

The logo for INDIUM, featuring the word "INDIUM" in a bold, white, sans-serif font. The letter "I" is stylized with an orange triangle pointing to the right, and the letter "D" has a small orange square at its top-left corner. The background of the entire image is a blurred industrial setting with orange robotic arms.

**INDIUM**

---

**Automating Multilingual  
Text Extraction:  
How teX.ai Streamlined  
Multilingual Data Extraction  
for Manufacturing Excellence**

SUCCESS STORY

## Client Overview

The customer is a multinational corporation with a rich legacy spanning over a century. Originating in Germany, the company has grown to serve a global clientele, with a presence in 120 countries and regions. Their innovative technologies and products are widely accessible, and they specialize in delivering goods for the Adhesive Technologies, Beauty Care, and Laundry & Home Care industries.

# Overcoming the Compliance Conundrum: Tackling Manufacturing Challenges in a Global Landscape

Maintaining impeccable quality and adherence to standards is paramount in high-stakes manufacturing. The client encountered a series of challenges that tested their precision and efficiency in a complex, multilingual setting.

- ▶ **Meticulous Manufacturing Process:** To maintain optimal quality and efficiency, the client had to ensure that every product adhered to a meticulous manufacturing process that required precise calculations and the integration of numerous chemical components.
- ▶ **Certificate of Analysis (CoA) Compliance:** Each chemical component was accompanied by a specific Certificate of Analysis (CoA) that needed to be accurately verified against established industry standards. This rigorous compliance check was essential for upholding product integrity.
- ▶ **Managing High-Volume Documentation:** The operation was challenged by the need to manage up to 1,000 PDF CoA documents daily. Handling and verifying this large volume of files manually imposed significant scalability challenges on the team of expert chemical engineers.
- ▶ **Navigating Multilingual Barriers:** The verification process was further complicated by the multilingual nature of the CoA documents, which were produced in English, German, Mandarin, and Thai. This linguistic diversity added an extra layer of complexity to ensuring uniform compliance across all documents.



## Innovative Business Blueprint: Empowering Effortless Document Extraction

Our client envisioned a system that could revolutionize the way data is extracted and processed from complex documents. Below are the key business requirements that laid the foundation for a next-generation document extraction solution:

- ▶ **Text Data Extraction:** The system needed to extract text data from PDF documents, seamlessly handling scanned images and digital PDFs.
- ▶ **Multilingual Support:** The extraction process had to manage multiple languages with different scripts, ensuring that a single PDF document could contain text in English, Thai, German, Mandarin, and other languages.
- ▶ **Tabular and Peripheral Data Extraction:** Besides standard text extraction, the system was required to extract tabular data and peripheral information from the documents, capturing all relevant details.
- ▶ **Front-end Web Application:** The client sought a user-friendly web application that easily allowed them to upload and edit documents.

# Revolutionizing Document Intelligence: The teX.ai Automated Extraction Breakthrough

Indium's comprehensive solution using teX.ai addressed a myriad of challenges, transforming complex PDFs into actionable data with precision and efficiency.

## **Multilingual & Redundant Data Handling:**

The documents featured multiple languages beyond English, non-readable headers and footers, and redundant data that needed to be filtered out. This complexity was effectively addressed to ensure clean and accurate data extraction.

## **Text Extraction Initiation:**

The text extraction process was kicked off using teX.ai, which converted the entire document into a uniform text format. This foundational step sets the stage for all subsequent data processing tasks.

## **Image Conversion with Tesseract OCR:**

Tesseract OCR was deployed to transform all images embedded in the PDFs into editable text. This conversion bridged the gap between scanned visuals and digital text, ensuring no information was lost.

## **Neural Network Object Detection:**

Advanced neural network algorithms were utilized to accurately detect and draw bounding boxes around key elements within the PDFs. This precise object detection ensured that only the required data was captured during extraction.

## **Multilingual Table Extraction:**

Tesseract, Tabula, and Camelot were successfully integrated to extract table data in non-English languages such as German, Mandarin, and Thai. This multi-tool approach guaranteed that structured data was accurately parsed across diverse languages.

## **Interactive Front-End Application:**

Indium's Product Development team developed an interactive web application that automatically processed PDFs from a dedicated email account. The application also measured OCR confidence levels, providing transparent and reliable output.

## **Versatile Data Delivery:**

Outputs were provided in both CSV and XML formats, offering flexible options for users to view and download data effortlessly.

## **Empowered Admin Control:**

Clients with admin access were granted the ability to edit and save outputs in their respective formats, ensuring seamless integration and customization.

The background of the slide is a photograph of a modern industrial factory. In the foreground, a large orange robotic arm is positioned on the right side, extending towards the center. The arm is mounted on a black base. In the background, other robotic arms and industrial equipment are visible, along with a high ceiling featuring skylights. The overall scene is brightly lit, suggesting a clean and active manufacturing environment.

## Business Impact in Action: Accelerating Manufacturing Efficiency with teX.ai

- ▶ **Accelerated Performance: 75% Reduction in Extraction Time**

teX.ai drastically cut the data extraction time from PDF files, ensuring faster turnaround and enhanced operational efficiency.

- ▶ **Enhanced Accuracy & Seamless Validation**

The text extraction process achieved superior accuracy, outperforming the client's legacy method. Clients could effortlessly validate outputs via the frontend application, comparing input and output side-by-side.

- ▶ **Adaptive Learning for Continuous Improvement**

The AI model continuously enhances its performance by incorporating feedback and improved edits from expert Chemical Engineers, leading to ongoing improvements in both accuracy and extraction quality.



## 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.

### USA

Cupertino | Princeton  
Toll-free: +1-888-207-5969

### INDIA

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

### UK

London  
Ph: +44 1420 300014

### SINGAPORE

Singapore  
Ph: +65 6812 7888

[www.indium.tech](http://www.indium.tech)



For Sales Inquiries  
[sales@indium.tech](mailto:sales@indium.tech)



For General Inquiries  
[info@indium.tech](mailto:info@indium.tech)

