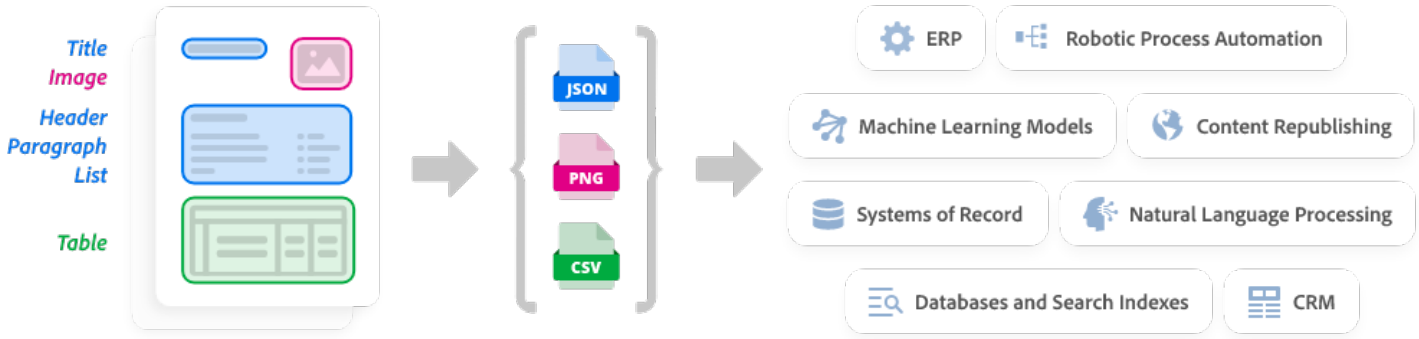
About PDF Extract API

Structured Information Output Format

Introduction

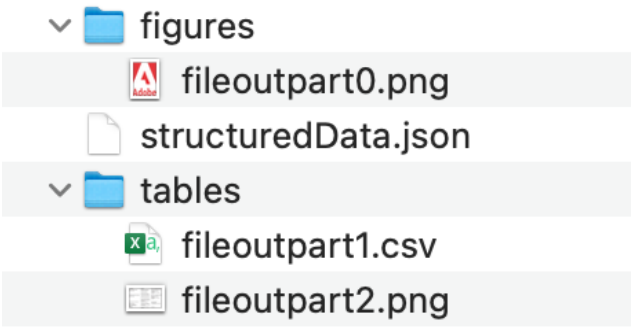
Extracted content is output in a structured JSON file, with tables optionally included as CSV or XLSX files and images saved as PNG files, so developers can easily store, analyze, and manipulate the data in a variety of downstream systems. Examples include databases, systems of record, CRM, ERP, NLP, RPA as well as ML models and analytic tools.



The output of an SDK extract operation is a zip package containing the following:

The structuredData.json file with the extracted content & PDF element structure. See the JSON schema for a description of the default output. (Please refer the Styling JSON schema for a description of the output when the styling option is enabled.)

A renditions folder(s) containing renditions for each element type selected as input. The folder name is either “tables” or “figures” depending on your specified element type. Each folder contains renditions with filenames that correspond to the element information in the JSON file.



© 2022 Adobe. All rights reserved.

List of key components

The following is a summary of key components in the extracted JSON:

- Elements: Ordered list of semantic elements (like headings, paragraphs, tables, figures) found in the document, on the basis of position in the structure tree of the document. The output does not include headers or footers. In addition, headings that repeat across pages are reported for the first occurrence only.
- Bounds: Bounding box enclosing the content items forming this element. Not reported for elements which don't have any content items (like empty table cells).
- Font: Font description for the font associated with the first character. Only reported for text elements.
- TextSize: Text size (in points) of the last character. Only reported for text elements.
- Attributes: Includes additional properties like line height and text alignment.
- Path: The Path describes the location of elements in the structure tree including the element type and the instance number. Element types are based on the ISO standard[1], a summary is included for convenience.
- Text: Text for the element in UTF-8 format, only reported for text elements. When inline elements are reported separately from parent block element, then this value has references to those inline elements.
- Figures: Identified as a Figure in the Path attribute, saved as a PNG in the figures folder with the filename identified in the filePaths attribute.
- Tables: Identified as a Table in the Path attribute, saved as a .CSV, .XLSX, and .PNG in the tables folder with the filename identified in the filePaths attribute.
- FilePaths: List of file paths to additional output files (images and spreadsheets)
- Pages: A list of properties for each page of the PDF including page number, width, height, and rotation.
- Reading Order: The reading order of content within columns, across page breaks, and inclusive of asides is represented by the order of the elements in the Elements array. In the normal mode, exceptions can occur for elements extracted from their container (eg. A reference link in the middle of a paragraph). However, the order is preserved in Styling mode where all Elements and their Kids are represented in the natural reading order.

How are businesses using PDF Extract API?

Some common and relevant use cases include: loan document workflows, content based process automation, data analysis, processing employee resumes, field service management, and streamline procurement processes.

[1] ISO 32000-2 defines PDF 2.0 and is the first PDF specification entirely developed under the ISO open consensus-based process.

© 2022 Adobe. All rights reserved.

SUMMARY OF ELEMENT TYPES

CATEGORY	ELEMENT TYPE	DESCRIPTION
Aside	Aside	Content which is not part of regular content flow of the document
Figure	Figure	Non-reflowable constructs like graphs, images, flowcharts
Footnote	Footnote	Footnote
Headings`	H	Beginning of heading
	H1	Heading Level 1
	H2	Heading Level 2
	etc	Heading Level X
List	L	Beginning of list
	Li	List Item
	Lbi	List Item label
	Lbody	List Item body
Paragraph	P	Paragraph
	ParagraphSpan	Denotes part of a paragraph. Reported when paragraph is broken (generally due to page break or column break)
Reference	Reference	Link
Section	Sect	Logical section of the document
StyleSpan	StyleSpan	Denotes difference in styling of text relative to the parent container
Sub	Sub	Single line of a multiline paragraph (e.g. addresses). Such paras are created in html using inside <p> tags
Table	Table	Beginning of table
	TD	Table cell
	TH	Table header cell
	TR	Table row
Title	Title	Title of the document. This is the most prominent heading which can define the whole document.

© 2022 Adobe. All rights reserved.