CASE STUDY

DATA DIGITIZATION USING AZURE OCR FOR A LEADING DIAGNOSTICS FIRM

Enabled scientists with access to 99.5% accurate digitized documents of manually-processed records

Client Background

Our client is a leading global diagnostics company that manufactures a wide array of innovative medical diagnostic assays. The company had about 50,000 Batch Record documents in scanned format from 8 product families. Each product family had 2 types of documents i.e. PBR (Production Batch Record) and FLR (Filling and Labelling Record). The client wanted a solution to digitize and index this information to equip scientists with real-time access to critical data records for quick statistical analysis. The key objectives included:

- Make PBR (Production Batch Record) and FLR (Filling & Labelling Record) documents searchable
- Extract defined entities from scanned Batch Record documents, including enhancements for managing handwritten entities
- Store extracted data in DB for indexing purposes that can be easily retrieved for analysis
- Perform validations to ensure that the extraction accuracy is within the desired range
- Build UI for business applications supporting proofreading feedback

Xoriant Solution | Key Contributions

Xoriant developed an OCR tool using Azure OCR to make reports searchable for clinical scientists with limited IT knowledge. We performed data

KEY BENEFITS

- 99.5% accuracy levels achieved in digitized documents with our Data Quality Improvement efforts

classification, data extraction, and the necessary quality checks to convert the manual records into accurate digital records. Overall, 200+ document versions were created across 3 document families. 40000+ documents were to be scanned and converted. Xoriant team delivered approximately 500 documents in the first phase of the project. Our efforts enabled scientists with access to manual records in digital format for seamless analytics, querying, integrations, and decision-making.

Handled document classification and data extraction using Azure OCR and AI-based software: Xoriant team studied a set of client's documents from diverse families. With Azure OCR, we built the OCR tool in a re-usable format, on time, and on budget. We performed machine learning-based classification to identify each document's type, version, page number. Then, we used Xoriant's AI-based proprietary tool to understand the document structure and for data extraction of printed text by applying OCR techniques. A digital document builder was used for bringing all extracted information in a structured JSON format. In addition, we created a No-SQL DB storage of the extracted information. Each document was digitized using specific configurations.

Performed quality checks for data accuracy: We conducted a series of data accuracy checks. Our AI-based proprietary solution was used for predicting the attributes that required corrections. The outliers were checked and rectified manually to ensure maximum data accuracy in the digitized documents.

Developed UI for quick document access: Our experts created an API and a simplified User Interface for fetching information stored in the database. We provided scientists with the ability to access the rendered form of the actual data in a familiar format. The extracted data can be exported by the scientists in an Excel or PDF format for further analysis.

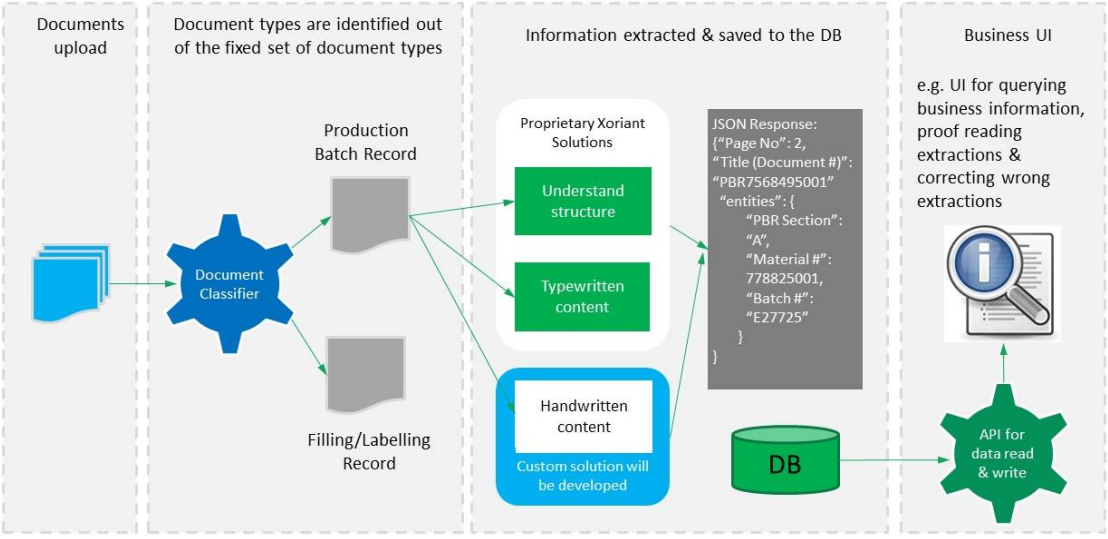
KEY BENEFITS

- Equipped scientists with the ability to quickly search historical documents and visualize data for making product improvements
- Minimized information search time from hours to seconds and improved investigation efficiency with a centralized data storage and retrieval system
- Enabled better quality management in the manufacturing process by equipping scientists with better analytical abilities
- Minimized resources needed for data mining with digital transformation and allowed more resources (scientists) to work on investigations
- Increased customer satisfaction with accelerated product development and improvement in the overall release quality

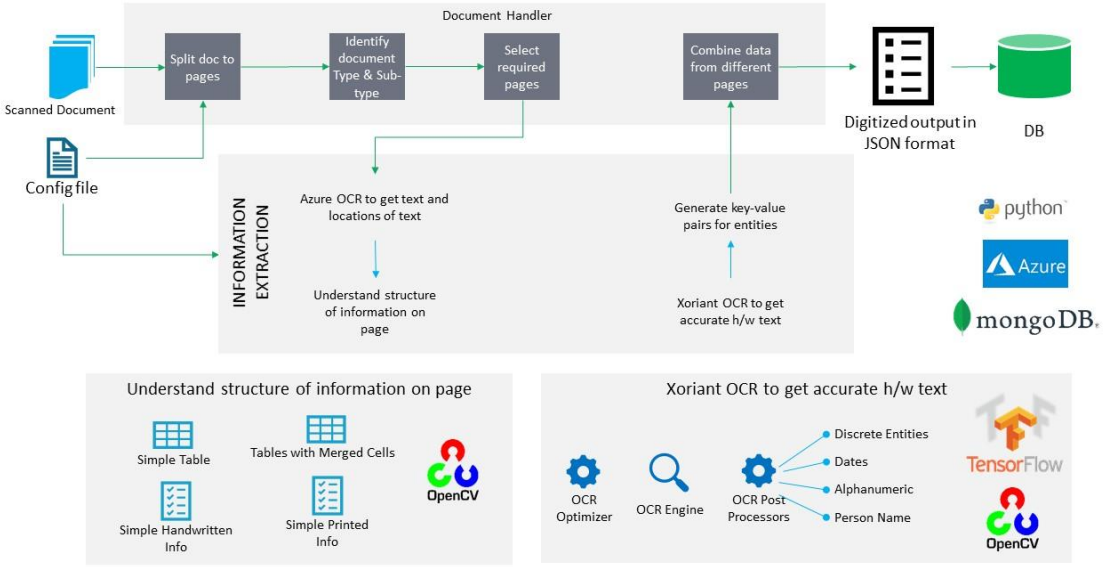
Client Testimonial

Xoriant enabled our team to access and analyze millions of data entries by resolving the complexities in this challenging project with their remarkable tech expertise. Instead of spending hours sifting through thousands of pages, our scientists can now find specific information in seconds. They use the saved time to complete their investigation and extend their analysis.

High Level Architecture



Xoriant Proprietary Tool for Document Digitization & Indexation



Technology Stack

AWS (S3, Redshift, EC2) | SFDC | Oracle Fusion | Tableau | CloudIO (FLOW) Python | Computer Vision (OpenCV) | Object Detection (inception model) | Image Processing (OpenCV) | Tesseract OCR | Azure OCR | Handwritten text detection models | No-SQL DB (Elasticsearch/MongoDB) | Virtual Machines on Azure | Kubernetes and containers

Xoriant is a product engineering, software development and technology services company, serving technology startups as well as mid-size to large corporations. We offer a flexible blend of onsite, offsite and offshore services from our eight global delivery centers with over 3600 software professionals. Xoriant has deep client relationships spanning over 30 years with various clients ranging from startups to Fortune 100 companies.