Success Story
Building and Construction

ABBYY®

Robust OCR service raises productivity for leading provider of construction data

Construction Management Data (CMD) provides information to construction firms on hundreds of thousands of past, current and future construction projects via their website. This requires collecting and converting complex paper and digital documents into searchable plans and specification documents, for an approximate total of 35 million pages per year. But CMD's solution for creating searchable PDFs required manual document preparation and could not handle large-scale construction plans. Additionally, it couldn't scale to meet their need to process and make searchable, years of past data, while accommodating growing volume and workloads related to current projects. Seeking a way to achieve higher productivity, automation and scalability, CMD found the solution in ABBYY Recognition Server.

> *"ABBYY was key to realizing our product vision – in terms of providing our customers with searchable documents, and automating our document processing internally."*
>
> *Dan DuBois, Vice President – Data Strategy, CMD*

Large-scale conversion and oversized documents demand robust and scalable OCR

Whether they're seeking business intelligence regarding their market, hunting for new project leads or searching for insights into creating bids, construction firms across North America look first to CMD for the information they need. Focused on non-residential commercial construction projects, CMD's website provides subscribers a knowledge base that includes a vast repository of searchable project plans. "We collect data on every stage of a project – from early planning, through contract award," explains Dan DuBois, Vice President – Data Strategy, CMD. "Our researchers acquire it in a variety of formats, paper and digital, and using OCR we convert it to searchable PDFs for integration into our knowledge base."

As DuBois says, the volume and complexity of the documents that CMD converts every year is enormous: "Over 100,000 document sets comprising approximately 35 million pages are processed every year – and each page can combine text, schematics, drawings and illustrations." The physical dimensions of many documents can also be very large. And Vic Mykulowycz, Senior Software Developer, CMD, says the company's previous OCR technology was inadequate for the construction industry. "For example," he says, "it couldn't handle large formats, and we process millions of construction plans that are up to 44"x36". So people had to manually pull and prepare information for our database." Plus, the old software's precision required unacceptable levels of manual verification. "We needed," says Mykulowycz, "to streamline the process and minimize manual intervention through automation."

CMD

About CMD

Construction Management Data (CMD) is a leading North American provider of construction information. CMD's diverse portfolio of innovative products and services includes national, regional and local project leads, marketing solutions and market intelligence that provides insight to construction industry professionals throughout the U.S. and Canada.

Learn more at www.cmdgroup.com

Seeking a solution for implementing large-scale automated OCR…

For CMD, it was vital that their new OCR solution deliver accurate full-text searches for all documents – whatever combination of text and graphics. With this in mind they obtained trial licenses for various, highly competitive, solutions and began a series of side-by-side tests. “We processed the same pages across multiple solutions,” describes Mykulowycz, “and carefully compared the text output to see how accurate it was - counting the number of errors made by each one. ABBYY Recognition Server was more accurate, had better speed and, very importantly, could process large pages.”

Given the volume of document conversion, CMD needed to put ABBYY Recognition Server through severe testing. So they arranged for an extended license through Conarc Inc., an ABBYY Partner known for its expertise in OCR and document management solutions.

Finding the answer with ABBYY Recognition Server

“Because of CMD’s huge throughput,” says Ben Holton, Senior Software Developer, Conarc, “we arranged for a trial license with significantly larger page volume than normal.” Having acquired their extended license, CMD installed 11 new physical servers and began running Recognition Server over them.

“We needed to see what the real capacity of the solution could be,” recalls Mykulowycz, “and what the optimal configuration of core licenses running over almost a dozen servers would be. Extensive testing followed. And then, during implementation, we needed to make changes to accommodate Recognition Server because of technical differences between our in-house software and the new ABBYY solution.” According to Mykulowycz, Conarc and ABBYY were prompt in their support. “They were very responsive. ABBYY even created special one-off releases that addressed our issues.”

Initially, CMD purchased a 72-core license for Recognition Server – coordinating the management of all the distributed CPU processing through a single server. The results, according to Mykulowycz, “met expectations.” But within a year, the company decided to create a historical archive from previously unconverted material. And as Ben Holton says, “That required a significant boost in resources. Plus, the construction industry’s rebound meant we would eventually need to apply those resources to meeting current production needs –and require significant scaling.”

Scaling even further: Achieving a 156-core OCR solution with Recognition Server

According to Mykulowycz, the new project required converting one and a half years of historic data into searchable PDFs, a formidable task: “Processing those 35 million-odd pages of information while keeping up with our current production load demanded precise, automated OCR.”

To accomplish conversion of the historic data, CMD purchased an additional 88-core license and dedicated it to the project. This was completed on schedule; at which point two more factors came into play: “Our business is seasonal,” explains Mykulowycz. “From February to June we are really busy. It’s when most of our current document processing happens. That, plus a big upturn in construction, meant we weren’t seeing the throughput we needed.”

To handle the rapidly growing number of plans and specifications, CMD combined its license cores under a single master. The resulting solution, dedicated to current production, comprised a single system utilizing multiple servers with a total of 156 networked cores. “Combining everything under one server manager has proven successful,” confirms Mykulowycz.

The results

Within 7 months of implementation, CMD’s Recognition Server-based solution completed the archival conversion project, and the additional processing power was rolled over into the main production system. “The system now copes with documents much faster,” Mykulowycz says. “And our document conversion process has been automated to a degree where human intervention is minimized. Another benefit,” he says, “is that Recognition Server provides the coordinates of words on a page. When we search for text, a red box surrounds the words that come up – highlighting results for our users.”

And as Dan DuBois confirms, Recognition Server has helped CMD realize its long-term goals: “ABBYY was definitely a key component in realizing what our product vision was, our roadmap. Both in terms of being able to provide our customers with searchable documents, as well as automating our document processing internally.”

Learn more at www.ABBYY.com/recognition_server

The Challenge:

Raise productivity and data quality by implementing an automated OCR service capable of handling high volumes of construction documents.

The Solution:

Using ABBYY Recognition Server, CMD automated the capture and conversion of millions of construction documents, including highly complex and oversized format project plans.

“Our document conversion process has been automated to a degree where human intervention is minimized.”

Vic Mykulowycz
Senior Software Developer
CMD

Conarc
30000 Mill Creek Avenue
Suite 475
Alpharetta, GA 30022
Tel 770.849.0508
Fax 770.448.1425

www.Conarc.com

ABBYY
North American Headquarters
880 N. McCarthy Blvd.
Suite 220
Milpitas, CA 95035, USA
Tel 408.457.9777
Fax 408.457.9778
sales@abbyyusa.com

www.ABBYY.com

Conarc, Inc. offers powerful, customized software solutions. These include iChannel – a powerful, web-based, collaborative enterprise content management system that makes your organization’s information accessible by anyone you choose to share it with – at any time, from anywhere.