INDIUM

Fast Tracking Text Extraction for an Oil & Gas Company with teX.ai
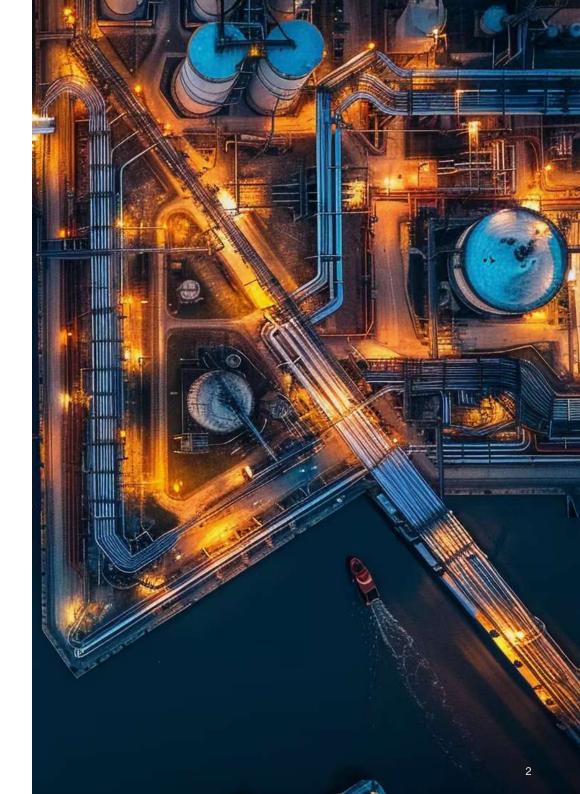
SUCCESS STORY

Client Overview

The client is one of the pioneers in the oil and gas industry, consistently pushing the boundaries of innovations to develop cutting-edge solutions that empower their customers to fuel progress in agriculture, industry, medicine, science, space, technology, and transportation. Their expertise extends beyond traditional energy exploration and extraction, encompassing engineering disciplines, computer science, geophysics, and metallurgy that fuel breakthroughs to create a winning formula for all stakeholders in such projects.

Data Extraction Challenges: Handling Complex Engineering PDFs

The client frequently dealt with an extensive collection of PDF documents containing highly detailed diagrams of drilling machine components, complex, nested tables, and other unstructured data formats.

Manually extracting relevant information from these PDFs was time-consuming, error-prone, and inefficient. The client sought an advanced, automated solution that could accurately extract and store the data in a structured format, enabling seamless retrieval, analysis, and reporting. The goal was to enhance operational efficiency by ensuring that critical data could be extracted and saved in a format that could facilitate further analysis downstream.

Key Challenges:

High Volume of Documents:

1. The client needed to process hundreds of PDF documents containing vital engineering data.
2. The number of pages per document varied widely from 2 to 100, making standardization difficult.

Inconsistent Data Distribution:

1. Not all pages within a document contain the required data.
2. This inconsistency made it challenging to automate data extraction without intelligent filtering.

Need for Custom AI Models:

1. A single extraction model was insufficient due to the variability in document structure.
2. Separate AI models had to be trained and fine-tuned to process each document type accurately.
3. These models were needed to handle text, tables, images, and technical schematics precisely.

Diverse Document Formats:

1. The client had to deal with five different document formats, each requiring a unique approach for extraction.
2. These formats included:
   - Engineering Drawings – Containing intricate diagrams and schematics.
   - Nested Tables – Hierarchical data structures embedded within tables.
   - Un-demarcated Tables – Tables without clear borders or separations, making traditional extraction methods ineffective.
   - Other free-form and complex layouts require specialized processing.

AI-powered Data Retrieval:
How teX.ai Streamlined Complex Engineering Document Processing

The client dealing with vast amounts of technical documentation is a critical yet challenging task. Engineers and analysts frequently work with PDFs containing intricate schematics, nested tables, and complex datasets. Extracting meaningful information from these documents manually is time-consuming and prone to errors. To address this challenge, Indium's AI-powered NLP accelerator, teX.ai, provided an advanced solution for automating data retrieval, ensuring accuracy, efficiency, and seamless processing of engineering documents.

We have highlighted how teX.ai successfully tackled key document processing challenges, such as validating file quality, extracting structured data from unstructured sources, and handling complex document formats, including chemical composition reports, survey files, and well-schematics.

Quality File Validation

- Ensuring the integrity and consistency of input files was a crucial first step. The solution implemented a file validation mechanism to verify document quality before processing, preventing errors and inconsistencies in extracted data.

Chemical Composition Extraction & Key-Value Pair Conversion

- Extracted chemical composition data from PDF files and converted it into structured key-value pairs for more straightforward analysis.
- These chemical composition reports span up to 10 pages per document, requiring advanced parsing and formatting techniques.

Automated Survey File Processing

- Implemented an AI-driven approach to identify and extract survey tables from multi-page documents automatically.
- Ensured that survey data was accurately categorized and extracted, enabling efficient data retrieval and analysis.

Well Schematics Extraction

- Developed an intelligent parsing system to identify and extract nested tables as separate entities from complex engineering PDFs.
- These documents contained a combination of intricate, good schematics and drilling equipment drawings, requiring a hybrid approach of image processing and table extraction.

Implementing teX.ai for Seamless Text Extraction for an Oil & Gas Company

It involves multiple AI models tailored to recognize and accurately process text, tables, and diagrams.

The AI engine was fine-tuned to detect patterns in unstructured documents, classify content, and convert raw data into meaningful insights.

Quality File Validation

▶ The Analysis table containing the chemical composition details was identified in the document and extracted using OCR.

▶ The time taken to extract is just a few seconds, and accuracy is more than 85%.

Public Files (Surveys)

▶ First, isolate the survey tables using the keyword search leveraging OCR.

▶ Survey details are then extracted using techniques such as Tabula or Camelot.

Well Schematics

▶ All the nested tables were extracted as separate tables and saved in CSV format.

▶ The nested tables are extracted in 2 stages, leveraging the FCN model at stage 1 and OpenCV in the next stage to detect rows in the table.

Deployment

▶ Once the AI models were built, and the required accuracy and performance tuning was complete, Indium deployed teX.ai with an admin interface built using Flask and containerization using Dockers.

Outcome Delivered: Speed, Accuracy & Efficiency

By leveraging AI-powered data retrieval, Indium's teX.ai successfully automated the extraction of structured data from complex engineering documents, enabling faster decision-making, reducing manual efforts, and enhancing operational efficiency in the oil and gas sector. Integrating teX.ai led to the following:

▶ **4x Faster Text Extraction:** Compared to traditional manual methods, teX.ai accelerated data extraction, drastically reducing processing time for complex engineering documents. This allowed engineers and analysts to focus on higher-value tasks rather than spending hours manually retrieving data.

▶ **80% Reduction in Human Intervention:** Automating text and table extraction minimized the need for manual validation and data entry, leading to substantial time and cost savings. This shift also reduced the risk of human errors, ensuring greater reliability in extracted data.

▶ **75% Improvement in Process Quality:** With AI-driven precision, the quality and consistency of extracted data improved significantly. The automated system ensured critical engineering information was retrieved accurately, enhancing decision-making and operational efficiency.

INDIUM

About Indium

Indium is an AI-driven digital engineering company that helps enterprises build, scale, and innovate with cutting-edge technology. We specialize in custom solutions, ensuring every engagement is tailored to business needs with a relentless customer-first approach. Our expertise spans Generative AI, Product Engineering, Intelligent Automation, Data & AI, Quality Engineering, and Gaming, delivering high-impact solutions that drive real business impact.

With 5,000+ associates globally, we partner with Fortune 500, Global 2000, and leading technology firms across Financial Services, Healthcare, Manufacturing, Retail, and Technology—driving impact in North America, India, the UK, Singapore, Australia, and Japan to keep businesses ahead in an AI-first world.