Low-Effort Document Indexing and Metadata Extraction with IDA

Indexing is a critical component of document management systems aimed at **structuring and categorizing documents to facilitate the retrieval of information**. It renders documents readily accessible and searchable by converting them into digital formats and tagging them with metadata such as dates, authors, and other relevant elements. There are various types of document indexing such as:

Metadata indexing

Metadata indexing involves cataloging descriptive information (metadata) about the document. This method is particularly useful for searching by document type than its content. Search systems can quickly filter through collections based on these attributes.

Full-text indexing

Full-text indexing involves analyzing and cataloging the entire content of a document. This allows for more comprehensive searches for words or phrases within the text of documents. This method is particularly useful for searching specific content.

Both methods have their use cases and can complement each other. Organizations need to decide on the right approach based on their specific requirements and the nature of the data they are managing.

Discover how the IDA software suite excels at enhancing the indexing process across various document types, significantly reducing expenses for manual labor.

Minimize manual efforts

- Zero-shot entity extraction, enabled by large language models (LLM)
- Outstanding text recognition accuracy for the most difficult scenarios, including handwriting

Reduce maintenance

- Low-effort training for changing document layouts
- No-code approach to user interfaces

Ensure compliance

- On-premises or private cloud deployment

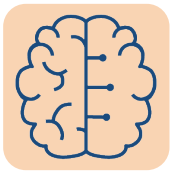
COMMON CHALLENGES

Digitization efforts often experience delays during the indexing phase. Achieving **accurate information retrieval** from distorted, poor-quality scans with machine-print and handwriting often requires substantial human involvement for validation and correction. Automating data extraction from **unstructured documents**, which can constitute up to 80% of an organization's document volume, has until now appeared to be a nearly impossible challenge. If OCR (Optical Character Recognition) results are not sufficiently accurate, the process often resorts to full manual execution.

Furthermore, **maintaining rule-based document classification and data extraction methods** for handling new or modified document layouts, as well as more detailed categories, can be both complex and costly.

THE SOLUTION

PLANET AI's Intelligent Document Analysis enables **low-effort auto-indexing and metadata extraction** of large document volumes. With its **rule-free approach to document classification and data extraction,** IDA requires minimal training data and low maintenance.



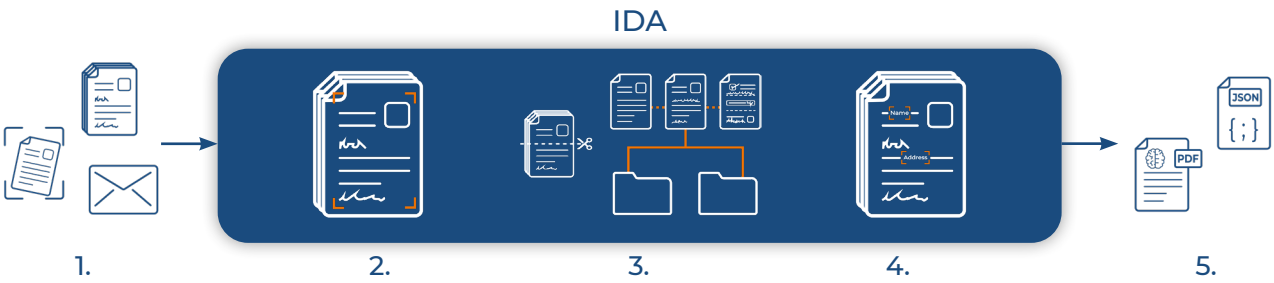
By leveraging large language models (LLM), IDA enables **zero-shot entity recognition and extraction**[BETA] **from unstructured documents**, which is accessible through a no-code, intuitive user interface. Misrecognitions in the AI's output ("hallucinations") are mitigated by verifying extracted data against the textual basis. To implement effective human-in-the-loop involvement, results are highlighted in their original positions.

IDA delivers **unmatched OCR and ICR accuracy**, minimizing the need for manual correction even in the most challenging scenarios. As a result, machine learning models are fed with the highest quality input.

IDA can be deployed **on-premises or in a (private) cloud** as either a Java application or by containerization using Docker to ensure compliance with privacy and security regulations.

HOW IT WORKS

IDA workflow for low-effort document indexing and metadata extraction:



1. Input: Physical and electronic documents via scanner, mailbox, email etc.

2. Recognition: OCR and ICR capability based on patented PerceptionMatrix

3. Document splitting and classification: Rule-free, few-shot learning separation of large consecutive documents and document categorization

4. Data extraction depending on document input:

a) For metadata indexing: Zonal data extraction to capture data fields from structured and semi-structured documents, such as forms

b) For full-text indexing: LLM-based entity extraction[BETA] for unstructured documents, such as contracts

5. Output: PDF or PDF/A (all conformance levels) with text layer containing recognition results and optionally highlighted metadata and/or JSON with metadata including positional information, confidence score etc.

CUSTOMER SUCCESS STORY

Document indexing finds application in various scenarios, such as records management within business process outsourcing (including scanning services), document and content management, as well as digital libraries and archives.

Our renowned client has been offering **business process outsourcing services** to healthcare providers, the public sector and enterprise customers for over 50 years. They struggled with an **automation rate of only 50% for document classification**. As a result, a substantial amount of manual work was required to correct and validate documents. IDA's few-shot learning approach enabled a swift initial setup and **pushed their automation rate to 90%, resulting in an 80% reduction in manual efforts**.