# Lost In A Sea Of Data?

Article contributed by dtSearch®

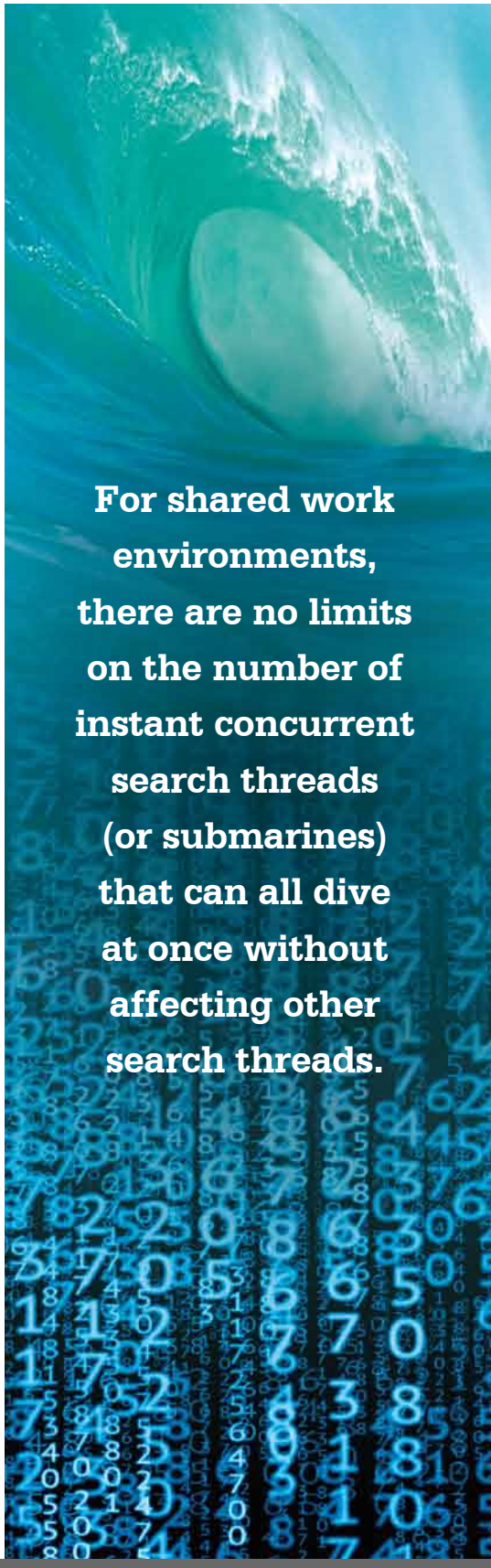
Ever feel like you are in a leaky canoe barely staying afloat atop terabytes of data? Transform that leaky canoe into a state-of-the-art exploration submarine with a search engine.

When you think of a search engine, you probably conjure a scan-the-Internet search engine such as Google. Enterprise search engines are in a different product category from your typical online search engine. With an enterprise search engine like dtSearch® one or more exploration submarines can instantly plumb oceans of data and emerge with treasures from the deep.

An enterprise search engine works its search magic after first indexing the data. While indexing requires a lot of effort, all you have to do is point to the folders, email stores, etc., to index, and the search engine will take it from there. When you review files and emails, you typically retrieve them in their associated applications: Microsoft Word, Excel, Access, PowerPoint, OneNote, Outlook, Adobe Acrobat Reader, etc. For efficiency, however, the search engine must bypass the associated applications' retrieval and go straight to the binary formats of files.

While associated applications optimize their file displays for readability, binary formats can look like a sea of binary codes, making it hard to make out any words at all. To identify the text, the search engine needs to apply the right parsing specification. These specifications can be hundreds of pages long and vary dramatically not only across different file types, but sometimes even within different versions of the same file type. While a file extension can suggest the file type, it is far from definitive, as it is not hard to save a PDF with a .DOCX extension or a Word document with a .PDF extension. For accuracy, the search engine needs to rely on the binary format itself to determine the parsing specification to apply.

After applying the correct parsing specification, the search engine is ready to start indexing. The index is an internal tool the search engine builds recording each unique word and number in the data and also tracks the location of each. Expect the search engine to go quite deep here in its indexing reach.



For example:

- Text that blends in with the background color in an associated application view, like aqua blue writing against an aqua blue background, may be nearly impossible to spot in an associated application view but is just regular text to a search engine
- "Hidden" metadata that may take a lot of clicking around in an associated application before you even know that it is there is on par with any other metadata to a search engine
- Indexing covers multi-tiered nested structures down to the innermost level, like an email with a ZIP or RAR attachment with an embedded Word document inside, with the Word document itself containing an Excel spreadsheet.

After indexing, the submarines are free to start their explorations. For shared work environments, there are no limits on the number of instant concurrent search threads (or submarines) that can all dive at once without affecting other search threads. Searching can span all data, or hone in on specific metadata only. Queries can range from simple "all words" or "any words" search requests to highly intricate word and phrase Boolean (and/or/not) along with proximity search formulations.

Search covers not only European languages, but also right-to-left languages like Hebrew and Arabic, and double-byte text like Chinese, Japanese and Korean. Fuzzy searching sifts through minor deviations in spelling such as mistypes (or mistypos) in an email. Concept searching finds synonyms of search terms. The search engine can also locate numbers and numeric ranges as well as dates and date ranges across different date formats. The search engine can even flag any credit card numbers that may have wormed their way into data.

A search engine has multiple options for relevancy-ranking, letting you sort and instantly re-sort search results by relevancy or other criteria. After a search, the search engine will dive to the bottom of the ocean to pull out a complete copy of retrieved items that you can then browse in full with highlighted hits. And if your ocean of data has new content flowing in and out, no problem. Instant concurrent searching can continue uninterrupted even while an index updates to reflect data additions, deletions or modifications.

Lost in a sea of data no more!

Article contributed by dtSearch®