



建設情報プロバイダCMDが、ABBYYの拡張性の高いOCRソリューション導入で、大規模な建設文書処理を効率化

企業紹介



企業名
CMD社について

所在地
米国オハイオ州

業界
建設業界

企業URL
www.cmdgroup.com

企業概要

大規模な量の建設文書を処理できる自動化されたOCRによる、業務の効率化とデータの質の向上。

事例概要

ABBYY RECOGNITION SERVERを導入することで、CMDは、複雑で大きいフォーマットの企画書を含む、何百万にものぼる建設文書のキャプチャおよび変換を自動化。

Construction Management Data社(以下CMD)は、ウェブサイト経由で数十万にも及ぶ過去、現在、未来の建設プロジェクト情報を建設会社へ提供している北米の建設情報プロバイダです。このサービスを提供するにあたり、年間3,500万ページもの複雑な紙文書やデジタル文書を、検索可能な企画書や仕様書に変換する必要があります。しかし、CMDの、検索可能なPDFを作成するソリューションは、手動の文書作成が必要で大規模な建設計画書に対応できていませんでした。その上、現在の文書を処理する仕事量が増え続ける一方で、何年もの過去のデータを処理して検索可能にする必要もありました。そこでCMDは、プロセスの効率化、自動化および拡張性を成し遂げる方法を模索し、ABBYY Recognition Serverに解決策を見つけました。

「ABBYYは、我々の顧客に検索可能な文書を提供すること、そして内部で文書処理を自動化することにおいて、弊社の事業のビジョンに気付かせてくれる重要な要素でした。」

-CMDデータ戦略部副社長 Dan DuBois氏

大規模な文書変換や大きなフォーマットの文書処理に、堅固で拡張性のあるOCRが必要

北米の建設会社は、マーケットリサーチが必要な時や、ビジネスになる新しいプロジェクトを探したり入札動向の情報が必要な時は、まず最初にCMDから情報を得ます。CMDのウェブサイトは、非住宅商業建設プロジェクトを主に扱っており、加入メンバーは膨大な検索可能なプロジェクトドキュメントを含む知識ベースにアクセスできます。「初期の建設計画から契約締結までプロジェクトのあらゆるステージのデータを収集しています。」とCMDのデータ戦略部副社長であるDan DuBois氏は説明します。「我々の調査員達は、紙やデジタルなど多様なフォーマットで情報を入手するので、知識ベースに統合するためにOCRでそれらを検索可能なPDFに変換する必要があります。」

CMDが年間で処理する文書は膨大な量です。およそ3,500万ページから成る10万以上の文書セットを処理しています。そして各ページはテキスト、図表、図面、絵が組み合わさっている複雑な文書です。また、紙文書もあるので、紙の物理的容量は非常に大きくなることもあります。CMDのシニアソフト開発者のVic Mykulowycz氏は、以前使っていたOCR技術は建設業界には不十分だったと言います。「例えば、大きなフォーマットにも対応できませんでした。我々は44 X 36インチにも及ぶ何百万もの建設計画書を処理するので、データベースから手動で情報を引き出したり準備したりする必要がありました。更に、以前のソフトウェアの精度は、処理しきれないほどの手作業での検証を必要としました。自動化によってプロセスを合理化して手作業を最小化する必要がありました。」

大規模なOCRソリューションを検討

テキストと画像のどんな組み合わせであっても正確なテキスト検索ができることが、GMDの新しいOCRソリューションの不可欠な条件でした。これを考慮して、GMDは業界シェアの高いいくつかのソリューションの試用ライセンスを取得し、一連の対照比較テストを行いました。「我々は同じページを各ソリューションで試しました。どれぐらい正確にテキストを出力したのかや各ページのエラー

ーを数えながら細かく比較しました。中でもABBYY Recognition Serverは特に正確でした。スピードも優れていましたし、何より、大きなフォーマットのページも処理することができました。また、処理する文書の量を考慮すると、GMDI 社BBYY Recognition Serverに厳しいテストをする必要がありました。そこでOCRと文書管理ソリューションの専門家として知られ、ABBYYのパートナーでもあるConarcに拡張ライセンスを手配してもらいました。」

「我々の文書変換処理は自動化されたおかげで手作業での処理が最小化されました。」

-CMDのシニアソフトウェア開発者
Vic Mykulowycz氏

ABBYY Recognition Serverの導入

Conarcのシニアソフトウェア開発者のBen Holton氏は、「CMDの膨大な情報量のため、通常よりも遙かに多いページに対応した試用ライセンスを手配しました。」と言います。拡張ライセンスを取得したCMDは11の物理サーバーにRecognition Serverをインストールし稼働させました。

Mykulowycz氏は次のように述べています。「我々はソリューションが実際どこまでできるか、11つのサーバーを運用するコアライセンスの最適形態は何なのかを知る必要がありました。広範囲なテストを行った結果、導入時には社内ソフトウェアとABBYYソリューションの技術的な違いからRecognition Serverを調整する必要があるとわかりました。ConarcとABBYYは迅速なサポートを提供してくれた上に、ABBYYは我々の問題に対処するため、特別な単発リリースさえ作成してくれました。」

最初にCMDはRecognition Serverの72コアライセンスを購入しました。1つのサーバーで全ての分散されたCPUを処理するというもので、結果、CMDの期待を満たしていました。しかし、その1年以内に同社は、過去に変換されていない資料からもアーカイブを作成することを決めました。Ben Holton氏は「それをするにはかなりのリソースが必要でした。さらに昨今の建設業界の回復を考慮すると、ゆくゆくは現在の要求を満たすためだけにそのリソースが必要だと予想されました。それは膨大なスケールアップ(拡張)が必要なことを意味していました。」

更なる拡張: 156コアのOCRソリューション

Mykulowycz氏によると、新しいプロジェクトは、1.5年の過去のデータを検索可能なPDFに変えることが必要で、それは膨大な作業量でした。「現在の資料の変換に加え、4,500万余りのページを処理する作業負荷を考えると自動化されたOCRが必要でした。」

過去のデータ変換を遂行するために、CMDは追加で88コアライセンスを購入し、そのプロジェクトに専念しました。ある程度までは予定通りに完了したのですが、2つの要素が浮上してきました。

Mykulowycz氏によると、建設は季節的な事業だといいます。「2月から6月が非常に忙しく、文書処理のほとんどが発生する時期です。加えて、建設業界の景気回復も考慮しておらず、実際に必要な作業量を見積もっていなかったことを意味しました。」

急速に数が増大する計画書や仕様書に対応するためCMDは1つのマスターCPUに全てのコアライセンスを結合しました。出来上がったソリューションは、1つのシステムから成り、合計156のネットワーク化されたコアライセンスを複数のサーバーに割り当てています。「1つのサーバーマネージャーに全てを結合したことは成功でした。」とMykulowycz氏は言います。

結果

導入から7カ月で、Recognition Serverを使ったCMDのソリューションは、過去の文書変換プロジェクトを完了しました。そして今度はそのリソースをメインのシステムに移行しました。

「システムは現在、更にスピーディーに文書を処理するようになりました。」とMykulowycz氏は言います。「我々の文書変換処理は自動化されたおかげで手作業での処理が最小化されました。もう1つの利点はRecognition Serverは単語の検索も簡単にできることです。テキスト検索する際、単語の周りを囲んだ赤いボックスが現れ、結果をハイライトしてユーザーに示してくれます。」

ABBYY

ABBYYジャパン株式会社
〒222-0033
神奈川県横浜市港北区新横浜2-3-12 新横浜スクエアビル14階 T 045.285.2410

www.abbyy.com/ja/