Escape the Data Riptide With Enterprise Search
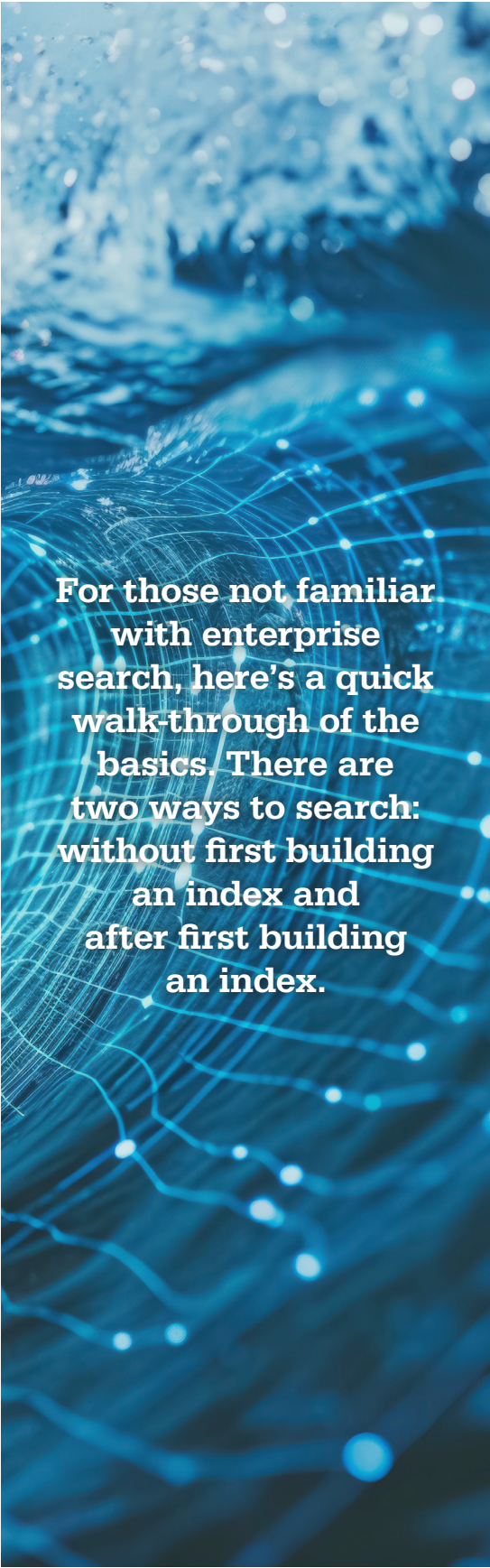
Everyone wants some extra free time to enjoy summer. But to get there often requires an escape from the riptide of ever-expanding enterprise data. Enterprise search is up to the job. For those not familiar with enterprise search, here's a quick walk-through of the basics. There are two ways to search: without first building an index and after first building an index.

While dtSearch®, for example, offers both options, indexed searching enables the instant multithreaded concurrent searching across terabytes that organizations need to tackle the data riptide. Indexing may sound like a lot of hot summer work. But all you need to do is check off the email archives, folders and the like to cover and the software will take it from there. It doesn't matter if files are remote like Office 365, DropBox or SharePoint. As long as the files appear as part of the Windows folder system, the indexer can handle them just like ordinary local files.

To efficiently index the data, the indexer bypasses retrieving each email, "Office" file, PDF and the like in its associated application and instead approaches everything in binary format. To accurately parse each file, the indexer needs to figure out the exact file format. But the indexer can use the binary format itself to determine the precise file format without reference to the file extension. That way, a PowerPoint saved with a .PDF extension or a OneNote file saved with an Access database extension won't trip up the indexer.

The indexer's reliance on the binary format offers other benefits as well. The binary format makes obscure metadata that might be very hard to spot in a file's associated application immediately visible to the indexer. The binary format also enables the indexer to handle recursively nested formats like an email with a ZIP or RAR attachment containing a Microsoft Word file that itself embeds an Excel spreadsheet. The indexer

> For those not familiar with enterprise search, here's a quick walk-through of the basics. There are two ways to search: without first building an index and after first building an index.
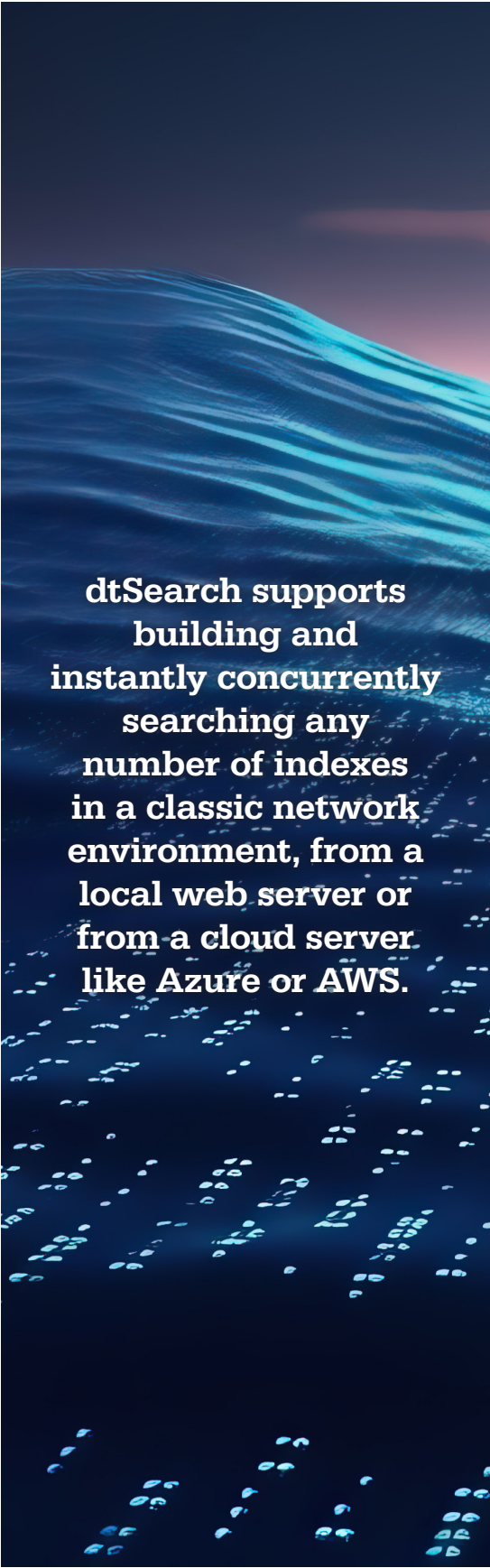
will handle everything there down to the innermost text and metadata. And the binary format lets the indexer detect and flag PDFs that are "image only" and require OCR processing such as through Adobe Acrobat Reader prior to full-text searching.

Additionally, the binary format makes it easy for the indexer to handle text like blue writing against a blue background or white writing against a white background that the display of a file in its associated application might obscure. This includes text that may appear under a black rectangle and look invisible in certain redaction programs but actually remains in the document. This also includes text that may appear deleted in certain track changes modes but that can persist as part of a file if track changes are not fully accepted.

The final index stores each unique word and number and its position in the data. With dtSearch, each index can hold up to a terabyte of text. And dtSearch supports building and instantly concurrently searching any number of indexes in a classic network environment, from a local web server or from a cloud server like Azure or AWS. While indexing is resource intensive, searching is much less so. dtSearch offers efficient multithreaded searching and can update indexes automatically to accommodate file modifications, new files and file deletions without interrupting network or web-based concurrent searching.

After indexing, choose from over 25 different search options. Less experienced end-users can enter basic "all words" or "any words" search requests like *whirlpool tsunami currents*. For more experienced end-users, dtSearch has precision phrase, Boolean (and/or/not) and proximity search options: *pacific tsunami* or *tidal wave* in a file that also mentions *ocean currents* but not *pond ripples* and has subject metadata including *riptide* within 14 words of surface calm. Concept searching will automatically extend a search to synonyms like *crosscurrent* for *riptide*. Fuzzy searching adjusts from 1 to 10 to look for typographical deviations like *riptibe* that can occur as a result of email mistyping or from OCR errors.
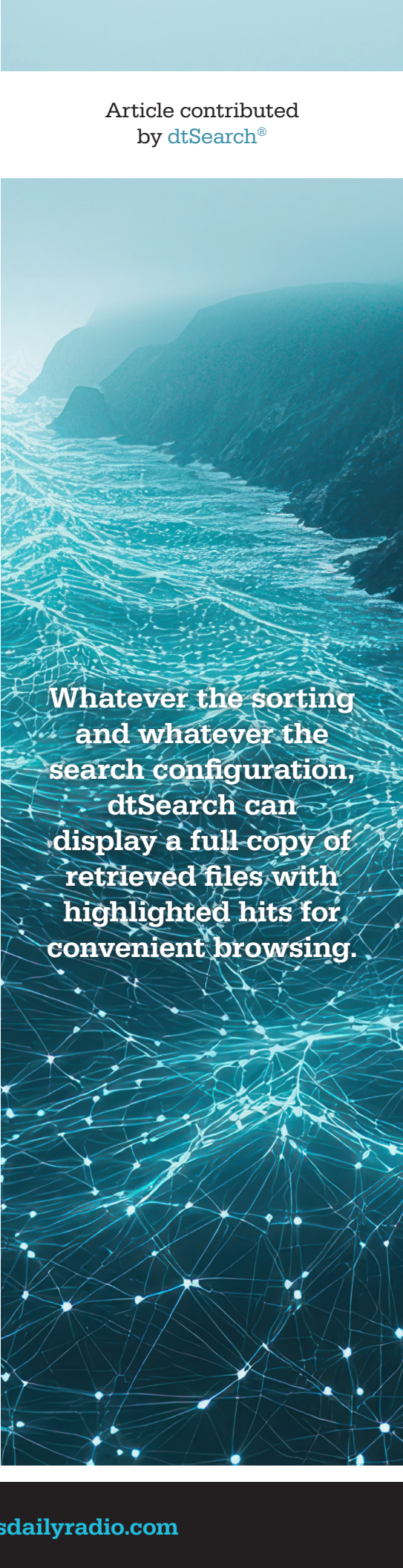
dtSearch supports building and instantly concurrently searching any number of indexes in a classic network environment, from a local web server or from a cloud server like Azure or AWS.

Article contributed
by dtSearch®

For international language text, dtSearch supports Unicode in files. A single file or email can cycle through English and other European languages, right-to-left text like Arabic and Hebrew and double-byte character Asian text. Unicode and dtSearch will track all of that. The software also supports date and date range searching across the full-text of files or in certain metadata. This type of searching will even pick up common date variants so a search for *date(2/3/24 to 3/4/26)* will pick up *7/1/25* as well as *July 1, 2025*. The software further supports number and numeric range searching. dtSearch can even identify any credit card numbers in text or generate and search for hash values across indexed data.

By default, the software relevancy ranks files based on hit term density and rarity. Take an "any words" search for *whirlpool tsunami currents*. If *tsunami* and *currents* are common across indexed data but *whirlpool* rarely mentioned, then files with *whirlpool* will get a higher relevancy rank, with files with the densest mentions ranking highest. Or users can add custom positive or negative variable term weighting across all text, positionally at the top or bottom of files, or in specific metadata. For a different window on search results, the software can instantly re-sort by a new metric like file location or file date. Whatever the sorting and whatever the search configuration, dtSearch can display a full copy of retrieved files with highlighted hits for convenient browsing.

So this summer, escape the riptide of enterprise data. dtSearch has fully-functional 30-day evaluation enterprise search downloads to get your organization started on instant concurrent searching across terabytes. And here's to using that extra time to get to the beach.

Whatever the sorting and whatever the search configuration, dtSearch can display a full copy of retrieved files with highlighted hits for convenient browsing.