

Low-Effort Document Indexing and Metadata Extraction with IDA

Indexing is a process in document management systems aimed at structuring and categorizing documents to facilitate the retrieval of information. It renders documents easily accessible and searchable by converting them into digital formats and tagging them with metadata such as dates, authors, and other relevant elements.

Discover how the IDA software suite excels at simplifying and streamlining the document indexing process.

Reduce manual efforts

Outstanding text recognition accuracy for the most difficult scenarios, including handwriting

Minimize maintenance

Low-effort training for changing document layouts

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. If the OCR results are not sufficiently accurate, the process often resorts to full manual execution. Additionally, **maintaining rule-based document classification and data extraction methods** for handling new or modified document layouts and more detailed categories can be both complex and costly.

THE SOLUTION

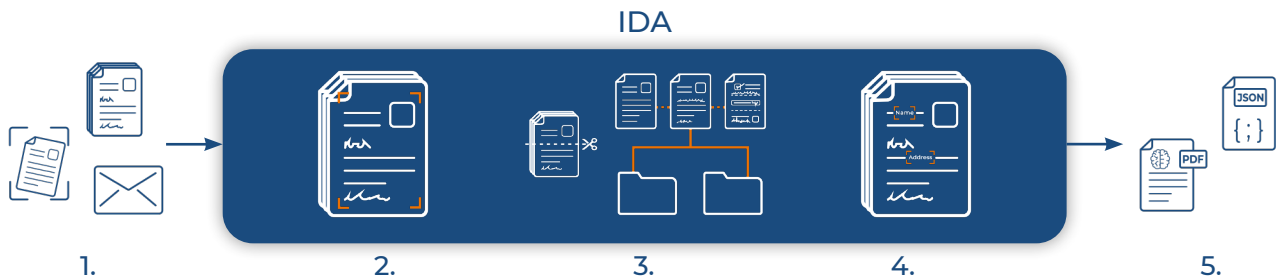
PLANET AI's Intelligent Document Analysis enables **low-effort auto-indexing and metadata extraction** of large document volumes. By combining the patented core technology with sophisticated machine learning capabilities, IDA delivers **unmatched OCR and ICR accuracy**, minimizing the need for manual correction even in the most challenging scenarios.

SOLUTION BRIEF: LOW-EFFORT DOCUMENT INDEXING

With its **rule-free approach to document classification and data extraction**, IDA requires minimal training data and low maintenance efforts. IDA can be deployed **on-premises or in a (private) cloud** as a Java application or containerization using Docker.

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. Metadata extraction: Zonal data extraction to capture data fields from documents

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**.