How contentCrawler improves the user experience of searching in Microsoft SharePoint

The business need

Enable full-text search across all documents within a SharePoint environment

Convert image-based PDFs, TIFF, and MSG files to searchable PDFs

Process newly-created documents added to SharePoint in real time, as well as a backlog of migrated content

Ensure all records are discoverable in order to comply with the Public Records Act and the Archives New Zealand Information Management Standards

About Manawatu District Council

Manawatu District Council serves a population of roughly 30,000 people in an area about two hours' drive from Wellington. Its main town, Feilding, has been awarded New Zealand's Most Beautiful Town 16 times thanks to its picturesque Victorian and Edwardian-style buildings. The job of the Council is to support local infrastructure, public services, and regulatory management systems.

"Having our documents processed by contentCrawler has made a massive difference to our users' search experience, allowing them to work more efficiently."

Mel Rush
Information Team Leader,
Manawatu District Council

Migrating non-searchable legacy documents to SharePoint

Manawatu District Council (MDC) had a problem. Its legacy document management system was no longer supported and this was causing major issues. Information Team Leader at MDC, Mel Rush, explained they "had times when it would just crash, and IT would struggle to get it rebooted again. We had one case where we lost a bunch of documents that had been scanned in – and that's just what we know about."

So, MDC switched to Microsoft SharePoint and Mel and her team began moving hundreds of documents over to that environment. But this brought with it a new set of problems.

"We were bringing nearly 200,000 documents from two older systems into the new SharePoint environment. One of our biggest concerns was people being able to find the content they were looking for," Mel explained. "Our legacy systems didn't play ball when we started our migration project, and some of the metadata did not align with the records." A lot of crucial information that gave the documents meaning was lost. "We had these arbitrary records sitting there, which you had to open in order to know what they were about."

One of the key selling points of SharePoint is its Google-style search technology that makes finding content easier with the use of filters. Mel and her team had been selling this feature as a real benefit to staff and knew non-searchable documents would have undermined its value.

SharePoint's search technology relies on metadata – of which MDC's legacy content had very little or none. "Therefore, we needed our documents fully text-searchable to allow staff to find them," Mel explained. "We also wanted these files to be converted to PDF to ensure all staff would be able to access them."

As well as impacting the value of its new SharePoint environment, non-searchable documents were a risk. "In order to be compliant with the Public Records Act and the Archives New Zealand Information Management Standards, we needed our records to be discoverable."

Mel and her team began looking for a solution to "OCR documents both within our new SharePoint environment and those migrated from legacy systems." Optical Character Recognition (OCR) technology analyzes image files for the presence of text and converts them to searchable documents.

"We had analyzed our SharePoint environment and knew there were a total of 76,061 files that needed to be processed, including PDF, TIFF, and MSG files," said Mel. Image-based PDFs, TIFFs, and MSG files do not have the text layer needed to be found by search technology. "The migrated content from legacy systems – nearly 200,000 files – were mostly TIFFS".

Researching text recognition solutions
Mel and her team began to investigate available solutions. "Cost is always really high on the list of what we need to accommodate," said Mel. "We were wary of buying something that would do what we wanted but also a hundred other things that we didn't need. We wanted to really focus on what our actual needs were, and so, when we began researching possible solutions through Google, contentCrawler came out on top because it was able to do exactly what we needed."

MDC trialed contentCrawler before making its final decision. "We were able to set contentCrawler up in our environment and have it process our actual content. By the time we'd deployed it fully, we had a good idea of what we were going to get out of it and how it was going to work."

How contentCrawler is making files searchable
contentCrawler is integrated with MDC's SharePoint environment, processing both new documents as they're added as well as any legacy documents. "Every time someone adds something new, contentCrawler will look at it within 24 hours. We also have a backlog running that is just chugging through that older content," explained Mel. "contentCrawler supports PDF, TIF, and MSG files and is able to compress our documents, which really suits our way of working."

A lot of building consent information was migrated to SharePoint as part of MDC's property file digitization project. "contentCrawler is looking directly at that site to OCR those files as quickly as possible."

For other organizations looking for better search
For other organizations looking for a way to enable effective search, Mel recommends reaching out to technology providers – like she did with DocsCorp – and ask questions.

"We weren't 100% sure if it was going to work for us, but DocsCorp had a great support system in place right from the start," said Mel. "DocsCorp not only wanted to help us use contentCrawler properly, they genuinely wanted to hear our feedback and use it to improve the product."

"Having our documents processed by contentCrawler has made a massive difference to our users' search experience, allowing them to work more efficiently."

Visit our website to learn more about contentCrawler and how it supports better search.