Adobe PDF Extract API Technical Brief

V1.0 – 10/26/21

Introduction

Since 1993, when Adobe first introduced Acrobat and the Portable Document Format (PDF), people have been leveraging this canonical, application independent file format that straddles both paper and electronic media to enable easy and effective document sharing. In fact, PDF has been so successful in proliferating the use of electronic documents and automating document workflows that today, close to 80% of information created and used by organizations is captured in unstructured forms such as PDF. With the recent advent of Artificial Intelligence and Machine Learning (AI/ML), Natural Language Processing, and Robotic Process Automation, it has become critical for organizations to unlock the information captured in unstructured PDF documents to streamline and automate business processes that depend on this information.

The Adobe PDF Extract API, included with Adobe's PDF Services API, is a web-based document service hosted in Adobe's Document Cloud that leverages Adobe Sensei AI/ML technology to unlock document content and structure from PDF documents of all kinds – including reports, contracts, marketing materials, specifications, and more. The PDF Extract API can be used to extract information from PDF documents that are created by authoring applications, as well as those that contain scanned pages. PDF Extract uses AI/ML technology to identify and categorize the various objects within documents – such as paragraphs, lists, headings, tables, and images – and extract the text, formatting, and associated document structural information which is then delivered in a resulting JSON file. Extracted table data can optionally be delivered within .CSV or .XLSX files, and extracted images are delivered as .PNG files.

PDF Extract is different from other AI/ML document extraction services in the following ways:

- It works with PDF documents of all types without requiring ML model training.
- In addition to extracting raw text it extracts complete text blocks, identifies the types of text blocks extracted, and provides a deep understanding of the document structural context – including natural reading order, document organization via multiple heading levels, the location of text on a page, as well as text size, font, and styling.
- It can parse complex tables that include cells spanning multiple rows and/or columns, and also captures table formatting.
- It identifies and extracts figures – images, charts, and vector-based graphics – embedded within documents and delivers them as stand alone .PNG files.
- Finally, PDF Extract is cloud-platform agnostic and can be easily integrated with solutions that operate in any cloud platform.

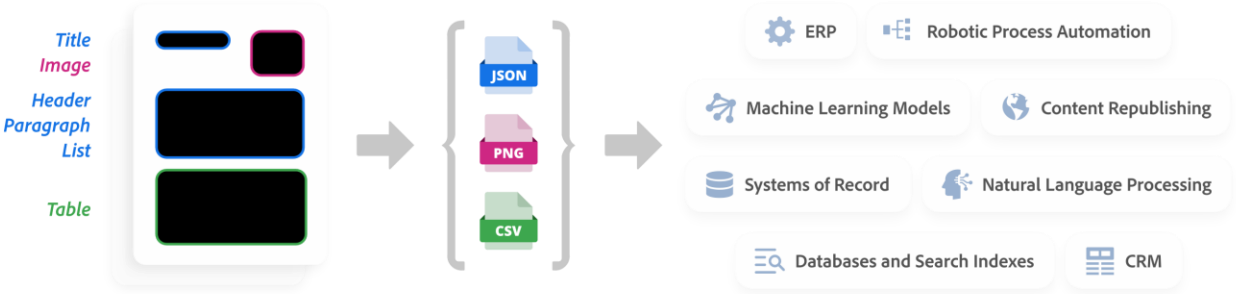
Security and Privacy

PDF Extract runs in Adobe's Document Cloud and inherits all the cloud platform features and benefits – including data privacy and security. Security practices are deeply ingrained into Adobe's internal software development, operations processes, and tools. These practices are

strictly followed by Adobe's cross-functional teams to help prevent, detect, and respond to incidents in an expedient manner. Adobe stays up to date with the latest threats and vulnerabilities and regularly incorporates the latest advanced security techniques into the products and services offered. Those services that touch customer content stay current with numerous industry certifications and comply with key government regulations, including GDPR. And all documents processed by PDF Extract are uploaded to Document Cloud and stored temporarily in cloud storage as part of normal service operations but are never stored permanently. This ensures that confidential information stays private. For more details on Adobe Document Services Security, please refer to the Security Overview document at: https://www.adobe.com/content/dam/cc/en/security/pdfs/AdobeDocumentServices_SecurityOverview.pdf.

Adobe PDF Extract Use Case Examples

Because PDF Extract can identify and interpret not just the raw text in a document but also document structural context, complex tables, and images, it enables a wide array of unique use cases as indicated in the diagram below.



Some of the use cases enabled by PDF Extract are summarized in this section.

Content Republishing

PDF Extract enables a wide array of content republishing use cases thanks to its ability to interpret document structure, reading order, page layout, and text and table formatting – as well as the ability to extract images. For example, PDF Extract allows the content of a given document to be extracted so the document contents can be translated into a different language, medium (e.g., audio), or data format (e.g., HTML) using automated translation tools, and the content can then be reconstructed as a fully transformed version of the original. PDF Extract also allows documents to be automatically parsed into their logical components which can each be stored and reused as building blocks to produce new, remixed documents – for example educational worksheets, quizzes, and tests.

Content Analysis and Insights

PDF Extract enables users to analyze not just the data contained in documents, but also the higher-order context of the document content. Through a deeper understanding of document structure – for example identifying text block types, associated headings, and extracting

paragraphs in their entirety, even when they span multiple pages or columns – PDF Extract enables more effective downstream processing of extracted content – including better results from NLP algorithms. PDF Extract enables a wide array of downstream outputs such as content summaries, key take-aways or recommendations, and even sentiment analysis across dozens, hundreds, or even thousands of PDF documents.

Content Aggregation and Search

In many cases, organizations want to create a knowledge base of the content from hundreds or thousands of documents for future reference and easier access. However, storing content within PDF containers is not the most efficient or effective approach. For example, while PDFs are searchable, finding all the paragraphs across hundreds or thousands of documents that relate to a specific topic or search query can be difficult. However, with PDF Extract the components of documents can be extracted in context, categorized, and organized within a knowledge base. This is possible thanks to PDF Extract's ability to extract complete paragraphs, interpret the structure and context of text paragraphs, and extract complex tables and images in industry standard data formats.

Process Automation

Information that can be used to automate business processes is often locked within PDF documents. Until recently, this data would have to be manually reentered to enable process execution. With PDF Extract, often working together with RPA technologies, manual data reentry can be avoided or eliminated by identifying and extracting the needed data and automatically populating it into downstream systems. PDF Extract's ability to identify data more effectively by understanding document structure and support for data extraction from complex tables makes it possible to automate more processes.

Content Extraction Capabilities

The table below summarizes the unique content extraction capabilities supported by PDF Extract.

Content Extracted	Description
Text Blocks	All the text within a document is extracted as holistic, logical text blocks and delivered within the resulting JSON file. PDF Extract works with PDF documents created by applications as well as scanned pages.
Text Block Types	PDF Extract identifies the type of each text block extracted in the resulting JSON. Text block types that can be identified include paragraphs, multi-leveled lists, titles, headings, asides, footnotes, and references (e.g., hyperlinks).
Complex Tables	Data contained in tables, including complex tables with cells that span multiple rows and columns, can be accurately extracted by PDF Extract. Header rows are identified, as is table formatting (e.g., text position within cell, border thickness, and

		background color). Table data is delivered in the resulting JSON file and can also optionally be delivered in a .CSV or .XLSX file. Table renditions can also be delivered as .PNG files to provide a visual representation to facilitate validation of extracted table data.
Text Formatting		Text size, font, and style – along with indentation, alignment, and justification – is captured for all text elements, including text within tables.
Figures		Figures – including charts, images, and vector graphics – are identified, extracted, and delivered as stand alone .PNG files.
Page Layout		The bounding box for each extracted object is provided in the resulting JSON to reflect the document layout.
Document Structure	Headings	Up to three levels of section headings are identified within the JSON output to capture the hierarchical structure of the document content.
	Reading Order	The natural reading order of the content within a document is reflected in the order of the JSON output, including paragraphs that span multiple pages and/or columns.
	Page Layout	The relative position of text blocks provided by the element bounding boxes can help to determine or validate their context within the document structure.
	Text Formatting	Text formatting can also be used to provide indication of document structure.

Using PDF Extract API

PDF Extract SDKs

PDF Extract is available in four different SDKs:

- Java
- .NET
- Node.js
- Python

PDF Extract can also be used as a REST API. A Postman API collection is provided to make this easier.

PDF Extract API Calls

PDF Extract provides several API calls to conveniently enable different extraction options. The table below summarizes the different API calls supported by PDF Extract and the corresponding outputs for each. Note that a "rendition" is a .PNG rendering of a table or figure (chart, image or vector graphic).

API Call	Output							
	Text Only (Including Block Type and Formatting) in JSON	Individual Character Bounding Boxes in JSON	Text Child Styling Information in JSON	Table Data in JSON	Table Renditions (Visual Representation in .PNG Files)	Table Data in .CSV Files	Table Data in .XLSX Files	Figure Renditions in .PNG Files
Extract Text	X							
Extract Text and Tables	X			X			X	
Extract Text and Tables (w/ Table Renditions)	X			X	X		X	
Extract Text, Tables (w/ Table Renditions) and Figures	X			X	X		X	X
Extract Text, Tables (w/ Table Renditions), Figures, and Character Bounding Boxes	X	X		X	X		X	X
Extract Text, Tables (w/ Table Renditions and .CSV files) and Figures	X	X		X	X	X		X
Extract Text, Tables, and Text Child Styling Information	X		X	X			X	

PDF Inputs

PDF Extract input can be any PDF file, including documents with scanned pages. For more information refer to the developer documentation: https://opensource.adobe.com/pdftools-sdk-docs/extract/latest/extract.html#api-limitations.

PDF Extract Outputs

PDF Extract outputs extracted document content in a single .ZIP file that includes the following files:

- JSON file, which includes the following data for each object identified in the PDF document:
  - Element type (text block type, table, figure as specified in the Path attribute)
  - Bounding box
  - Text (for elements that contain text)
    - Text font, size, and style
    - Indentation, alignment, and justification
  - Table data
    - Cell-by-cell data
    - Table formatting
    - Filenames for .CSV, .XLSX, and .PNG rendition files
  - Figures

        - Filenames for .PNG files
    - Elements are listed sequentially in their natural reading order
- .CSV or .XLSX files and .PNG renditions for each table identified (optional)
- .PNG files for each Figure extracted (optional)

Summary

Adobe PDF Extract API provides a powerful tool to extract and leverage the content captured in PDF documents in modern, ML/AI-enabled workflows and solutions. Its ability to extract text, complex tables, and images – coupled with an understanding of document structure – enables advanced, powerful content republishing, analysis, and process automation use cases. With no ML model-training required and support for the most popular programming languages (including a REST API), PDF Extract can be easily used together with components running in other cloud platforms across a wide array of solutions. Most importantly, PDF Extract is made by Adobe, the company that not only invented PDF but has been developing innovative PDF solutions for almost 30 years, making it a safe and trusted choice.

Next Steps

Below are some helpful links with more information about PDF Extract:

PDF Extract Product Home Page: https://www.adobe.io/apis/documentcloud/dcsdk/pdf-extract.html
PDF Extract Developer Documentation:
https://www.adobe.io/apis/documentcloud/dcsdk/docs.html