Wrap Up the Year's Data Then Tie a Bow on It With Enterprise Search

Article contributed
by dtSearch®

It's been an eventful year. Wrap up the year's data then tie a bow on it with enterprise search. That way, everyone in your organization can start the new year on top of all relevant data.
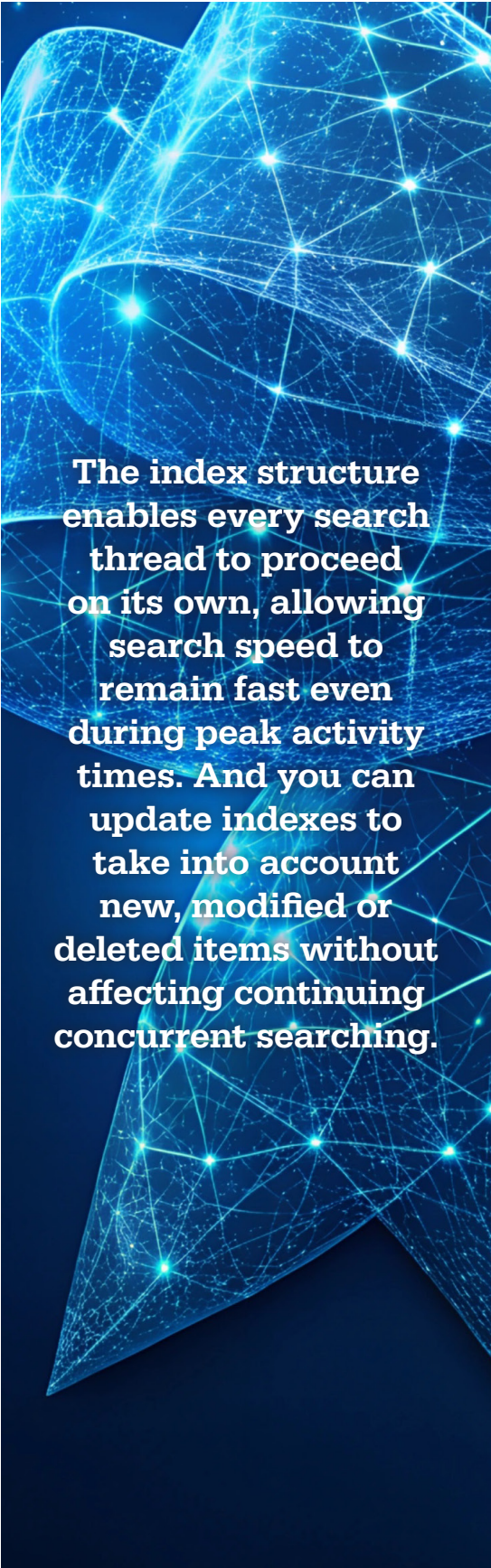
Enterprise search like dtSearch® can run in a classic network environment, from a local web server, or from a server in the cloud. In all configurations, enterprise search lets multiple end-users concurrently sift through massive data repositories. The software includes efficient multithreaded search processing and display of individual search results showing the full text of retrieved files with highlighted hits for easy review.

All products generally work the same way whether running from an on-premises network, from an on-premises web server, or from a cloud server. Enterprise search instantly and concurrently searches through terabytes only after first indexing the data. But indexing couldn't be easier. Just tell the indexer the email archives, file folders and the like to cover, and the indexer will take it from there.

The files themselves can be local inside the Windows folder system. Or the files can be remote like OneDrive / Office 365 documents, SharePoint attachments or DropBox documents and just appear as part of the Windows folder system. No need to identify file types. dtSearch, for example, can on its own figure out if files are local or remote PDFs; Microsoft Word, Access, Excel, PowerPoint or OneNote; Outlook, Exchange or other email files; or even ZIP or RAR compressed.

To make this determination, the software uses the information inside the binary format. That way, a mismatched file extension like a Word document saved with a .PDF extension or a PDF saved with a .DOCX

The index structure enables every search thread to proceed on its own, allowing search speed to remain fast even during peak activity times. And you can update indexes to take into account new, modified or deleted items without affecting continuing concurrent searching.

extension will not affect the process. The product line can even handle multilevel file data. An email can have a ZIP or RAR attachment with an Excel spreadsheet inside and a Word document embedded in the spreadsheet, and the indexer will handle the whole chain.
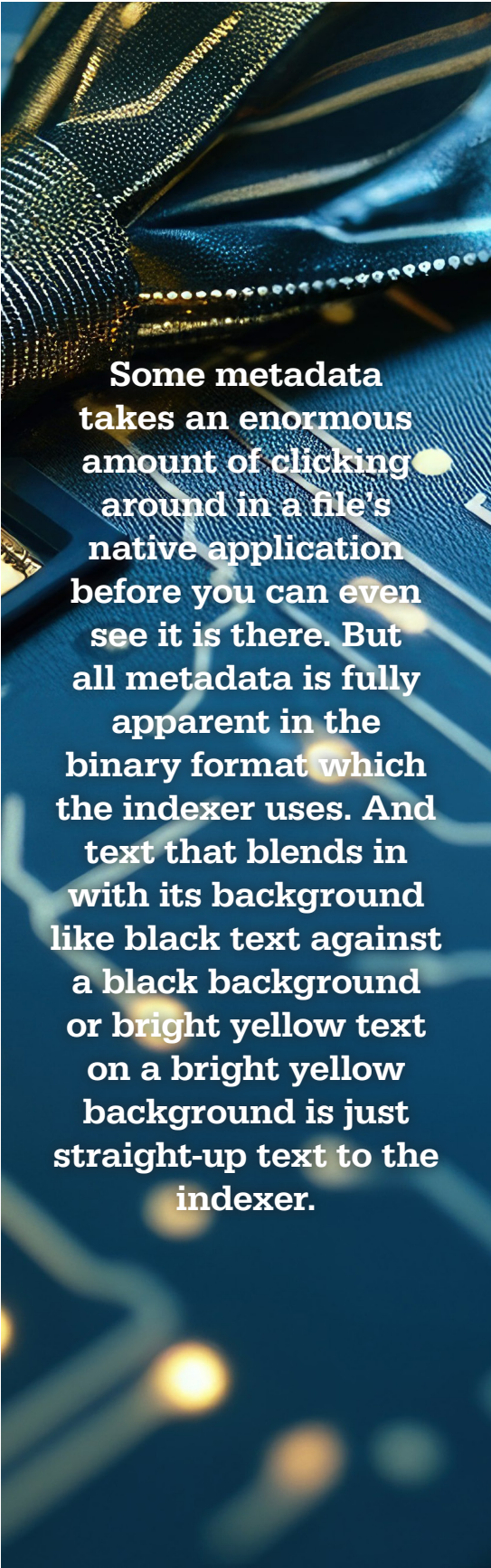
With dtSearch, a single index can hold up to a terabyte of text and there are no limits on the number of indexes the software can create and end-users concurrently search. While indexing is resource-intensive, searching is resource-light making it easy for multithreaded search requests to seamlessly scale even during peak data-access times. And concurrent searching can continue at the same time as indexes update to reflect new files added, deleted or modified.

Suppose you wanted to search for: *wrap up the year's data then tie a bow around it*. You could do any "all words" search finding only items that contain all the above keywords. Or you could do any "any words" search finding items that contain even one of the central keywords, like *wrap*, *data*, *tie* or *bow*. Or you could do a more intricate structured query like looking for the phrase *2024 data* and the phrase *tie a bow* with no mention of the phrase *holiday party*. Or you could do a proximity search such as locating *tie a bow* within 27 words of *2024 data*.

By default, searching will cover the full text of all files plus all metadata. Or a query formulation can require that certain components appear in specific metadata. The software also supports date searching like looking for *12/31/24* or date range searching like looking for *date(1/15/24 to 12/31/24)* either as part of the full text of files or limited to specific metadata. Date searching can pick up not only dates in the *12/31/24*-type format but also automatically extend to other formats like *December 31, 2024* or *Dec 31, 2024*.

Concept searching will let you search for *present* and find *gift*. Fuzzy searching adjusts from 1 to 10 to sift through typographical or OCR errors, like *present* mis-OCR'ed or mis-typed as *presemt*. Stemming finds different endings on the same root word like *presented*, *presenting* and *presents* for *present*. The software can

Some metadata takes an enormous amount of clicking around in a file's native application before you can even see it is there. But all metadata is fully apparent in the binary format which the indexer uses. And text that blends in with its background like black text against a black background or bright yellow text on a bright yellow background is just straight-up text to the indexer.

also search for numbers and numeric ranges, and can even identify any credit card numbers that might be lurking in enterprise data.
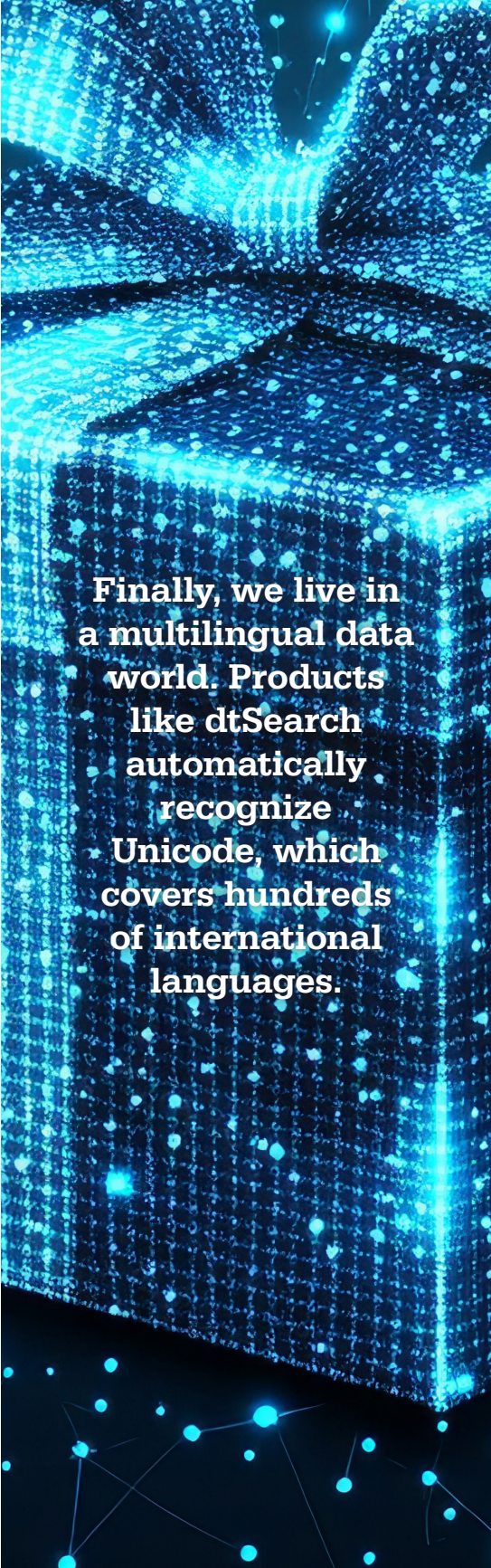
By default, dtSearch will apply so-called vector-space relevancy ranking to search results. Take any "any words" search for *wrap*, *data*, *tie* or *bow*. If *wrap* and *data* are common across indexed content but *tie* and *bow* relatively rare, then files with *tie* or *bow* will get a higher relevancy rank and files with the densest mention of these will get the highest relevancy weight.

Variable term weighting lets an end-user apply custom term weighting like giving *tie* a positive weight of 3, *wrap* a positive weight of 8 but only if it appears in certain metadata or at the top or bottom of a file, and *2023* a negative weight of 7. Or for a different view on the data, instantly re-sort by some non-relevancy metric like filename or file location. Whatever the sorting, everyone will see a full copy of retrieved files with highlighted hits for convenient browsing.

Finally, the product line works with Unicode which covers hundreds of international languages. A single file can go from one Unicode encoding to the next to the next and the product line will track the whole progression, even covering right to left languages like Hebrew and Arabic and double-bye Chinese, Japanese and Korean text.

So wrap up the year's data then tie a bow around it. Visit dtSearch.com for fully-functional 30-day enterprise search evaluation downloads.

Article contributed
by dtSearch®

Finally, we live in a multilingual data world. Products like dtSearch automatically recognize Unicode, which covers hundreds of international languages.