Searching Emails

Article contributed
by dtSearch®

This article starts with how enterprise search works generally, and then goes over some options specifically for emails.

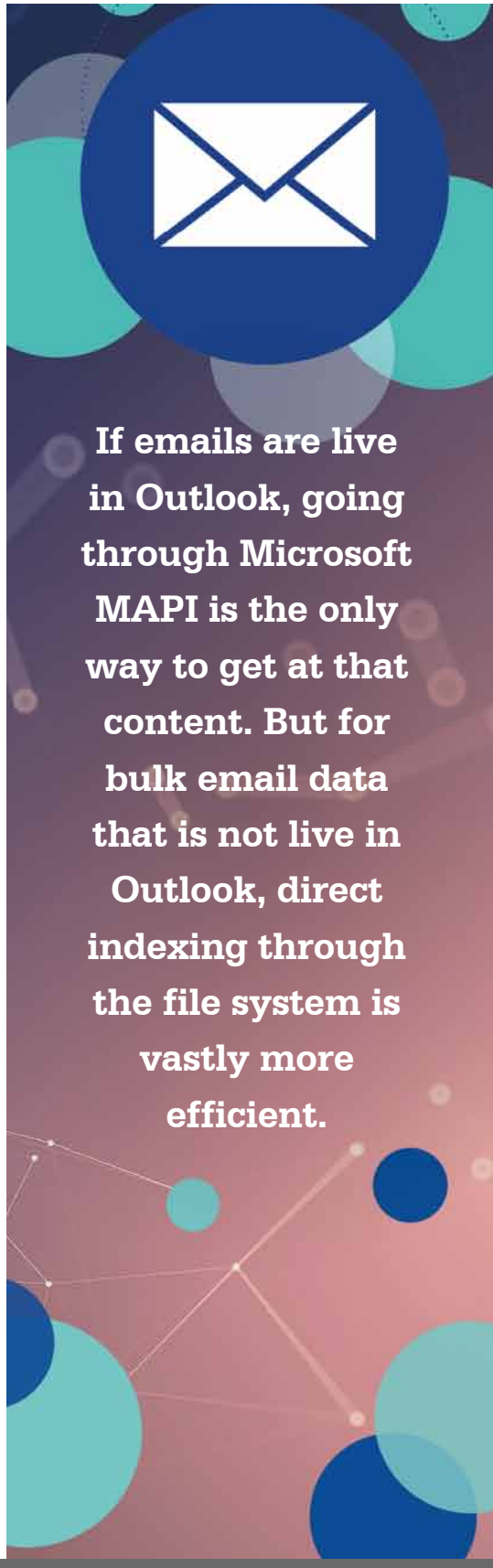
How Enterprise Search Works Generally. Enterprise search like dtSearch® instantly searches terabytes after first indexing the data. An enterprise search index is not like a back-of-a-reference-book index. Rather, an enterprise search index is just an internal tool for pre-processing unique words and numbers across the data and recording the location of each in the data. While indexing is a lot of work for the search engine, after indexing, a single search request can instantly span terabytes.

Indexing couldn't be easier. Just point to the folders and the like to index, and dtSearch will do the rest. When you review a file, you typically look at the file in its associated application, browsing a OneNote file in OneNote, a PDF in Adobe Acrobat Reader, etc. For a search engine to iterate across terabytes of data, however, the search engine needs to review each file in its binary format, bypassing associated application retrieval completely.

A file in its binary format view looks more like a jumble of codes than readable text. To distill out the text and metadata, the search engine has to apply the right parsing specification. File parsing specifications can be enormous, some hundreds of pages long, so applying the right one is key. You might think that a search engine like dtSearch would use the file extension to determine the correct file format, so a .DOCX would indicate a Word format, and a .PDF extension would indicate a PDF. But it is all too easy to have files that end up with mismatched extensions. The only surefire way to identify the file format is to look inside the binary file itself.

As long as cloud-based files like Office 365, SharePoint-synced files and the like appear in the Windows file system, the search engine can work with them just like any other data. After indexing, a single individual can search the data, or multiple end-users can simultaneously search the data. dtSearch supports multithreaded search with no built-in limit on the number of concurrent search threads. And concurrent search can continue while indexes automatically update to reflect new content.

dtSearch has over 25 different search features for precision searching. The search engine has numerous options for relevancy-ranking search results, as well as the ability to instantly re-sort by a completely different metric like file date or file location for a different window on search results. After a search, the search engine can display the complete text of retrieved files with highlighted hits.



If emails are live in Outlook, going through Microsoft MAPI is the only way to get at that content. But for bulk email data that is not live in Outlook, direct indexing through the file system is vastly more efficient.

Turning to Emails. dtSearch can index and search email files accessible through the Windows file system just like any other file data. Alternatively, dtSearch has an extraction utility to convert Outlook and Exchange files to .MSG, and then work with them that way. For emails that are live in Outlook, dtSearch can go through Microsoft's MAPI protocols to access the data. If emails are live in Outlook, going through Microsoft MAPI is the only way to get at that content. But for bulk email data that is not live in Outlook, direct indexing through the file system is vastly more efficient.

dtSearch can index the whole email plus the text of all attachments as an integrated unit, or can separately index the emails and the attachments. Attachment support covers both individual attachments and compressed attachments like RAR and ZIP. Nested file support ensures that dtSearch will comprehensively handle embedded attachments, like an Excel spreadsheet nested inside a Word document. After a search, dtSearch lets you optionally select and copy a specific email from a larger archival format, or select and copy an individual attachment out of a RAR or ZIP archive.

All of the 25 different search features apply to emails and any other formats that dtSearch indexes. But there are a few search options that are especially noteworthy for emails.

- By default, a search will cover the full text of an email plus all metadata. But dtSearch also lets you limit a search request or a portion of a search request to specific metadata. That way, if you want to limit a query to just emails that ABC sent or received, dtSearch can do that.

- dtSearch offers date and date range searching. These automatically extend to popular date formats, so a date range search for April 2, 2023, through May 1, 2023, will automatically pick up a different format such as 4/14/23 that may be in the full text or metadata.

- Because mistyping is common in emails, it is a good idea to activate fuzzy searching when sifting through emails. Fuzzy searching adjusts from 0 to 10 and works with just about any other text search option to broaden a search request to accommodate small typographical errors.

- In addition to general text search features, dtSearch also has the ability to locate email addresses anywhere in indexed data. The email addresses can of course appear in email metadata. But dtSearch can also identify them in the body of emails.

- Finally, dtSearch has the ability to flag credit card numbers that may be lurking in email as well as other data.

Article contributed
by dtSearch®

Attachment support covers both individual attachments and compressed attachments like RAR and ZIP. Nested file support ensures that dtSearch will comprehensively handle embedded attachments, like an Excel spreadsheet nested inside a Word document.