IDA Extraction

Precise data extraction for less manual work

IDA Extraction automates data capture from structured and unstructured documents without complex rule sets. Whether forms, contracts, or free-text documents: The combination of few-shot learning and LLM technology extracts precisely the data you need. The result: Drastically reduced manual effort for validation and workflow setup, significantly higher straight-through processing rates, and high-quality data as the foundation for downstream processes.

TWO EXTRACTION OPTIONS

IDA Extraction Assistant (ExA)

- For structured and semi-structured documents
- Key-value pair extraction

IDA LLM Entity Extraction

- For unstructured documents
- Zero-shot named entity recognition

BENEFITS

Wide scope of scenarios

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

Leveraging unmatched OCR quality

IDA Extraction is built on **IDA Recognition**, the OCR engine for outstanding results in the most difficult scenarios. Even with distorted scans, poor image quality, and difficult-to-read handwriting, IDA Recognition delivers the high-quality text foundation that's critical for reliable data extraction. Why this matters: Data extraction quality depends entirely on input data quality. IDA Recognition captures machine-printed and handwritten text, checkboxes, tables, and historical scripts as the perfect foundation for downstream processes.

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 (alternatively REST API) facilitates swift integration.

IDA EXTRACTION ASSISTANT

Few-shot learning capabilities

The IDA Extraction Assistant (ExA) uses few-shot learning that analyzes visual and textual features. With just a few sample documents, the system is ready for deployment and extracts data from documents 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

Users without programming skills can independently create, train, and customize extraction models. The browser-based user interface makes machine learning accessible to everyone.

SYSTEM REQUIREMENTS

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

For 64-bit systems
Linux: Ubuntu 18.04-25.10, Debian 11-13, CentOS 8-10, Red Hat 8.x-10.x, LEAP 15.4-15.6, 16.0; SLES 15 SP 4-7
Windows: 10, 11
Windows Server: 2016, 2019, 2022, 2025
Docker

At least **12 GB hard disk storage**
At least **16 GB RAM**

Refined zonal data extraction

ExA extracts key-value pairs from structured documents. Simply specify the data fields you want to capture – not just text, but also checkboxes and numeric values. IDA provides positional information that proves valuable for downstream tasks like validation or archiving.

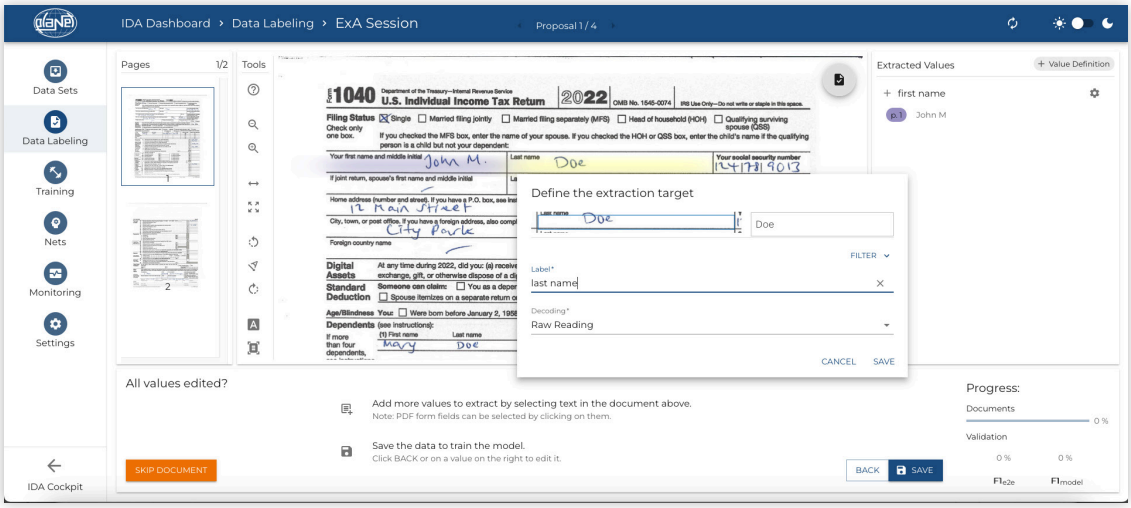
Model Training

ExA's graphical interface eliminates the need for complex dataset preparation. Train models directly in the browser – without technical setup. Currently, ExA performs best on structured and semi-structured documents such as forms.

An **automatic pre-training** accelerates the process: ExA automatically recognizes recurring anchor points in your documents and groups similar layouts. Result: You label only one sample document per layout group instead of every single document.

How it works:

- At least 3 documents with identical layout form a pattern
- ExA automatically suggests extraction fields
- You validate or correct the suggestions
- AcroForm fields are prioritized and automatically recognized



ExA labeling interface

Training recommendations:

- Minimum: 5 documents per document class
- Optimal: The more training documents, the better the model
- For structured forms: Few examples are sufficient

Users can optionally utilize the **Layout Language Model (LayoutLM)** to improve extraction results. This model prioritizes visual cues and contextual comprehension and has proven to be particularly effective for processing semi-structured documents. Based on the document categorization performed by IDA Classification, documents can be routed to different extraction models.

With IDA 5.3, ExA also serves to **validate LLM Entity Extraction**. Create ground-truth sets and test different prompt variants objectively – for reliable LLM extraction in production environments.

IDA LLM ENTITY EXTRACTION

Keyword-based queries

Automate data extraction from unstructured documents through simple queries and without prior training. Formulate what you're looking for: A label combined with a keyword, a keyword list, or a brief description.

Verification of extracted data

AI hallucinations are a known problem with LLM-based systems. IDA prevents them through an additional verification step: The data extracted by the LLM is cross-checked against the OCR transcription from IDA Recognition.

Prompt validation

LLM-based extraction previously came with uncertainties: Unpredictable model performance and lack of comparability between prompts made manual screening necessary. With IDA, you can validate LLM prompts against defined test data and thus compare prompt variants before going into production.

Invoice module[BETA]

IDA Extraction includes a module for invoice processing based on LLM Entity Extraction. It provides pre-configured entities and is ready for immediate deployment.

REQUIREMENTS

To utilize large language models (LLMs) on-premises, a dedicated **LLM Server** is necessary:

For 64-bit systems

- **Docker (Ubuntu-based)**
- At least **40 GB GPU memory** (can be split across multiple GPUs)
- At least **6.5 GB hard disk storage**
  + at least 20 GB for LLM
- At least **64 GB RAM**

Important: Disk storage and RAM depend heavily on the chosen models. Note that a CPU-only mode is not possible.

The LLM Server can also connect to **OpenAI models** and then acts as a relay. This means that no GPUs are necessary.

No-code solution

Users without programming skills can independently create prompts. The module can be easily customized via the browser interface.

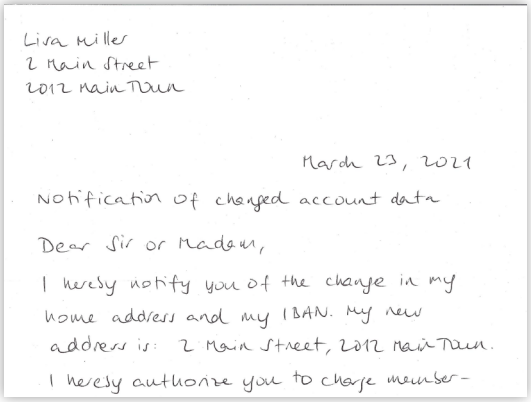
How does it work?

IDA LLM Entity Extraction 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.

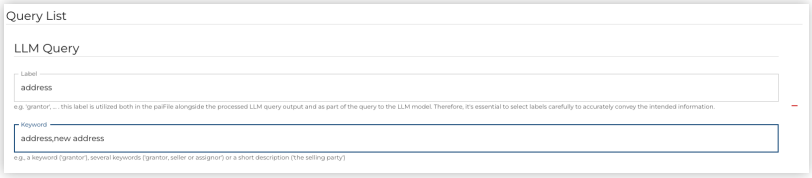
An **LLM query** consists of:

- Label: Name of the data field to be extracted
- Definition: Keyword, keyword list, or brief description
- Verification: Cross-check with Entity Finder and OCR transcription

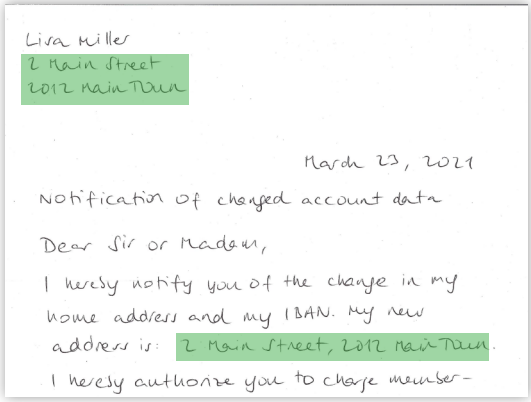
Input document



+



Output document



+

Experience IDA Extraction for yourself and contact us for a personalized demo.

Book demo