Enterprise Search Through Terabytes: Why Your Findings Will Surprise You

Article contributed by dtSearch®

You know that game where kids dive to the bottom of a pool to pick up toys? It is hard to see what the toys are when they are sitting there on the floor of the pool. It's only when the kids fish them out of the water that it becomes clear what each toy really is.

It's the same thing with data. It's not humanly possible to read through terabytes of text in a few sittings. But let's say you try. You can retrieve email after email in an email reader, "Office" file after "Office" file in each's native application, and still miss something that enterprise search can reveal.

This article will cover how enterprise search instantly finds things that even reading through files and emails you might miss. And this article also touches on the extent of items enterprise search can retrieve. Hint: it goes far beyond words.

When you edit a file in its native application, that application displays the file. When you exit the native application, it saves the file in binary format. Looking at this binary format directly, you might have trouble making out any text at all amid all the binary codes. File format specifications can run hundreds of pages long for each binary file type. And enterprise search must match the exact right specification to each file type for accurate results.

You might think enterprise search uses the file extension to determine the file type. However, it is all too easy to save a PowerPoint with an Access database extension. The more accurate method of file format

> **You can retrieve email after email in an email reader, "Office" file after "Office" file in each's native application, and still miss something that enterprise search can reveal.**

identification is to look inside the binary format itself. Therefore, while a OneNote file with a PDF extension might fool someone browsing through files, it won't impede enterprise search's ability to correctly process that file.
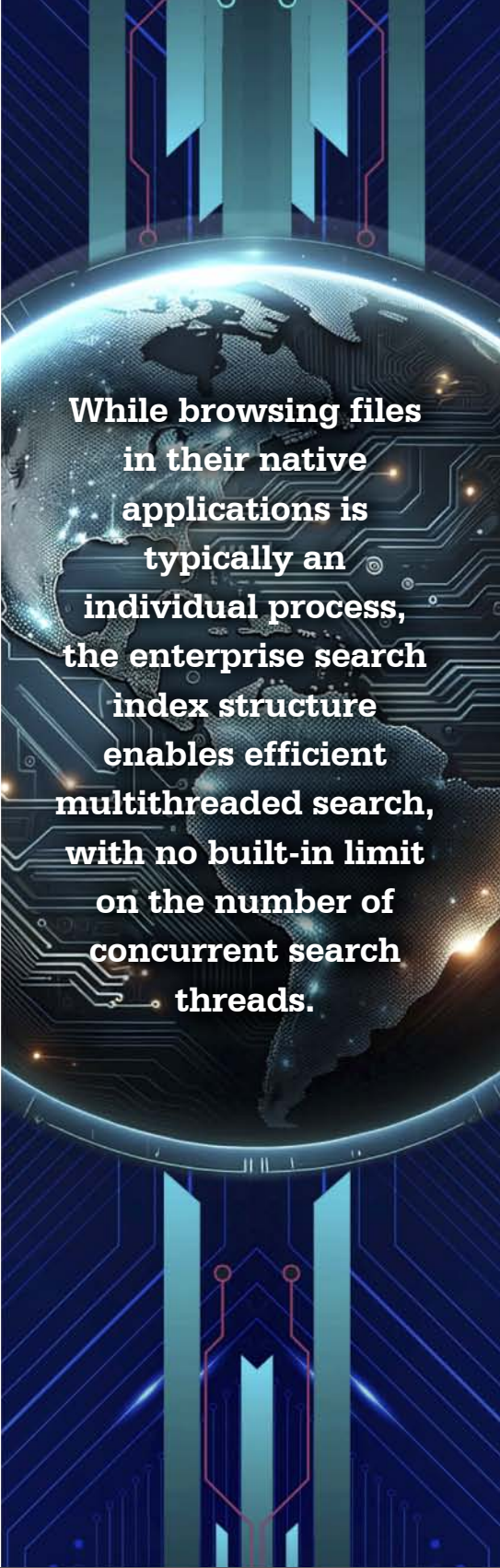
When you think of files, you probably picture stand-alone files. But you can have an email with a ZIP or a RAR attachment that itself contains an Excel spread-sheet with a Word document nested inside. Even assuming you drilled your way down to the embedded Excel file, you could still miss the sub-nested Word document. Fortunately, enterprise search can seam-lessly identify the entire structure down to the inner-most file just from the binary format.

You can click and click around files in their native applications and still miss certain metadata. Even worse, certain text can be present but invisible, like black writing against a black background or text behind an image. But all of that "hidden" text and metadata is fully apparent in the binary format and hence available to enterprise search.

Once enterprise search identifies the text and meta-data from binary files, it builds a search index holding each unique word and number and the position of each in the data. While indexing is a lot of work for enter-prise search, all you need to do is point to the folders and the like you want the index to cover, and the software will take it from there. The indexer can even work with web-based formats and remote Office365 or SharePoint documents that present through the Win-dows file system.

A single index can hold up to a terabyte of text, and there are no limits on the number of indexes that enter-prise search can generate and instantly search. While browsing files in their native applications is typically an individual process, the enterprise search index structure enables efficient multithreaded search, with no built-in limit on the number of concurrent search threads. As

While browsing files in their native applications is typically an individual process, the enterprise search index structure enables efficient multithreaded search, with no built-in limit on the number of concurrent search threads.

data updates, enterprise search can use the Windows Task Scheduler to update its indexes automatically without affecting continuing concurrent searching.
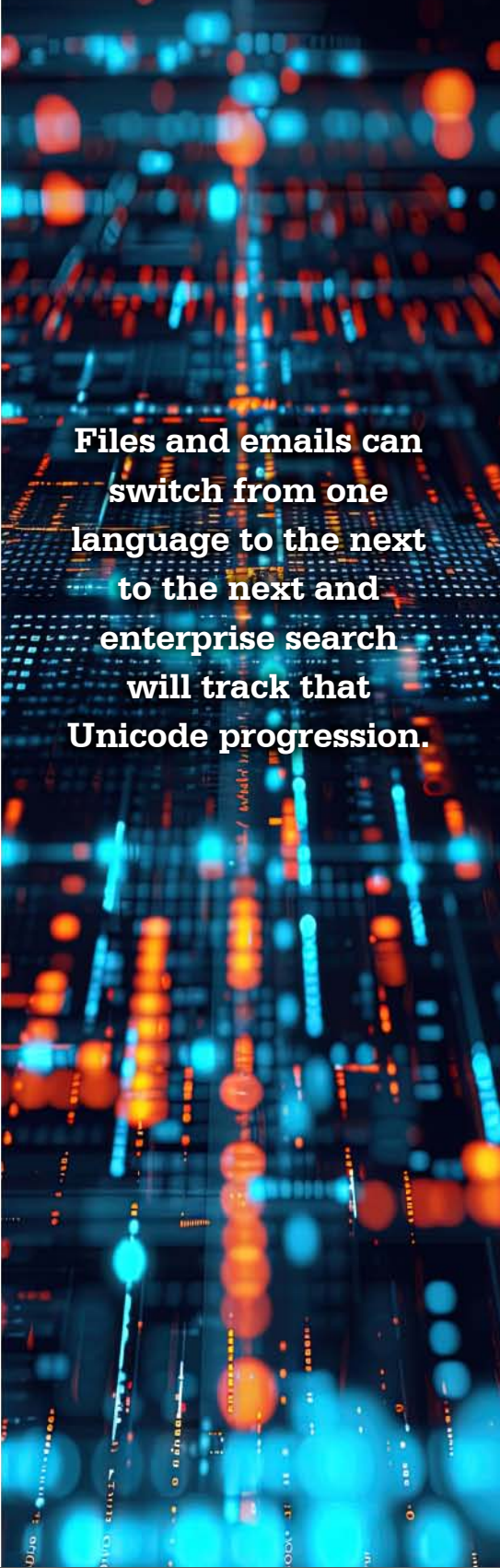
After indexing, enterprise search supports over 25 different individual and concurrent instant search options. Enter unstructured text, and leave it to the software to rank retrieved files and emails by relevance using "vector-space" search term rarity and density. Or craft a precision search request covering multiple words and phrases in Boolean (and/or/not) and proximity configurations. A search request can also include a number or a numeric range element or a date or a date range element — automatically extending across common date formats (March 15, 2024 vs. 3/15/24).

A search request can further span all text or require that certain items appear in particular metadata. Concept searching extends a search term to synonyms. Fuzzy searching adjusts from 1 to 10 to sift through minor misspellings that may occur in OCR'ed files or email mistypings. Enterprise search can even flag any credit card numbers across indexed data.

Have international language text documents? Enterprise search will automatically track Unicode in files and emails spanning hundreds of international languages including English and other European languages; right-to-left languages like Arabic and Hebrew; and double-byte Chinese, Japanese and Korean text. Files and emails can switch from one language to the next to the next and enterprise search will track that Unicode progression. Enterprise search can also search for specific Unicode emojis 😊

After a search, the software can display retrieved files with highlighted hits. For a different view on search results, the product can instantly re-sort results, such as re-sorting relevancy-ranked search results by file date or file location. So, get that instant individual or concurrent enterprise search going across terabytes. You never know what it might dredge up.

Files and emails can switch from one language to the next to the next and enterprise search will track that Unicode progression.