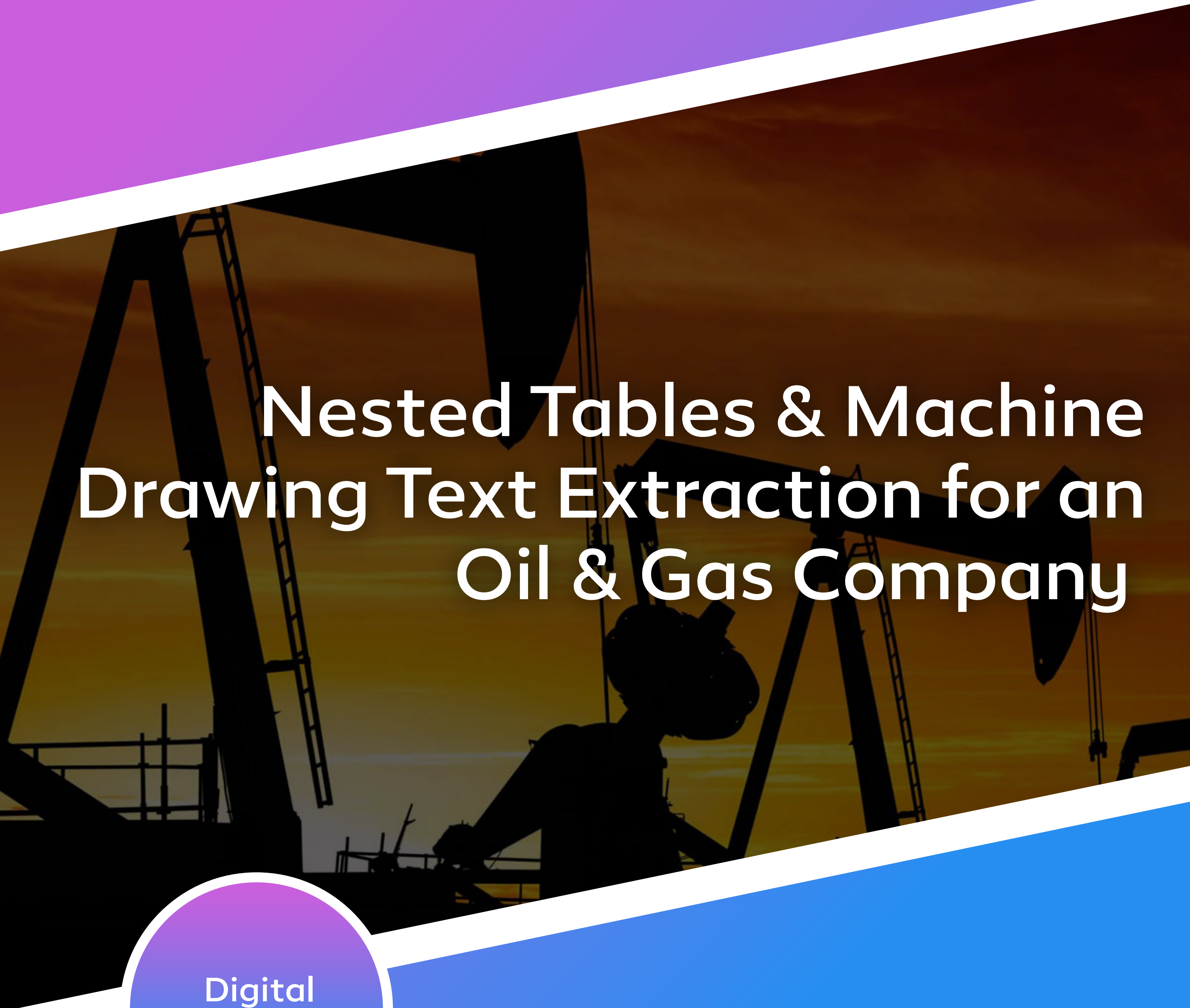

Nested Tables & Machine Drawing Text Extraction for an Oil & Gas Company

Digital Services

Success Story

www.indiumsoftware.com

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 Requirements

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.

Solution Overview

- **Quality File Validation**
  o Extraction of chemical composition file and Converting it to a key-value pair.
  o These chemical composition type PDF documents were in 3 different formats which are 10 pages long.
- **Survey Files**
  o Automatic identification of Survey(s) tables from multi-page documents followed by extraction.
- **Well Schematics**
  o Identify and extract the nested tables as separate entities. These documents had a combination of nested tables with complex drilling equipment's drawing.

Domain

Oil & Gas Industry

Technologies

The solution was built leveraging Python and several of its libraries.

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

- 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%.

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%.

Input

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

HEAT TREATMENT

Key : Element

Value : Composition Data

Format 1

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

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

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

14926.58	88.570	331.430	7059.68	7112.18	11147.46 N	5320.09 W	5985998.31 N	366556.20 E		12351.64	Tie-On Point
14991.00	91.060	328.760	7059.89	7112.39	11203.29 N	5352.20 W	5986054.68 N	366525.04 E	5.667	12415.89	Open Hole ST Point
15022.03	89.050	329.480	7059.86	7112.36	11229.92 N	5368.13 W	5986081.58 N	366509.58 E	6.881	12446.80	
15117.38	90.900	332.790	7059.90	7112.40	11313.41 N	5414.15 W	5986165.85 N	366464.99 E	3.977	12542.00	
15210.07	95.850	331.060	7054.45	7106.95	11395.02 N	5457.68 W	5986248.19 N	366422.87 E	5.656	12634.43	
15308.02	95.300	331.430	7044.93	7097.43	11480.49 N	5504.57 W	5986334.45 N	366377.44 E	0.676	12731.79	
15356.00	95.300	331.430	7040.50	7093.00 -	11522.45 N	5527.42 W	- 5986376.79 N	366355.31 E	0.000	12779.51	Projected Survey

All d[redacted] to True North.

[redacted]

Comments

Sample of the Survey Data Format

Output

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0														
1														
2														
3														
4	14926.58	88.570	331.430	7059.68	7112.18		11147.46 N	5320.09 W	5985998.31 N	366556.20 E		12351.64	Tie-On Point	
5	14991.00	91.060	328.760	7059.89	7112.39		11203.29 N	5352.20 W	5986054.68 N	366525.04 E	5.667	12415.89	Open Hole ST Point	
6	15022.03	89.050	329.480	7059.86	7112.36		11229.92 N	5368.13 W	5986081 .58 N	366509.58 E	6.881	12446.80		
7	15117.38	90.900	332.790	7059.90	7112.40		11313.41 N	5414.15 W	5986165.85 N	366464.99 E	3.977	12542.00		
8	15210.07	95.850	331.060	7054.45	7106.95		11395.02 N	5457.68 W	5986248.19 N	366422.87 E	5.656	12634.43		
9	15308.02	95.300	331.430	7044.93	7097.43		11480.49 N	5504.57 W	5986334.45 N	366377.44 E	0.676	12731.79		@
10	15356.00	95.300	331.430	7040.50	7093.00	-	11522.45N	5527.42 W	~ 5986376.79 N	366355.31 E	0.000	12779.51	Projected Survey	
11	All data is in	Survey) unless otherwise				and coordinates are relative								

Output of the Extracted Survey Data

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.

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%.

INDIA

Chennai | Bengaluru | Mumbai
Toll-free: 1800-123-1191

USA

Cupertino | Princeton
Toll-free: 1 888 207 5969

UK

London

SINGAPORE

+65 9630 7959

General Inquiries
info@indiumsoftware.com

Sales Inquiries
sales@indiumsoftware.com