



Processor



Key-Value Pairs



PROCESSOR > GUIDES > EXTRACTION > KEY-VALUE PAIRS

Transforming document extraction with machine learning



PSPDFKit Processor has been deprecated and replaced by [Document Engine](#). To migrate to Document Engine and unlock advanced document processing capabilities, refer to our migration guide. Learn more about these enhancements on our [blog](#).

The extraction of key-value pairs involves two tasks:

- ✧ Use OCR technology to recognize unstructured information and text in a document.
- ✧ Use machine learning, specifically deep learning, to make sense of the unstructured information by composing links between different parts of the extracted text.

A combination of both approaches is necessary to achieve the best results in data extraction. For this reason, Nutrient recognizes text and key-value pairs based on a hybrid approach of the following methods:

- ✧ Heuristics
- ✧ Mathematics
- ✧ Machine learning (ML)

This approach produces superior results compared to traditional optical character recognition (OCR) and pure ML approaches.



ASK AI

Traditional approaches

This section explains the traditional approaches to key-value pair extraction.

Traditional OCR

Extracting data with the traditional OCR approach is based on heuristics. The biggest limitation of the traditional OCR approach is that it needs to use a different template for each document type. This works well for simple documents with structured data. However, extracting data with the traditional OCR approach doesn't perform well with unstructured or semi-structured documents.

Extracting data with this approach suffers from the same limitations as traditional OCR engines that have difficulties recognizing text in the following contexts:

- ✧ Colored backgrounds
- ✧ Glaring
- ✧ Skew
- ✧ Text in tables and graphics
- ✧ Handwritten text

Lastly, data extraction solutions relying only on traditional OCR are difficult to scale.

Machine learning and deep learning

Data extraction solutions that leverage machine learning and deep learning use artificial intelligence (AI) technologies to mitigate traditional OCR limitations. These deep learning approaches are usually a combination of different techniques, such as convolutional neural networks, long short-term memory layers, transformers, and graph neural networks.

Data extraction relying only on machine learning and deep learning often fails for documents with a lot of noise and dotted lines.

Nutrient data model

By automatically recognizing the document type, Nutrient adapts to the context and determines the extraction approach that makes the best use of available resources. Nutrient recognizes the document type based on adaptive layout understanding and natural language processing (NLP) technologies.

This hybrid approach includes heuristics, mathematics, and machine learning (ML), and it address the usual weaknesses of the traditional OCR and pure ML engines.

The Nutrient data model enables you to extract data from documents with excellent results. Nutrient’s hybrid approach performs better than traditional OCR and pure ML engines, especially for documents with the following features:

- ✧ Noise
- ✧ Dotted lines
- ✧ Broken characters
- ✧ Text on colored backgrounds
- ✧ Underlined text
- ✧ Skewed text
- ✧ Text in graphics and tables

Was this helpful?

☒ YES

☐ NO

Questions? [Contact us](#)

