The Perfect Holiday Gift for IT – Enterprise Search

Article contributed
by dtSearch®

While we wrap up 2022 and look ahead to 2023, what is the IT department likely to hear? "Help! I can't find the file I need! I know it is out there. I needed it like, yesterday!" The gift every IT professional deserves this holiday season is enterprise search.

Enterprise search isn't a find-the-best-shoe-sale-on-the-Internet search engine like Google. Rather, enterprise search encompasses software such as dtSearch® for precision search *inside* an organization's emails, Office files, PDFs, web-based formats, etc. Instant concurrent search across terabytes of mixed data can run from a classic Windows network, a local web server, or a cloud server. (Online search can operate statelessly to scale seamlessly.) After a search, end users can browse the full text of retrieved items with highlighted hits.

But is there a catch? The catch is that to enable instant concurrent searching across terabytes, enterprise search first has to index the data. An index is just an internal mechanism for pre-processing each unique word, number and the location of each in the data. A single index can hold up to a terabyte and there are no limits on the number of indexes enterprise search can build or concurrently search.

But wait! I see the wheels turning: "OK, enterprise search would take a lot of burden off of IT, but doesn't indexing take a huge amount of effort?" Indexing is a lot of work, but for the software itself, not for IT. To kick off indexing, all IT has to do is point to the folders and the like to index, then let the software do the rest. In fact, below is a quick list of all the things that IT *does not* have to worry about with enterprise search indexing:

No worry #1: file type identification. Parsing each item requires identifying the correct file type and applying the right parsing specification. However, enterprise search can determine on its own the file type as it reviews each item in binary format.

No worry #2: mismatched file extensions. The binary version reveals the correct format, making the file extension irrelevant for purposes of determining file type. In other words, PDFs saved with .DOCX extensions and OneNote files saved with .PDF extensions are no problem.

No worry #3: multilevel nested content. Enterprise search can automatically handle recursively embedded formats, such as an email with a ZIP or RAR attachment containing a PowerPoint with an Excel spreadsheet fully embedded inside.

No worry #4: "hidden" content. This category includes obscure metadata that takes a lot of clicking around in a file's native application to discover. It also includes "hidden" text blending in with the background color inside a file's native application, such as white on white or black on black writing. Because enterprise search approaches files in their binary formats, all of the above is searchable as ordinary text.

No worry #1: file type identification.

No worry #2: mismatched file extensions.

No worry #3: multilevel nested content.

Reprinted with permission of IT Business Net, itbusinessnet.com

No worry #5: PDFs that require OCR. Some PDFs look like regular text-based PDFs but really consist of just an image for the main body. But enterprise search can flag these "image only" PDFs in the indexing process, letting IT know that these need to run through an OCR processor such as Adobe Acrobat to turn them into full-text searchable PDFs. Following OCR, enterprise search can display the complete original image with highlighted hits superimposed for easy navigation.

No worry #6: file additions, deletions or modifications. IT can set indexes to automatically update. Crucially, updating an index to reflect new content can proceed without affecting concurrent searching.

Indexing lets enterprise search perform more than 25 types of full-text and metadata search requests. These include simple search requests consisting of "all words" or "any words." These also include more structured search queries like *(holiday party or seasonal party) and (office party w/7 caterer) and not (entertainment services).*

Concept search extends any search term to synonyms of that search term. Fuzzy searching is adjustable from 1 to 10 sifts through typos. With a low level of fuzzy searching, a search for *party* would also pull up *parqy*. Stemming retrieves different word ending on the same route word, locating *parties* and *partying* for *party*. Metadata "only" search requires the presence of certain terms in specific metadata.

Date search can identify dates or date ranges even across different formats such as December 15, 2022 or 12/15/22. The software can also search for numbers or numeric ranges, and even generate hash values for indexed files and optionally search on these. Lastly, Unicode support enables searching in the hundreds of languages that the Unicode standard spans.

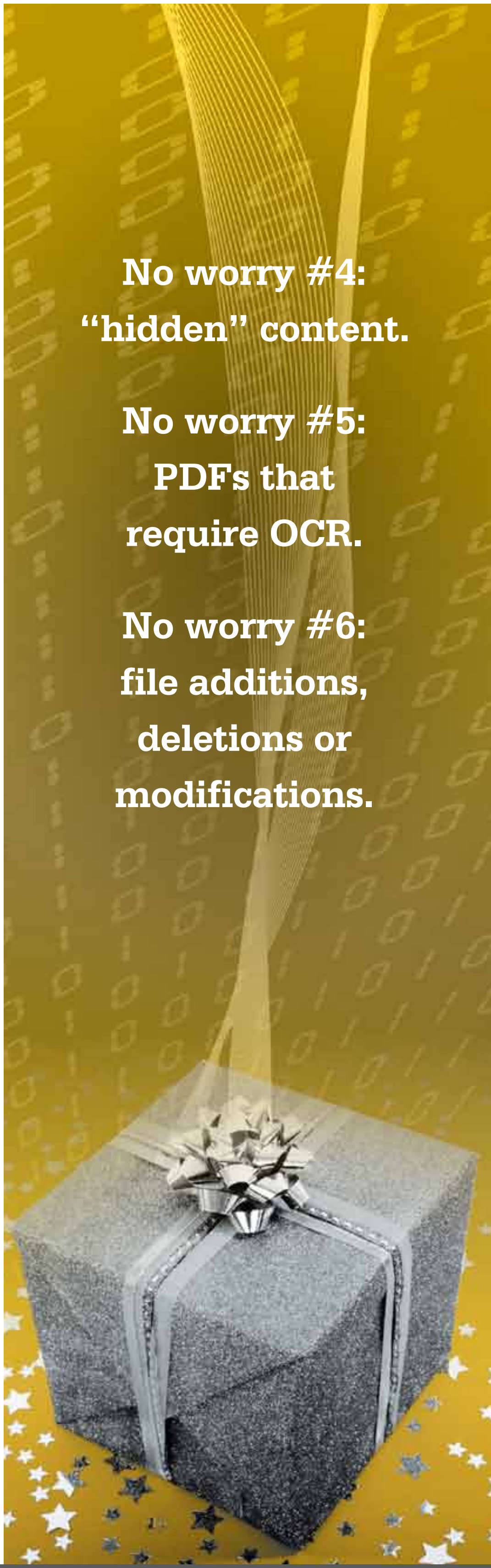
After a search, enterprise search applies vector-spaced relevancy-ranking. What that means is if *holiday* appears in millions of indexed files but *seasonal* only appears in a handful of files, then *seasonal* would get a much higher relevancy ranking. Files with the densest *seasonal* mentions along with mentions of other search terms would rank highest of all.

Supplementing the default relevancy-ranking, the software also supports positive and negative variable term weighting across the full text of files, in certain metadata only, or positionally at the top or bottom of files. End-users can also simply click to instantly re-sort by a completely different criterion, such as file data, file location or file size, for review with highlighted hits.

Special bonus: credit card identification. Relevant to paying for that holiday party, enterprise search can further find any credit card numbers that may have made their way into ordinary files. What the search engine does is run any series of numbers that might represent a credit card past an internally available credit card verification algorithm. The end result alerts IT to the presence of credit card data in "open" archives where these should not be.

And there you have it – now you know the perfect gift to give IT this year!

Article contributed
by dtSearch®



No worry #4: "hidden" content.

No worry #5: PDFs that require OCR.

No worry #6: file additions, deletions or modifications.

Reprinted with permission of IT Business Net, itbusinessnet.com