Application Development Services | CASE STUDY

Rules Based OCR Solution

telliant
Intelligent Software. Delivered.

"Client was seeking a partner to design and develop a sophisticated rules based OCR solution that could be easily integrated to their existing suite of ECM Products...."

The Client

The Client is an enterprise content management (ECM) software company providing document management, archival, electronic signature, and electronic receipt management products to the financial services industry. Their products address the challenges of the regulatory and compliance requirements. They offer the latest technologies in electronic document management including high-speed scanning, workflow management with high level encryption. They leverage the web-centric Microsoft.NET technologies to deliver browser based document management functions to their clients.

The Business Challenge

The client was looking for an application development partner to integrate their existing enterprise content management software with Optical Character Recognition (OCR) technology. While their product offers sophisticated solutions to scan, store and retrieve documents, their products lacked the ability to perform character recognition which is fast becoming a must-have technology for ECM systems. Having OCR technology integrated into their product will enable the customer's products to become much more competitive and offer outstanding value to end users.

The customer was interested in the following functionality for their solution:

- **Interface to an OCR Engine:** The customer identified ABBY FineReader as the OCR Engine of choice
- **Administration Component:** An application to maintain the required configuration parameters
- **Rules Processing Component:** A .NET API component to access the rules and call the OCR Engine
- **Fuzzy Search:** The rules processing engine should apply fuzzy search to find and extract the data
- **Hosting the OCR Engine:** The OCR Engine will be hosted on a multi-core environment running concurrently

Telliant was chosen by the Customer due to the following factors:

- Experience with developing ICR/OCR technologies and ECM solutions
- Proven track record in .NET Development
- Understanding the success factors related to outsourcing
- Provide outstanding problem solving skills
- Offer professional and personal customer service

They chose the ABBY FineReader tool for integration to increase their ECM capabilities. The current performance of their search engine lacked the speed and accuracy that their customers expected to access, search, and archive their documents. With increased use of pdf, digital, video and image documents, their ECM needed additional functionality and features enabled by the OCR to increase performance, with security at the forefront of the new application.

> "The challenge was to create a set of applications that performed at peak efficiency while providing a vast range of functionality that was easy to use and administer."

> "Our services included full lifecycle software development from analysis, design, development through testing and implementation."

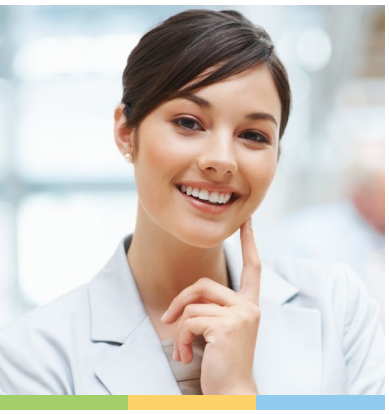
The Engagement

The Client chose the Telliant team for its demonstrable experience in building high performance applications and for its proven track record in forming a true development partnership.

The engagement included:

- Gathering and Analyzing the client requirements.
- Assembling an experienced and talented offshore delivery team.
- Following a modified Agile Project management with regular status updates.
- Design, development, testing, performance benchmarking, tuning of the application.
- Seamless interaction with the Client 's development team.

telliant
Intelligent Software. Delivered.



TECHNOLOGY ENVIRONMENT

Major Technology Components:

- Microsoft .NET 3.5
- VB.Net
- SQL Server 2008
- J-Query
- CSS, JavaScript
- AJAX
- WCF
- Crystal reports

Third-party tools:

- ABBY FineReader OCR SDK
- Infragistics Grid
- PDF Rasterization component

Corporate Office:

Telliant Systems
3180 North Point Pkwy
Suite 108
Alpharetta, GA 30005
USA

Tel: 678.892.2800
Fax: 678.892.2809
Email: info@telliant.com

Solution Highlights

The overall solution included a very user-friendly configuration application and high-performance rules based OCR processing engine interfacing to the ABBY FineReader OCR SDK. The following are the key highlights of the solution:

Configuration Elements: The various configuration elements include a processing plan with all the settings required for processing a particular group of documents using OCR. Users can create a variety of document types and the corresponding processing plans. An attribute rule is used to extract an individual value from the OCR text results. An attribute represents a piece of metadata about the document that can be found within the document's text. Validation rules improve the accuracy of extracting values from OCR text using the attribute rules. Validation logic for date, currency, and numeric attribute types is automatically applied.

Web Based Designer Application: ASP.NET web based application for maintaining the various configuration elements of the OCR Solution. The application was designed to be visual and user-friendly with a 'test-first' approach and allows definition of the various fields within a document that needs to be recognized. This is achieved by uploading the scanned image samples for the various types of documents.

Performance And Accuracy Charts: In addition to the test-first approach of the designer, the user has the ability to test the accuracy and performance of an entire processing plan. The system creates bar graphs to illustrate the statistics of each. Developed using Crystal Reports, these integrated performance metrics allow the Client to showcase their products capabilities.

OCR Rules Processing Component: The OCR rules processing component provides a .NET API for performing OCR on a document and extracting values from the text. The component loads the rules for parsing text values from the relational database, calls the ABBY FineReader OCR Engine to perform character recognition, and then parses the resulting XML text into a set of attributes based on the rules. The attribute values are then returned to the caller.

API Public Interface: This API accepts the document bytes, ProcessingPlanID, DocumentTypeID (if known) as parameters, and returns the results as a .NET DataSet. This Dataset contains the document type or attribute information that was extracted as a result of the processing. This information is used in the external Client application to automatically apply the corresponding attribute rules. The API calls are done through WCF calls (Windows Communication Foundation).

Re-Analyze Function: With this feature, the user has the ability to test the accuracy and performance of the entire processing plan. By selecting the "Re-analyze" button, the user recreates the statistics for both accuracy and performance for a selected ProcessingPlan or all the ProcessingPlans and store these results in their respective tables in the database. It may take several minutes to reprocess and OCR all the images/samples/document in the processing plan, an indicator for viewing progress is included.

Fuzzy Search Algorithm: The rules processing engine employs fuzzy search logic for matching text when searching for a particular label in the OCR results. This allows the phrase to be found within the text even when the OCR engine has imprecise recognition results. To determine if a word with the OCR text is a match, its similarity to the phrase word is quantified using the Levenshtein distance algorithm.

PDF Rasterization Component: The PDF rasterization component utility enables implementation of varied settings and properties for output images such as image type, size, resolution, and quality of images. Define PDF page range to be converted freely. This component converts to images all PDF pages, odd pages, even pages and, defined page ranges with ease and success.

Results Achieved

The benefits achieved by the Client included:

- Completion of the project on time with no outstanding issues
- Seamless project management and interaction of Client and offshore teams
- Delivered a sophisticated solution that is easy to configure and administer
- Increased market value of the products with the addition of the OCR technology
- Easy integration of the OCR feature with existing products