IDA Extraction

Smart data retrieval to reduce manual work

IDA Extraction provides **advanced information retrieval** for both structured and unstructured documents. Utilizing machine learning capabilities like **smart zonal data extraction**, **large language models (LLM)**, and **LayoutLM**, it enables accurate capture of data items within documents. This significantly reduces the need for manual labor in validating results or setting up rule-based workflows, resulting in **enhanced straight-through processing**.

PRODUCT CONFIGURATIONS

IDA Extraction Assistant (ExA)

- Ideally suited for structured and semi-structured documents
- Example use case: forms processing

IDA LLM Entity Extraction$^{\text{BETA}}$

- Ideally suited for unstructured documents
- Example use case: full-text document indexing

KEY FEATURES

Wide scope of scenarios

IDA Extraction utilizes a range of technologies from **machine learning and natural language processing (NLP)**, such as intelligent zonal extraction and large language models (LLMs). Its applications range from processing structured forms to analyzing documents containing unstructured text.

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.

Versatile output formats

Extracted data items are output in a **JSON** format for easy access in subsequent tasks during downstream processing. Additionally, results can be highlighted in their original locations in an output **PDF**.

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.

IDA EXTRACTION ASSISTANT

Few-shot learning capabilities

The Extraction Assistant (ExA) 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.

No-code training

ExA empowers users **without technical expertise** to create, train, and customize data extraction models using a graphical interface accessible via web browser.

SYSTEM REQUIREMENTS

The **IDA Server** is required to process input documents and also provides a browser interface.

For 64-bit systems
Linux: Ubuntu 18.04 - 23.10, Debian 11, 12; CentOS 8, Red Hat 8.x, 9; LEAP 15.x, SLES 15 SP 4-5
Windows: 10, 11
Windows Server: 2016, 2019, 2022
Docker

At least **12 GB hard disk storage**

At least **16 GB RAM**

Refined zonal data extraction

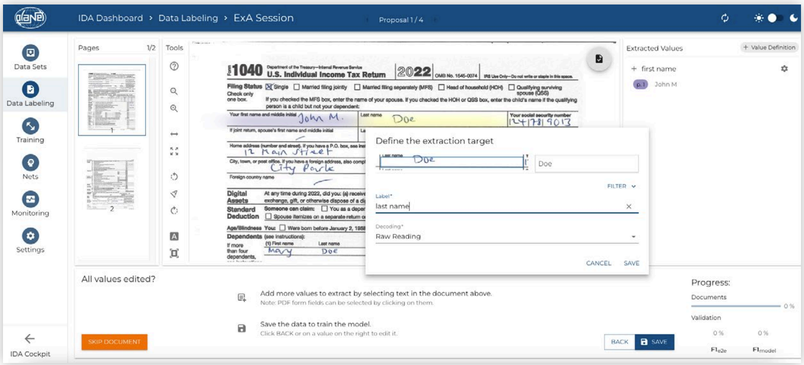
ExA utilizes an **advanced key-value pair extraction method** that empowers users to easily specify the data fields they wish to capture. These fields are not limited to text but can also encompass checkboxes and numerical values. IDA provides positional information, which proves valuable for subsequent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

Model Training

With the **Extraction Assistant**, IDA provides a graphical interface that allows users to train models without having to prepare complex datasets. Currently, ExA performs best on **structured and semi-structured documents** such as forms or invoices.

Users can optionally utilize the **Layout Language Model (LayoutLM)** to enhance their extraction results. This model prioritizes visual cues and contextual comprehension, and has proven to be particularly effective helpful for processing semi-structured documents. 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.



Label*: last name

Decoding*: Raw Reading

All values edited?

Add more values to extract by selecting text in the document above.
Note: PDF form fields can be selected by clicking on them.

Save the data to train the model.
Click BACK or on a value on the right to edit it.

ExA labeling interface

To expedite the manual labeling process, the assistant performs an **automatic pre-training** to detect 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 are simply required to verify or correct the proposed extracted fields. Constructing a sample requires at least **three documents with identical layouts**. Acroform fields will be prioritized and automatically suggested for use as a data field.

IDA LLM ENTITY EXTRACTION^{BETA}

Keyword-based Named Entity Recognition

With IDA LLM Entity Extraction^{BETA}, you can automate data extraction from unstructured documents **simply by providing queries**. This includes labeling the key elements to be extracted and a brief description or a list of keywords. No training is required beforehand.

Verification of extracted data

To prevent AI hallucinations, the resulting extracted data is **cross-verified** with the transcription from IDA Recognition.

No-code solution

IDA LLM Entity Extraction^{BETA} empowers users who lack technical skills or proficiency in formulating prompts. The module can be easily customized via a **browser interface**.

SYSTEM REQUIREMENTS

To utilize large language models, it is necessary to have a dedicated server, known as the **LLM Server**.

For 64-bit systems

Docker (Ubuntu-based)

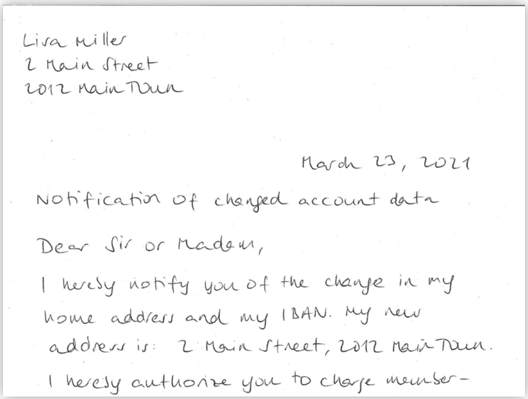
At least **40 GB GPU memory** (can be spread out across multiple GPUs)
At least **6.5 GB hard disk storage**
+ at least 20 GB for LLM
At least **64 GB RAM**

The required hard disk storage and the necessary RAM are both significantly dependent on the models intended to run on the LLM Server.
No CPU-only mode is possible.

How does it work?

IDA LLM Entity Extraction^{BETA} is **best suited for handling unstructured documents** that lack fixed layouts or data points, such as contracts or cover letters. Based on the document categorization performed by IDA Classification, documents can be routed to different extraction models.

Input document



Lisa Miller
2 Main Street
2012 MainTown

March 23, 2021

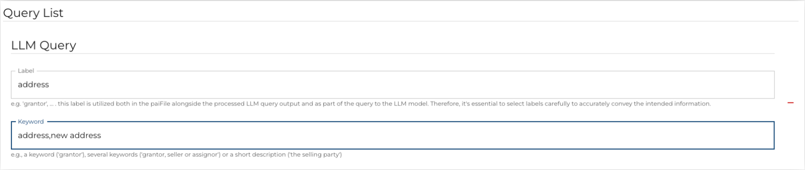
Notification of changed account data

Dear Sir or Madam,

I hereby notify you of the change in my home address and my IBAN. My new address is: 2 Main Street, 2012 MainTown.

I hereby authorize you to charge member-

+



Query List

LLM Query

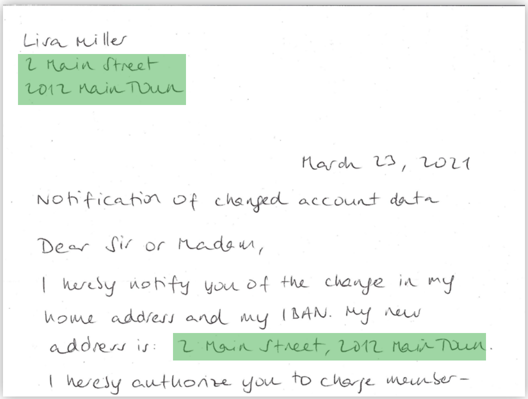
Label: address

e.g. 'grantor', ... this label is utilized both in the paiFile alongside the processed LLM query output and as part of the query to the LLM model. Therefore, it's essential to select labels carefully to accurately convey the intended information.

Keyword: address,new address

e.g., a keyword ('grantor'), several keywords ('grantor, seller or assignor') or a short description ('the selling party')

Output document



Lisa Miller
2 Main Street
2012 MainTown

March 23, 2021

Notification of changed account data

Dear Sir or Madam,

I hereby notify you of the change in my home address and my IBAN. My new address is: 2 Main Street, 2012 MainTown.

I hereby authorize you to charge member-

+

An LLM query consists of a label paired with a keyword, a list of keywords, or a brief description. The results produced by the LLM are **subsequently cross-checked with the Entity Finder and the transcription**. This ensures high accuracy and prevents any incorrect results, often referred to as “hallucinations”.

For more information, please refer to the software documentation.