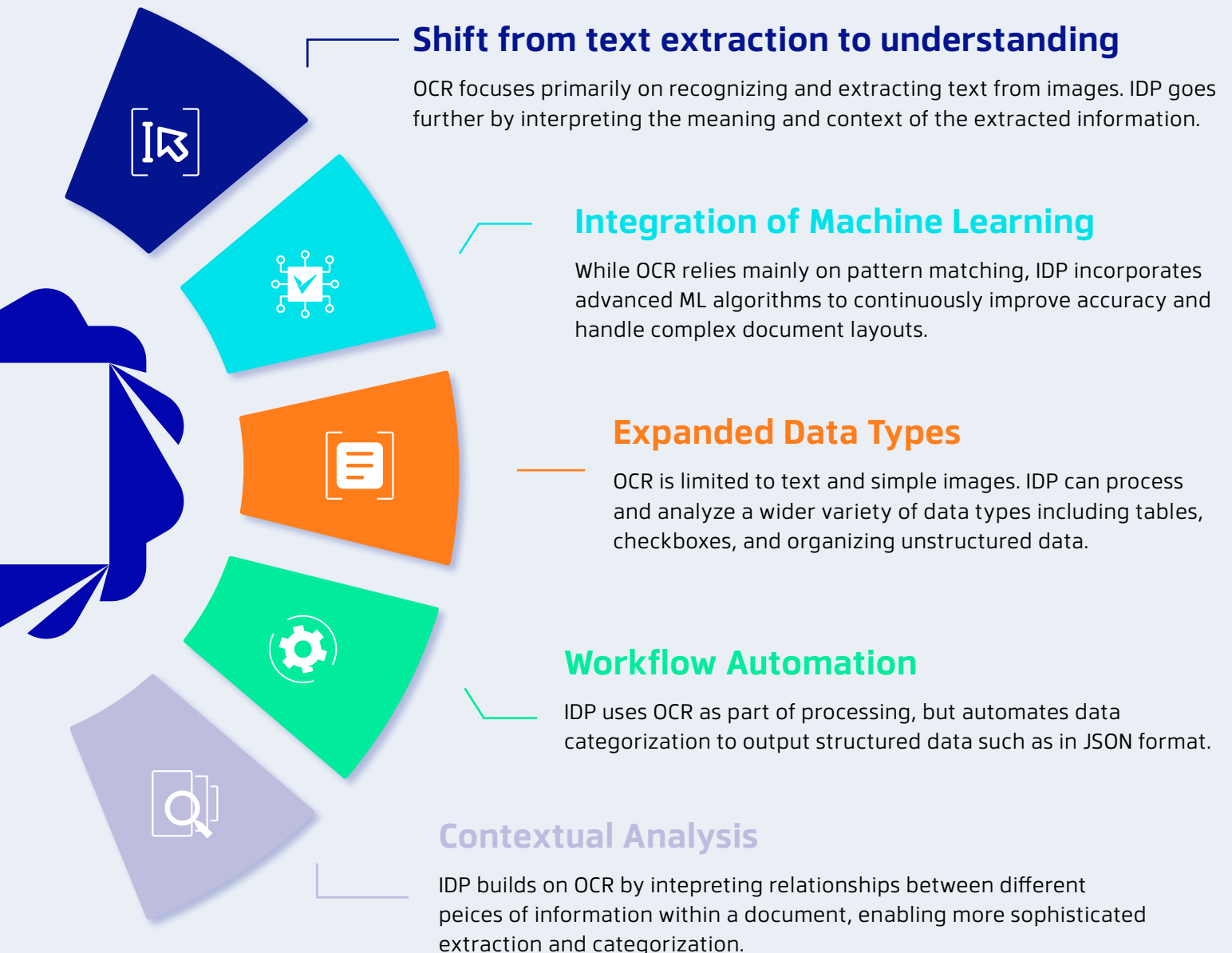


Enhanced Productivity

Like all automation tools, the benefits of data extraction grow as your business learns to integrate more use cases with the technology. Start with one document workflow, and build on your success.




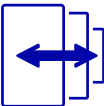

Why IDP?

The journey from OCR to IDP



Optical character recognition (OCR) is not Intelligent Document Processing (IDP), but it is a component of IDP. OCR reads and extracts all machine-printed text from images, documents, or embedded file attachments. The output can be delivered as plain text or structured data.

Comparison Chart

OCR	VS	IDP
Turns image of text into machine-readable format		Makes use of OCR, but takes that technology a step further
Used to extract data from documents		Allows identification, categorization, extraction, structuring and validation of data from documents
Based on templates that are expensive to build		Able to understand complex documents with tables, pictures and documents
Can be used for simple structured documents		IDP is template free
Might include some AI		Uses AI and Machine Learning technology

The limitation of OCR is that once OCR has converted a document to a digital text file, it still requires another solution, such as human eyes or IDP, to process the text and understand what to do with it. For example, if a legal software application needs to process a large volume of legal documents to generate a summary for quick user review, the LLM needs to understand what information is located in areas such as body text, headings, and tables to be effectively extracted for use. If only OCR is used to digitize the entire document, the reality remains that these key pieces of contextual information are not captured.

IDP powers the extraction of these key pieces of information by interpreting the meaning and context of information using technology such as AI, Machine Learning, and pattern recognition.

Key Challenges in PDF Data Extraction

PDFs are notoriously difficult to extract data from due to a variety of factors, including:



**FRAGMENTED
TEXT STORAGE**



**COMPLEX
LAYOUTS**



**DIVERSE
CREATION METHODS**

(e.g., scanned images, hybrid formats).

Organizations must overcome these obstacles to perform analytics, train AI models, automate workflows, and ensure compliance in industries with stringent privacy requirements.

To meet these needs, an effective IDP solution must:



Extract data in usable formats
for diverse workflows



Meet data privacy and
security requirements



Minimize manual input, such as
configuration and post-processing

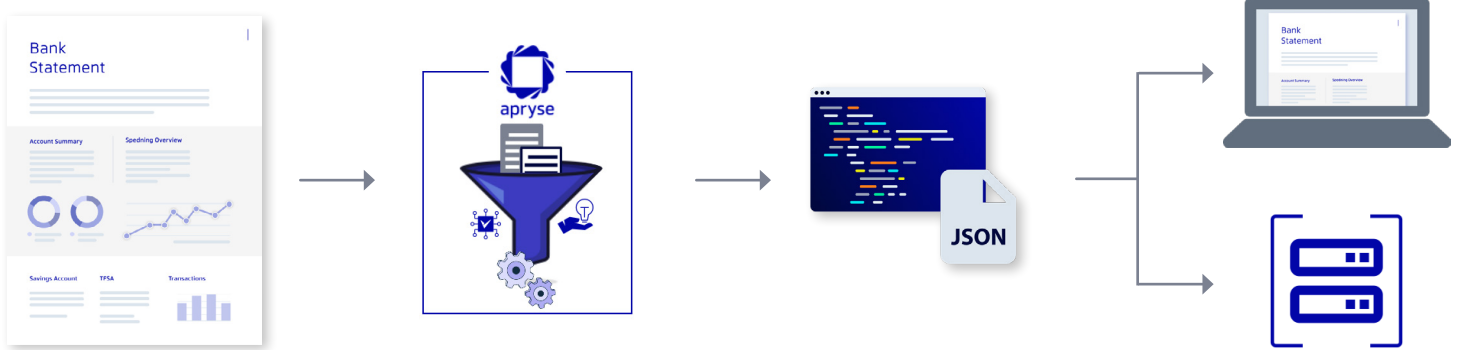


Handle diverse PDF styles and
multi-page documents

How IDP Works

The following steps are based on Apryse technology.

Other IDP solutions may follow different steps.



STEP 1: Pre-Processing

First, the IDP system performs the same pre-processing operations as OCR, including deskew, despeckling, and noise reduction. These pre-processing steps help to reduce errors in the digitized text. Apryse's pre-processing is powered by LEADTOOLS technology for greater accuracy and reliability.

STEP 2: Extraction

This step uses intelligent tools to extract relevant data from the PDF. These tools include OCR, templating, and AI to identify and extract data. It's because of this main step that the terms 'IDP' and 'Extraction' are sometimes used interchangeably.

STEP 3: JSON Output

Next, the IDP system outputs the extracted data to JSON, which is a lightweight data format allowing the user to import or connect the data to the desired application, such as a BI tool, management software, or database.

STEP 4: External Use

With the IDP process complete, the JSON data is now ready to be used in another application. For example, invoice data could be used to generate work orders, or student exam data could be used to populate grades.

VIDEO EXPLAINER

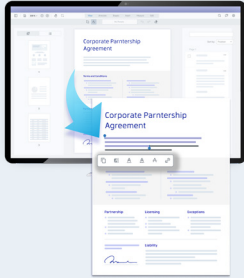
Click the link below to check out this Apryse video for a clear explanation of how Apryse IDP converts PDF content into structured JSON data.

[Watch Now](#)



Apyrse IDP Solutions

Developers and leaders looking to adopt IDP need more than just any IDP technology. It's important to source a robust solution that meets the needs and requirements of the organization, not just the extraction process. These needs include data privacy, integration, and reliability.



OCR

Apyrse OCR helps technology providers and enterprises easily add powerful text extraction to their applications. With multilingual support, seamless integration, and 8-10x faster performance than our previous OCR engine, you can efficiently automate document workflows while maintaining precision.

[Demo](#)

Form Extraction

Form Extraction uses templates to mark form fields for extraction, allowing users to programmatically fill and extract data from forms with JavaScript.

[Demo](#)

Barcode

IDP is for more than text. Our powerful barcode extraction SDK solution for technology providers and enterprises is designed to add seamless and efficient barcode reading capabilities to your applications.

[Documentation](#)

Table Extraction

This technology uses our custom built AI models to extract complex tables accurately and output the data in multiple formats.

[Demo](#)

Document Structure Recognition

In this mode of operation, the full logical structure is discovered, including paragraphs, lists, tables, headers, footers, images, graphics, like in a typical word processor. This enables more advanced IDP by automating the process of identifying content by its context on a page.

[IDP Documentation](#)

Why Apyrse IDP Stands Out



Privacy First: Fully self-hosted, on-premises SDK ensures complete control over data, ideal for regulated industries.



Developer-Friendly: Reduces workload with pre-built capabilities that integrate seamlessly into existing workflows. Available in most popular languages, simplifying implementation for developers.



Flexibility: A growing suite of extraction methods and data types supports a wide variety of use cases and customizable solutions.



Scalability: Built to process large volumes of documents with consistent performance.

With advanced OCR and support for barcodes, tables, forms, and unstructured content, Apyrse handles even the toughest document workflows, offering the tools to manage sensitive information and diverse formats with confidence and precision.

Enterprise-Grade Privacy and Security

Fully self-hosted, on-premises SDKs ensure complete control over your data, making them ideal for industries like healthcare, finance, and government.

Take Control While Limiting Cloud Spend

Extract text, tables, headers, and layouts with precision comparable to leading cloud-based tools. With Apyrse, companies already dealing with exorbitant cloud bills can limit reliance on managed services and can manage infrastructure costs

Effortless Automation of Table and Layout Extraction

Eliminate manual configuration and complex parsers with intelligent algorithms that adapt to your documents' structure.

Purpose built for handling complex PDF documents

Reliable performance in PDF extraction, multi-page documents, and even challenging edge cases like split rows or dynamic column mappings.

Reduce Developer Workload

Streamline workflows with minimal post-processing effort. Extract structured data that's ready for immediate analysis or integration into your systems. Simplify the process of extracting data from PDF format.

Scalability without compromise

Built to handle enterprise-scale workloads, Apyrse SDKs deliver consistent performance whether processing hundreds of millions of document.

Effortless Integration

Quick to implement and fully customizable, our SDKs integrate seamlessly into your existing systems and workflows.

Data Extraction Buyer's Guide

So, you're interested in procuring a data extraction solution to empower your document workflows with automation and precision. However, the different data extraction solutions on the market match best with certain use cases, and have different configuration requirements, ranging from setting up a template for each document type to complex AI model training.

Selecting the right extraction technology requires an understanding of three document types: structured, semi-structured, and unstructured documents. Selecting the right extraction technology requires an understanding of three document types:

Structured

Semi-Structured

Unstructured

Structured Documents

Documents such as forms, payment slips, time cards, passports, and surveys have a consistent layout and fixed format. Data remains in fixed locations in the document.

SOLUTIONS:

Using a combination of template-based extraction and OCR capabilities, key-value pairs and table data can be extracted based on user-defined zones and coordinates. Create a template for each provider to set up the automation.

CHALLENGES:

Creating a new template for each provider, form type and version can be labor intensive, and as documents are redesigned, the automation may break. This can lead to challenges managing a large template library. Pre-built templates in software may not anticipate user needs.

RECOMMENDED PRODUCT:

Apryse Template Extraction,

Apryse IDP SDK for some use cases.

Semi-Structured Documents

Documents such as invoices, purchase orders or bills of lading present a fixed data set, but format can vary. Across multiple invoices, the location of data points can vary, with different naming conventions and table layouts.

SOLUTIONS:

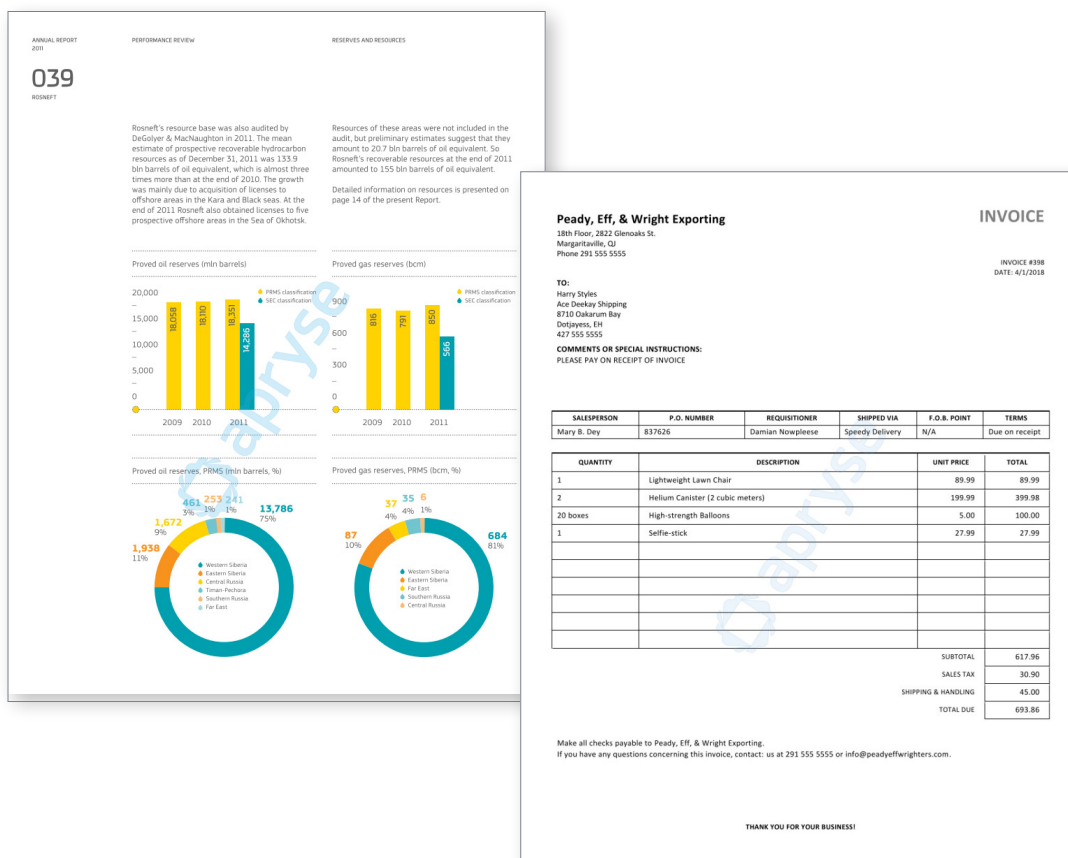
Because template-based solutions can't handle these variations, robust machine learning and natural language processing (ML and NLP) algorithms are necessary to interpret the context of each field. OCR is used to enable this processing.

CHALLENGES:

With third-party SaaS solutions, sensitive data is transmitted outside the organization to the cloud and subject to potential risks. Custom development of complex algorithms is a considerable challenge.

RECOMMENDED PRODUCT:

Apyrse IDP SDK



Unstructured Documents

Documents such as contracts, letters, articles, and memos are the most challenging documents to interpret and extract data points from. These documents are characterized by free-flowing, verbose content that can have information presented anywhere, in any format. This ebook is an example of an unstructured document, as pieces of information are presented throughout.

SOLUTIONS:

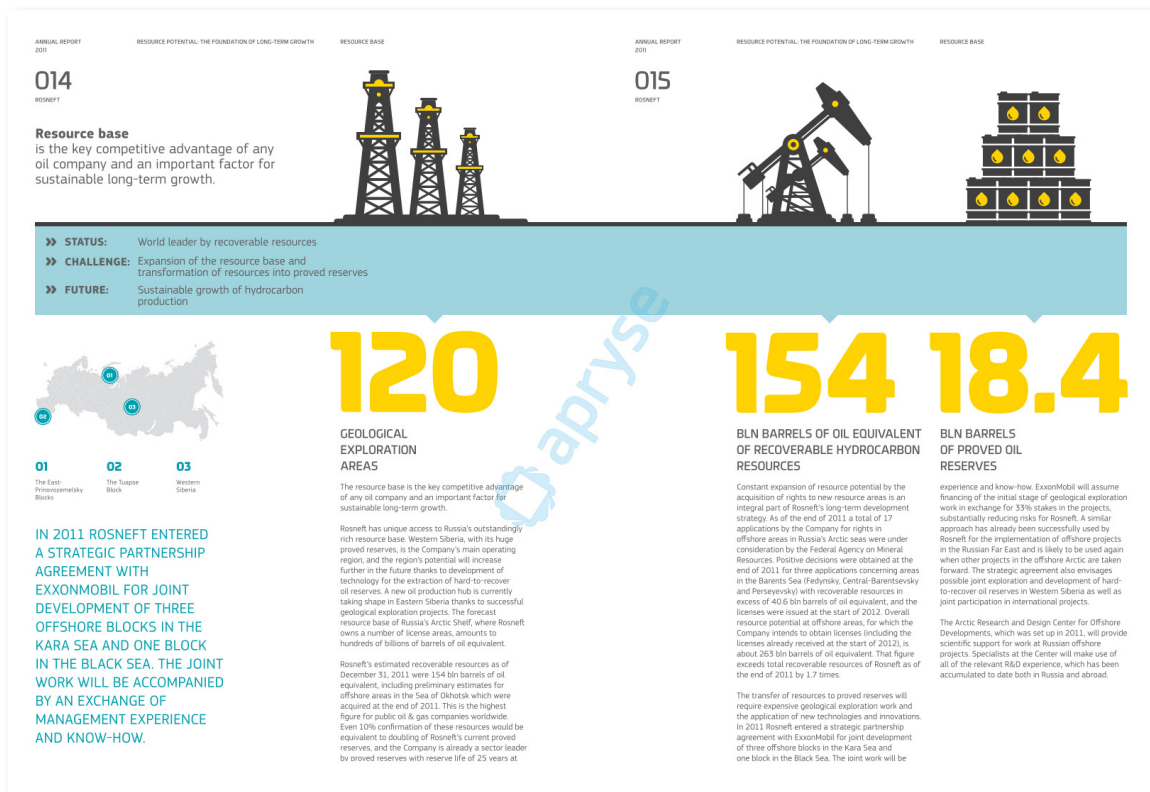
To handle these complex documents, machine learning, OCR and computer vision are required to interpret diverse components including charts, graphs, and images.

CHALLENGES:

Due to the complexity of these solutions, significant upfront investment may be required to train solutions that can effectively process unstructured documents. It may be difficult to achieve ROI on some applications of extraction.

RECOMMENDED PRODUCT:

Apryse IDP SDK



Alternative Solution: Template Extraction

While Apryse's IDP solution is AI-powered and designed to process documents without requiring predefined templates, Template Extraction takes a different approach:



Template-Driven:

Requires predefined templates for specific document types (e.g., ACORD forms, invoices), while IDP processes unstructured data without templates.



Focused Precision:

Ideal for cases where the document format is standardized, such as ACORD forms, invoices, or contracts.

This specific approach can be a cost-effective alternative to IDP for template-heavy industries.

Apryse Template Extraction automates data extraction from specific templates like ACORD forms, ensuring consistent and reliable results. High-volume, template-driven workflows benefit from this solution, which allows creation of custom template libraries for structured data extraction. Designed exclusively for Windows, it outputs data in JSON format, integrates with Java and .NET, and leverages advanced preprocessing capabilities from LeadTools to enhance accuracy and efficiency.

The screenshot displays the Apryse Template Extraction software interface. On the left, a sidebar shows a list of templates, with 'FinCEN Form 107' selected. The main window shows a preview of the document, which is a 'Registration of Money Services Business' form. Below the preview, a table displays the extracted data. The table has columns for 'Field', 'Type', 'Value', and 'Confidence'. The data is organized into sections corresponding to the form's parts: Part I (Filing Information), Part II (Registrant Information), Part III (Owner or Controlling Person), and Part IV (Money Services and Product Information). The bottom of the interface shows a list of states and territories, with checkboxes for each, indicating the selection process for the document's jurisdiction.

Field	Type	Value	Confidence
Form Number	Text	107	100
Form Title	Text	Registration of Money Services Business	100
Form Date	Text	01/01/2000	100
Form State	Text	CA	100
Form Title	Text	Registration of Money Services Business	100
Form Date	Text	01/01/2000	100
Form State	Text	CA	100

Apryse Template Extraction user interface

Supplier Considerations

In addition to selecting an SDK based on features and functionality, it's important to select a vendor that supports the product. With a true partnership, your project will be supported for success. Consider:



Support

Ensure that your selected vendor provides help resources, such as comprehensive documentation, code samples, and product demos, as well as live customer support.



Security Certifications

Ensure that your chosen vendor meets applicable security standards, such as SOC2, to reduce vulnerabilities.



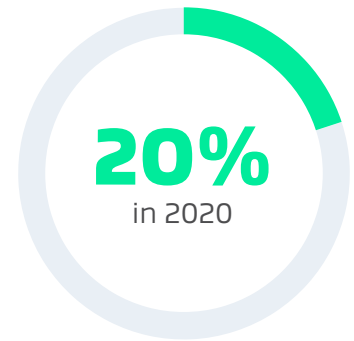
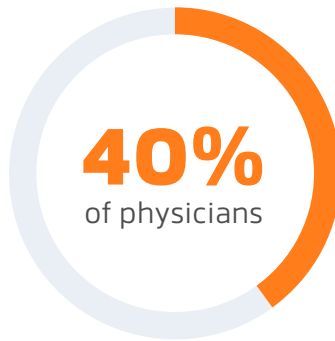
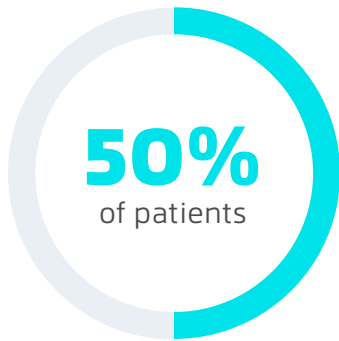
Other Products

If you plan to expand capabilities in future to include other document processing SDKs beyond data extraction, being able to stay aligned with one vendor can help save time and ease integration and compatibility challenges.

Real World Use Case Examples

Improving Healthcare Software Development with IDP Data Extraction

Telehealth solutions such as mobile apps are on the rise, with nearly 50% of patients in 2023 report using mobile healthcare apps, and 40% of physicians reporting 'telehealth' as a skill, doubling from 20% in 2020.



But with stringent regulations surrounding the extraction and processing of data from documents within healthcare, app developers in the healthcare space are faced with a challenge. With Apyrse's IDP solution, developers have access to a powerful tool that can revolutionize data extraction processes, leading to improved patient care and operational efficiency.

With benefits such as API integration, structured data output, and security and privacy compliance, ***Apyrse IDP benefits day-to-day healthcare app functions in many ways:***

Efficient Electronic Health Record (EHR) Management

Facilitates extraction of data from PDF-based patient records, enabling healthcare providers to populate and update EHRs with unprecedented speed and accuracy

Streamlined Insurance Claims Processing

Automates extraction of pertinent data from insurance forms and related documents

Enhanced Clinical Research

Swiftly identifies key data points, such as patient demographics, treatment protocols, and outcomes, enabling researchers to analyze larger datasets in less time, accelerating the pace of discovery and innovation in healthcare

Improved Patient Engagement

Fosters patient engagement by providing individuals with access to their medical records and relevant information in a user-friendly format

Compliance and Security

Prioritizes compliance by securely extracting and handling sensitive patient information from PDF documents

Reduces the risk of human error and ensures adherence to regulatory requirements

Employs advanced encryption and data protection measures to safeguard patient data throughout the extraction and processing workflow, providing healthcare organizations with peace of mind and ensuring the integrity of patient information

Enhancing AI Model Training with Apyrse IDP Data Extraction

Data extraction and organization plays a pivotal role in the success of AI model training, especially in finance and banking. Without quality data an accurate AI model cannot be created to perform the automated task.

A few use cases that rely on the extraction of unstructured data to train ML models for data monetization include:

Business Intelligence and Analytics Software:

Business intelligence and analytics platforms often extract unstructured data from various sources, such as social media, customer reviews, and text documents, to provide insights into market trends, customer sentiment, and emerging opportunities.

Customer Service Applications:

Call centers become much more efficient and lower their costs when they can aggregate data from support tickets, customer emails, SLA documents, and more to quickly solve their customers' problems.

Compliance and Risk Management Software:

In support of regulated industries like finance and healthcare, compliance and risk management solutions extract insights from unstructured legal documents and regulatory texts to ensure compliance with laws and regulations.

Importance of Data Extraction in AI Model Training

Data Quality: The quality of data is directly proportional to the performance of AI models. Even the most sophisticated algorithms cannot overcome the limitations of poor or inaccurate data. Data extraction ensures that data is clean, consistent, and error-free.

Data Relevance: Gathering only relevant data is crucial. Extracting irrelevant or redundant information can lead to extended training times and reduced model accuracy. A well-structured extraction process helps in filtering out unnecessary data.

Data Volume: Depending on the complexity of the AI model, a substantial volume of data may be required. Proper data extraction facilitates efficient data management, storage, and accessibility, thereby enhancing the effectiveness of the training process.

Importance of Data Organization in AI Model Training

Once data is extracted, the next step is to organize it effectively for AI model training. Data organization encompasses structuring, labeling, and categorizing the data, and is indispensable for several reasons:

Feature Engineering: Well-organized data simplifies the process of feature engineering, which involves selecting the most relevant attributes (features) and transforming the data into a format suitable for the model. This enhances the model's predictive capabilities.

Training Efficiency: Structured data accelerates the AI model training process. When data is organized consistently, the model can quickly grasp patterns and relationships, reducing training time.

Model Generalization: Properly organized data fosters better model generalization. This means the AI model can make accurate predictions on new, unseen data, as it has learned from a well-organized, diverse dataset.



Next Steps

Now that you have extracted the relevant data from this eBook PDF by reading it, you're ready to enable advanced document processing capabilities.

Embed Apryse's advanced OCR and data extraction to unlock intelligent document processing, enabling data-driven decisions across your clients' organizations. Automate tasks, extract insights from complex PDFs, and streamline workflows with precision and scalability.

Ready to learn more about Apryse IDP SDKs?

[Contact Sales](#)



Connect with Apyrse

Stay updated with all things Apyrse by following us on:

