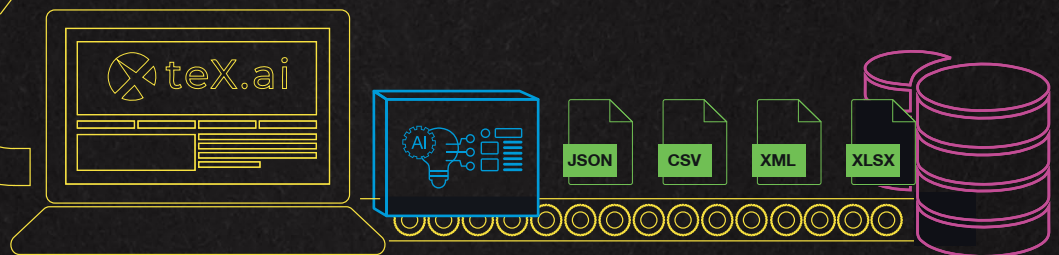# Nested Tables & Machine Drawing Text Extraction for an Oil & Gas Company

- 4x faster automated text extraction using teX.ai
- The need for human intervention was reduced by over 80%
- The quality of their process had increased by over 75%

teX.ai™

Customer Background

The Client is one of the pioneers in the oil and gas business, with a focus on innovation to find ways to help their customers to fuel progress in agriculture, industry, medicine, science, space, technology, and transportation. The combination of engineering disciplines, computer science, geophysics, and metallurgy help create a winning formula for all stakeholders in such projects.

Business
teX.ai

Domain
Oil & Gas Industry

Technology

The solution was built leveraging Python and several of its libraries Category-Wise Key Phrase OCR:

Tesseract, Tesserocr, OCRmyPDF, PyTesseract

Preprocessing and Post Processing Tools:

xPDF, Poppler, OpenCV, Pandas, Json

Table Detection and Extraction:

Camelot, OpenCV, LSD (line segment detection), csv, TensorFlow, FCN (Fully Convolutional Networks), CNN (Convolutional Neural Networks)

Application Deployment:

Flask, Docker

Key Highlights

- The accuracy level of 80% in identifying customer's problem area and classifying them.

- 20% reduction in same type complaints from customers within 3 months.

teX.ai™

Business Requirement

Given the document intensive nature of business, the client generally had to deal with numerous PDF documents dealing with complex drilling machine parts diagrams and data in nested tables and various other formats. Their requirement was to extract data and save in a format that could facilitate further analysis downstream.

Challenges

- Client had hundreds of PDF documents and each of these PDF documents had pages ranging from 2 to 100 pages. In some cases, the required data was not present in all of the pages of the PDF documents.

- There were 5 different formats of documents consisting of engineering drawings, nested tables, un-demarcated tables, etc. This requires model creation for each of the document format.

Objective

To leverage teX.ai for the automated text extraction process with an accuracy target of over 80% and requiring less than 50% of the current time taken

teX.ai ™

Solution Overview

Quality File Validation

- Extraction of chemical composition file and Converting it to a key-value pair.

- These chemical composition type PDF documents were in 3 different formats which are 10 pages long

Survey Files

- Automatic identification of Survey(s) tables from multi-page documents followed by extraction.

Well Schematics

- Identify and extract the nested tables as separate entities. These documents had a combination of nested tables with complex drilling equipment's drawing.

Approach & Implementation:

teX.ai was leveraged to process text for all the 3 use cases

Quality File Validation

- The Analysis table which contained 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 more than 85%.

teX.ai ™

Input

Format 1

CHEMICAL ANALYSIS

Al	.46	B		<.004	Bi		<.00003	C		.022	Ca		<.002
Co	.08	Cr		18.6	Cu		.06	Fe		17.81	Mg		<.005
Mn	.05	Mo		3.02	Nb + Ta		5.12	Ni		53.7	P		.006
Pb	.0002	S		.001	Se		<.0003	Si		.09	Ti		1.01

Key : Element   Value : Composition Data

HEAT TREATMENT

Format 1

Format 2

Chemical Analysis

Key : Element

C	%	Si	%	Mn	%	P	%	S	%	Al	%	B	%	Bi	%	Ca	%
0.025		0.1		0.07		0.008		<0.001		0.52		<0.004		<0.00003		<0.002	
Co	%	Cr	%	Cu	%	Fe	%	Mg	%	Mo	%	Nb	%	Ni	%	Pb	%		
0.14		18.4		0.06		18.06		<0.005		3.12		5.03		53.4		0.0002			
Se	%	Sn	%	Ta	%	Ti	%	NbTa	%										
<0.0003		0.001		<0.01		0.93		5.04											

Value : Composition

Combined Element Codes

Format 2

Format 3

Elements	UOM	Method	
Ni → Key : Element	%	XRF	53.8 → Value : Composition
Cr	%	XRF	17.8
Fe	%	XRF	BAL
Nb+Ta	%	XRF	5.02
Mo	%	XRF	2.89
Ti	%	XRF	0.94
Al	%	XRF	0.50
C	%	CS	.021
Co	%	XRF	0.30
Mn	%	XRF	0.07
Si	%	XRF	0.04

Format 3

Output:

File   Edit   Format   View   Help

```
{
    "Al": "46",
    "B": "<,004",
    "Bi": "<,00003",
    "C": "022",
    "Ca": "<.002",
    "Co": "08",
    "Cr": "186",
    "Cu": "06",
    "Fe": "17.81",
    "Mg": "<.005",
    "Mn": "05",
    "Mo": "302",
    "NbTa": "512",
    "Ni": "637",
    "P": ".006",
    "Pb": "0002",
    "Se": "<.0003",
    "Si": "09",
    "Ti": "1.01"
}
```

Key-Value Pair Output in JSON Format

teX.ai ™

Public Files (Surveys):

- First isolated the survey tables using the keyword search leveraging OCR.
- Survey details are then extracted using techniques such as Tabula or Camelot.

Input

Sample of the Survey Data Format

Output

Output of the Extracted Survey Data

teX.ai ™

Business Impact

4x faster text extraction from the source documents, by leveraging teX.ai in the automated process flow.

The need for human intervention was reduced by over 80%.

The quality of their process had increased by over 75%.

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 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 complete, Indium deployed teX.ai with an admin interface built using Flask and containerization using Dockers.

teX.ai ™

Get in touch to see how
teX.ai can help you!

✉ info@tex-ai.com

📞 +1 (888) 207 5969 (Toll-free)

🌐 www.tex-ai.com

📍 Suite 100, 19925 Stevens Creek Blvd,
Cupertino, CA — 95014

teX.ai ™