

Need to search terabytes of enterprise data? Tips for getting quickly to that 4-leaf clover

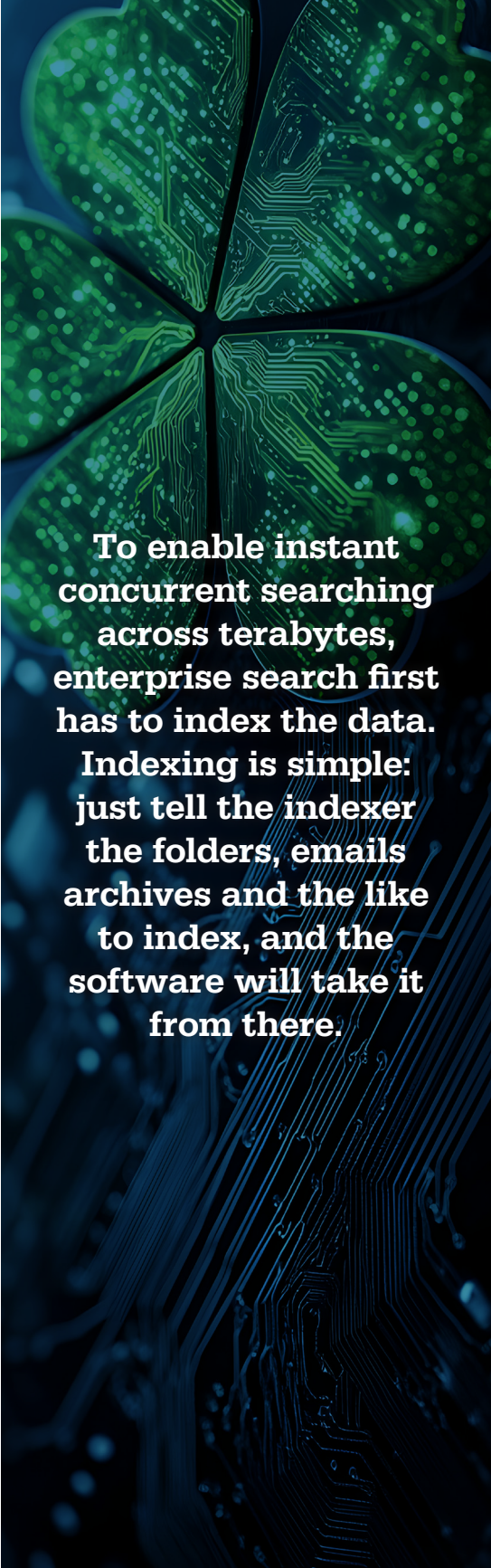
If you find yourself in a springtime clover field hunting for that rare 4-leaf clover, the journey is the reward. Not so if you and your team are hunched over your desks hunting 4-leaf clovers in terabytes of enterprise data. While combing through millions of files is never a “walk in the park,” enterprise search makes this process exponentially more pleasant.

To enable instant concurrent searching across terabytes, enterprise search first has to index the data. Indexing is simple: just tell the indexer the folders, emails archives and the like to index, and the software will take it from there. (This article uses dtSearch for its specifics on enterprise search but there are other comparable products on the market.) Tip: the files to index can be local or remote like SharePoint attachments, OneDrive / Office 365 files, etc. that appear as part of the Windows folder system.

The indexing process records each unique word and number across the data, and the location of each in the data. To get this information, enterprise search goes straight to the binary formats of files, bypassing the retrieval of each in its originating or associated application. While the indexer needs to know the exact right file type to correctly parse a file, the software can figure this out independently through the binary format. Tip: the indexer can determine the file type from the binary format regardless of file extension; a PDF can have a OneNote file extension and an Access database a PowerPoint extension without affecting file type determination.

A single index can hold up to a terabyte of text and there are no limits on the number of indexes the

Article contributed
by dtSearch®



To enable instant concurrent searching across terabytes, enterprise search first has to index the data. Indexing is simple: just tell the indexer the folders, emails archives and the like to index, and the software will take it from there.

software can build, and multiple end-users instantly and concurrently search. Searching itself can run in a classic Windows network environment, from an on-premises web server (typically Intranet for enterprise search), or from the secure cloud like Microsoft Azure or AWS. While indexing is resource-intensive, searching is not, making instant multithreaded searching easy to scale. Tip: index updates can proceed without interrupting continuing multithreaded searching, making it seamless to accommodate new, modified and deleted content.

Indexing has a very broad reach. It can cover multilevel nested data like an email with a ZIP or RAR attachment including a Word document with an Excel spreadsheet embedded inside. Enterprise search further encompasses all metadata – even metadata that might be really hard to spot in a file’s originating application. And it covers all text, including text that blends in with its background in a file’s originating application, like shamrock green text against a shamrock green background. Tip: redacted text that remains in the file even if not visible by default in the file’s associated application remains fully searchable.

After indexing, dive through the data using over 25 different search features. An “any words” search for *clover meadow shamrock* would find any file or email that contains even one of these words. An “all words” search for *clover meadow shamrock* would retrieve only files or emails that contain all 3 terms. A *clover meadow shamrock* phrase search would look for this exact phrase. The software also enables highly intricate Boolean (and/or/not) and proximity search formulations. Tip: while searching by default spans the full text of all data, the software also lets you limit a search or search component to specific metadata.

Concept searching finds synonyms. (I had no idea that *trefoil* was a synonym for *shamrock*.) Fuzzy searching adjusts from 1 to 10 to sift through typographical or OCR mistakes, like *shanrock* for *shamrock*. Additionally, the software can find numbers and numeric ranges as well as dates and date ranges, including automatically picking up common date variants like *Mar 17, 2025*, *March 17, 2025* and *3/17/25*.



Indexing has a very broad reach. It can cover multilevel nested data like an email with a ZIP or RAR attachment including a Word document with an Excel spreadsheet embedded inside.

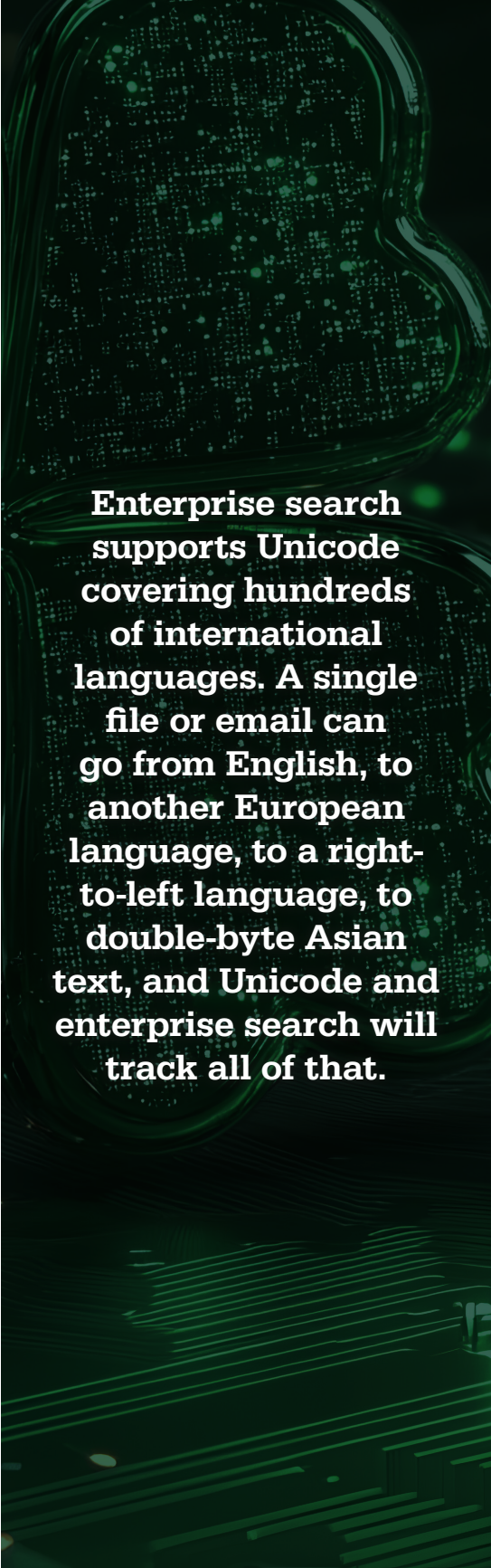
Tip: enterprise search can even identify credit card numbers hiding in indexed data.

Enterprise search supports Unicode covering hundreds of international languages. A single file or email can go from English, to another European language, to a right-to-left language, to double-byte Asian text, and Unicode and enterprise search will track all of that. Tip: enterprise search also lets you search for specific Unicode emojis like 🍀 and 🍀 .

Default relevancy-ranking gives a higher weight to less common words across indexed data. In any “any words” search for *clover meadow shamrock*, if *clover* and *meadow* are common but *shamrock* rare, *shamrock* files will get a higher relevancy rank, with the densest *shamrock*-mentioning files coming out on top. Or customize term weighting, giving *shamrock* a positive weight of 8, *clover* a positive weight of 3 and *meadow* a negative weight of 7 for occurrences anywhere or just in specific metadata. Tip: for a different perspective on search results, instantly re-sort by a completely different metric like filename or file date.

Whatever the sorting, view a complete copy of retrieved files with highlighted hits. 4-leaf clovers found!

Article contributed
by [dtSearch®](#)



Enterprise search supports Unicode covering hundreds of international languages. A single file or email can go from English, to another European language, to a right-to-left language, to double-byte Asian text, and Unicode and enterprise search will track all of that.