IDA Extraction

Few-shot learning intelligent data extraction

IDA Extraction offers **few-shot learning** for smart zonal data extraction from documents. Thanks to its **advanced machine learning capabilities**, the setup and maintenance of data extraction workflows are dramatically accelerated compared to rule-based or manual approaches. Combined with PLANET AI's **exceptional OCR and ICR capability**, the need for manual corrections decreases significantly, resulting in enhanced straight-through processing.

KEY FEATURES

Refined zonal data extraction

IDA Extraction utilizes an advanced key-value pair extraction method that empowers users to easily specify the data fields they wish to capture. These fields are not limited to text but can also encompass barcodes, checkboxes, and numerical values. IDA provides positional information, which proves valuable for subsequent downstream tasks, such as validation.

Few-shot learning capabilities

IDA Extraction features few-shot learning capabilities that consider both visual and textual features on documents. By presenting the system with just a few training documents and specifying the desired data fields, it can extract those from pages it hasn't seen during training. This approach not only reduces the time to value but also minimizes the effort required to adapt to changing document layouts.

SYSTEM REQUIREMENTS

For 64-bit systems

Linux: Ubuntu 20.04 - 23.10, Debian 11, CentOS 8, Red Hat 8.x, LEAP 15.x, SLES 15 SP 4-5
Windows: 10, 11
Windows Server: 2012, 2012 R2, 2016, 2019, 2022
Docker

At least **9 GB hard disk storage**
+ approx. 0.9 GB for full functionality
+ up to 0.5 GB to store logging data

No-code training

The IDA Extraction Assistant (ExA) empowers users without technical expertise to create, train, and customize data extraction models using a graphical interface accessible via web browser.

Leveraging unmatched OCR quality

IDA Extraction is based on **IDA Recognition**, an optical (OCR) and intelligent (ICR) character recognition engine that delivers outstanding results even when dealing with the most difficult scenarios. IDA Recognition captures machine-printed and handwritten text, checkboxes, tables, and historical scripts, even in poor-quality scans with rotated or skewed print.

Having high-quality input data is crucial for data extraction tasks as it directly affects the quality of the extraction output.

Easy deployment and integration

IDA is deployed either on-premises or in a (private) cloud as a Java application or containerization using Docker. The gRPC API facilitates seamless and swift integration.

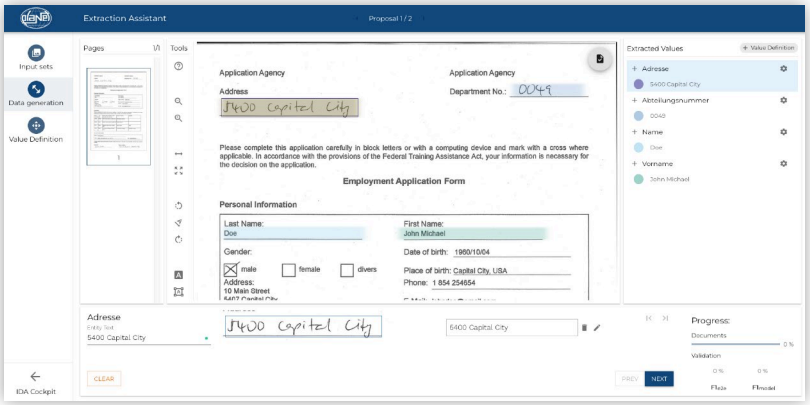
Versatile JSON output format

Extracted data fields are output in a JSON format for easy access in subsequent tasks during downstream processing.

MODEL TRAINING

IDA provides the **"Extraction Assistant" (ExA)**, a graphical interface that allows users to train models without needing programming skills or having to prepare complex datasets. Currently, ExA performs best on **structured and semi-structured documents** such as forms or invoices. Based on the document categorization performed by IDA Classification, documents can be routed to different extraction models.

As a general guideline, a **minimum of five documents** per document class is recommended. However, it is important to note that having a larger number of training documents typically leads to a better model. Users can define an unlimited number of data fields to extract.



ExA labeling interface

To accelerate the manual labeling process, the assistant conducts an **automatic pre-training** to identify recurring anchor points within the documents. As a result, ExA groups the training documents for improved efficiency. After labeling one of these sample documents, users only need to validate or correct the suggest extracted fields. An automatic pre-training requires at least **three documents of the same layout**. Acroform fields will be prioritized and automatically suggested for use as a data field.

For more information, please refer to the software documentation.