Separating the Wheat From the Chaff: Enterprise Search Edition

Article contributed
by dtSearch®

It's harvest time. While no one should count on me for actual farming tips, I can tell you how enterprise search can help separate the wheat from the chaff in your organization's data, letting you jump right to the most relevant kernels of information.

How does enterprise search do that?

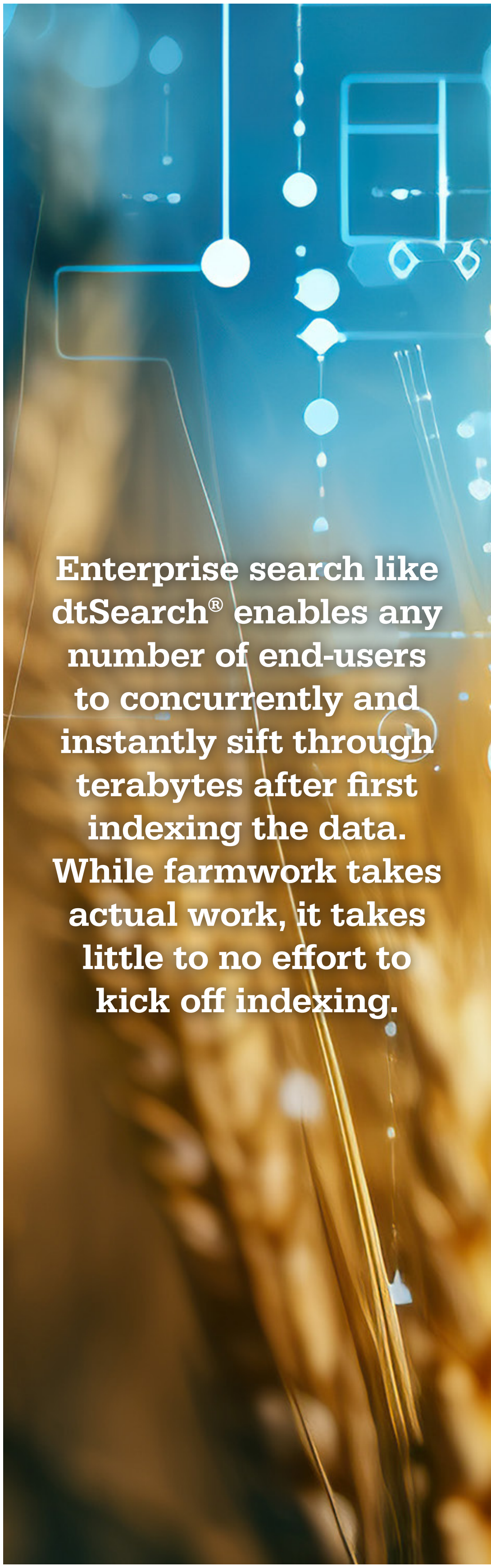
Enterprise search like dtSearch® enables any number of end-users to concurrently and instantly sift through terabytes after first indexing the data. While farmwork takes actual work, it takes little to no effort to kick off indexing. Just tell the software which folders, email archives and other data to index, and the indexer will take it from there.

Can the indexer work with cloud storage?

No need to tell the indexer if files are local or remote like One Drive / Office365, SharePoint or DropBox. So long as the indexer can see the items as part of the Windows folder system, the indexer can work with them. The indexer needs to determine the file format of an item to correctly parse it. However, the indexer can on its own figure out if an item is a PDF, PowerPoint, Excel, Access, Word, OneNote, Outlook, Exchange, etc. file, and handle it accordingly.

How does the indexer do that?

The indexer looks inside the binary format of each item to determine the file format. That way, a misplaced file extension like a PDF with a .DOCX extension will not affect file handling. And the indexer goes deep, extending through multilevel nested files like an email with a ZIP or RAR attachment that has a Word



Enterprise search like dtSearch® enables any number of end-users to concurrently and instantly sift through terabytes after first indexing the data. While farmwork takes actual work, it takes little to no effort to kick off indexing.

Reprinted with permission of USA Daily Chronicles, usdailychronicles.com

document with an Excel spreadsheet embedded inside, parsing complete text and metadata.

What is index capacity?

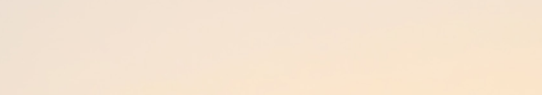
A single dtSearch index can hold up to a terabyte of text, and there are no built-in limits on the number of indexes the software can create and simultaneously search. The index structure enables multiple end-users to concurrently search the data, with each query operating independently so as not to impact other search threads. That way, searching stays snappy even during busy harvest times. As data grows, the indexer can update its indexes automatically through the Windows Task Scheduler to accommodate new, modified or deleted files without affecting ongoing searching.

But how does indexed search get to the wheat while ignoring the chaff?

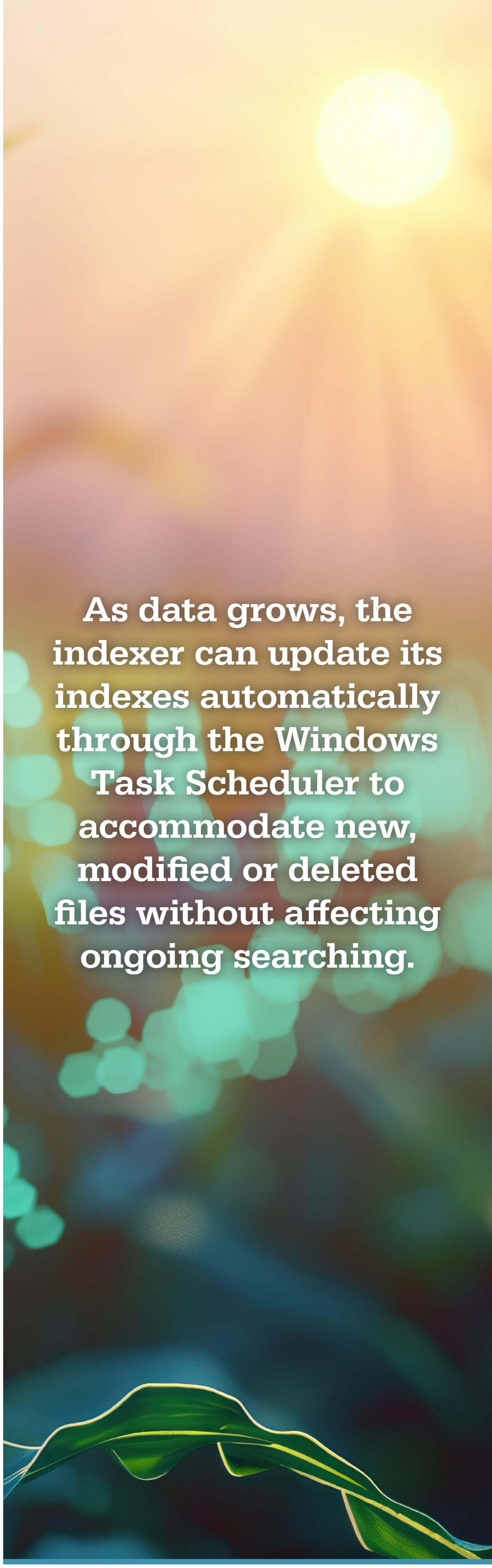
dtSearch offers over 25 different search features for precision queries. For basic searching, just check off “any words,” “all words” or “exact phrase.” An “any words” search for *wheat chaff kernel* would find any item that contains even one mention of *wheat*, *chaff* or *kernel*. An “all words” search for the same words would find only items that contain at least one mention of all 3 of these search terms. An exact phrase search for *wheat kernel* would find files that contain this precise expression.

And for more complex search formulations?

That is where Boolean and/or/not and proximity search requests come in. You could do a search for *wheat kernel* along with *chaff* in a document that also mentions *South Dakota* or *Nebraska* but not *Sioux City* or *Omaha*. Or you can enter a proximity element like *wheat kernel w/12 threshing machine*, looking for the first phrase within 12 words of the second phrase in either direction. Or you could require that *wheat kernel* appear within 34 words before *threshing machine*. Or add on metadata requirements, like any of the above searches in an email from sender *Spelt Farmer* to recipient *Barley Farmer*.

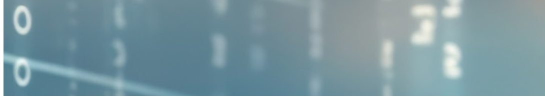


Article contributed
by dtSearch®



As data grows, the indexer can update its indexes automatically through the Windows Task Scheduler to accommodate new, modified or deleted files without affecting ongoing searching.

Reprinted with permission of USA Daily Chronicles, usdailychronicles.com



Article contributed
by dtSearch®

What about dates?

A search can include various date parameters like *July 31, 2022* or date range *July 1, 2021 to August 22, 2023* anywhere in a file or in specific metadata. Date searching will also pick up common date variants like *Aug 5, 2022* or *8/5/22*. The software can likewise look for numbers or numeric ranges anywhere in the data or just in specific metadata. And dtSearch can further identify any credit cards that may appear in the data – just in case a credit card used to buy farming equipment accidentally winds up in open enterprise data.

What about international languages?

Enterprise search automatically supports hundreds of international languages. The key is Unicode, which works not only with European languages but also double-byte Asian text like Chinese, Japanese and Korean as well as right-to-left text like Hebrew and Arabic. Just as there is no need to tell dtSearch the type of file it is indexing, there is no need to tell dtSearch the file’s language. In fact, a single file or email can cycle through multiple different languages, and Unicode and dtSearch will track that progression. In supporting Unicode, dtSearch also lets you search for individual Unicode emojis, such as the bread loaf emoji 🍞

Enterprise search automatically supports hundreds of international languages. The key is Unicode, which works not only with European languages but also double-byte Asian text like Chinese, Japanese and Korean as well as right-to-left text like Hebrew and Arabic.

But how do you separate the wheat from the chaff?

Sometimes a search request will take you straight to what you are looking for. Other times, even a precision query will retrieve too many items for easy review. That’s where relevancy ranking comes in. Take the “any words” search for *wheat chaff kernel*. With default vector-space relevancy-ranking, if *wheat* and *kernel* are common across indexed content but *chaff* is rare, then *chaff* hits will get a higher relevancy score, with denser references scoring even higher. Or add your own custom weightings, giving *wheat* a positive weight of 6, *kernel* a negative weight of 3, and *chaff* a positive weight of 7 but only in certain metadata or near the top or bottom of a file. For a new window on search results, instantly re-sort by a completely different metric like filename or file date. In all cases, the software will display the full text of items with highlighted hits for convenient browsing.

Final thoughts?

Try enterprise searching to extract the wheat from the chaff across terabytes of your organization's data. Start now by downloading a fully-functional 30-day evaluation copy from dtSearch.com

Article contributed
by dtSearch®

Reprinted with permission of USA Daily Chronicles, usdailychronicles.com