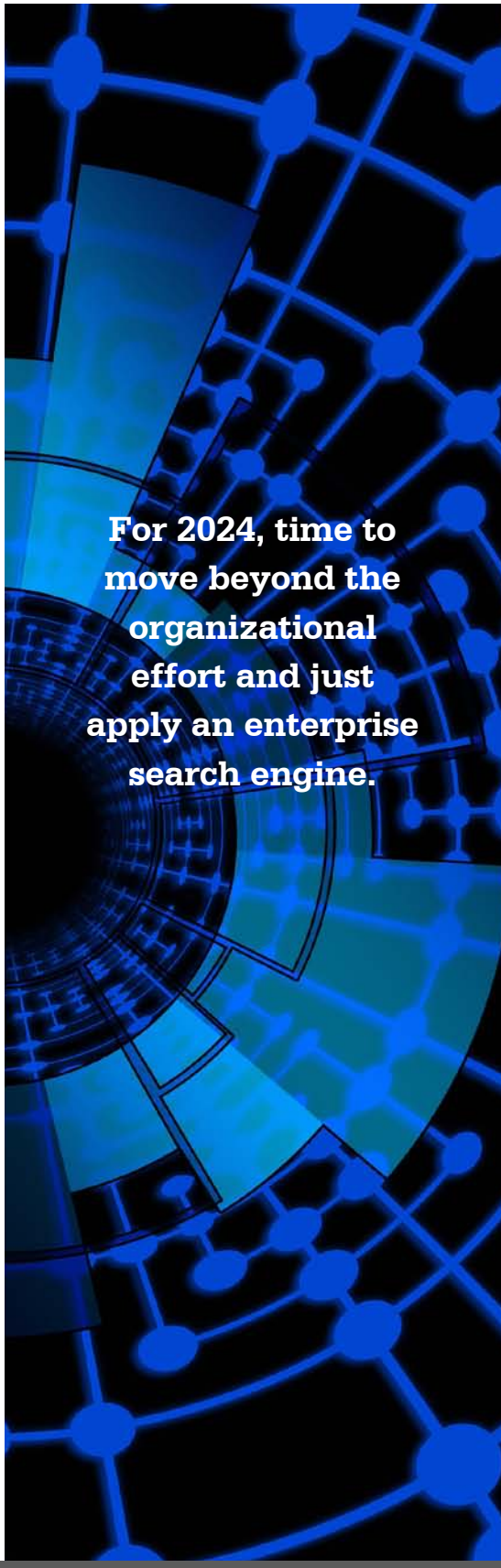
For 2024, Out With the Old Way of Organizing Enterprise Data to Find Things

Article contributed by dtSearch®



How do you navigate through your enterprise data? The old way: to meticulously sort everything into folders, subfolders, and sub-subfolders, and then go item by item ensuring that each filename reflects the contents of that file for visual scanning. And if you suddenly get in terabytes of outside data to process, game over. But even under the best of circumstances, organizing large volumes of enterprise data quickly goes from tedious to unmanageable.

For 2024, time to move beyond the organizational effort and just apply an enterprise search engine. An enterprise search engine is a different animal from a canvass-the-entire-web search engine like Google. An enterprise search engine digs deep into your organization's own data, instantly searching terabytes across the full text of content. While different enterprise search engines have different features and employ different terminology to describes those features, for discussion purposes this article uses dtSearch® as its springboard.

A search engine instantly searches terabytes after indexing the data. Indexing records each unique word and each unique number across the data, and the location of each in the data. To start the search engine indexing, all you need to do is point to the folders and the like to index. The search engine will take it from there. A single index can hold a terabyte across multiple data repositories, and there are no limits on the number of indexes the search engine can create and instantly query.

In building its index, the search engine reviews the binary version of all files and emails in selected folders. The search engine automatically works with popular file types like Microsoft Word, Excel, Access, PowerPoint and

OneNote; PDF; ZIP or RAR containers; email formats like Outlook/Exchange; and web-based data formats. The search engine can even handle multilevel nested formats, like an email with a ZIP attachment including a Word document with an Excel spreadsheet fully embedded in the Word file.

As a technical matter, the search engine has to identify each file type before parsing it, as different binary formats have very different parsing specifications. The component that does this file recognition and parsing goes by the name of document filters. Of course, in the real world, end-users can save PDFs with .DOCX extensions and Word documents with .PDF extensions. For that reason, the search engine will look inside the binary format to accurately recognize the file type.

The binary format view of a file is a very different window into the file than you would see looking at the file in its associated application. Text, such as black writing against a black background or white writing against a white background, that may have been hidden when viewing the file in its originating application is now on par with any other text. Metadata that may have taken an enormous amount of clicking around to find in the originating application is now just as apparent as any other text.

After indexing, the search engine supports multithreaded concurrent search across a classic network, via an "on premises" web server, or from the cloud. Online search can run in a completely stateless manner, making it very easy to scale. As data evolves, the search engine can automatically update its own indexes without affecting instant concurrent searching.

Over 25 different search options enable precision text retrieval. These range from basic "all words" / "any words" search requests to exacting Boolean and proximity word, phrase and metadata-specific formulations. Fuzzy search sifts through minor typographical errors than can arise in emails or OCR'ed text. Concept search expands a search to cover similar search terms. Unicode support covers not only European languages but also Chinese/Japanese/Korean double-byte text and right-to-left languages like Hebrew and Arabic.

Article contributed by dtSearch®

After indexing, the search engine supports multithreaded concurrent search across a classic network, via an "on premises" web server, or from the cloud.

Along with word-oriented queries, the search engine can also find numbers and numeric ranges, dates and date ranges spanning different date formats, as well as certain number patterns. For example, the search engine can flag any credit card numbers lurking in data. (In doing so, the search engine takes any digit sequence that might represent a credit card number and runs it past a credit card validator embedded in the software.)

The default relevancy ranking is via a vector-space algorithm. In a search for *commercial, jet* and *engine*, if *commercial* and *engine* are all over the data, but *jet* appears in just a few files, then *jet* would have a higher relevancy rank, and files with the densest mentions of *jet* would get the highest relevancy ranking of all. Advanced users can also apply their own variable term weighting, giving *commercial* a positive weight of 8 and *military* a negative weight of 7 across all text, or giving extra weight if terms appear at the top or bottom of files or in specific metadata.

After a query, the search engine can display a full copy of retrieved items with highlighted hits for convenient search results browsing. Or just click to apply a different sorting mechanism for a different window into search results. For example, after reviewing relevancy-ranked returns, instantly re-sort by file date or file location.

In sum, skip the tedium of organizing your data to locate what you need. For 2024, just install a search engine.

About dtSearch® dtSearch has enterprise and developer products that run "on premises" or on cloud platforms to instantly search terabytes of "Office" files, PDFs, emails along with nested attachments, databases and online data. Because dtSearch can instantly search terabytes with over 25 different search features, many dtSearch customers are Fortune 100 companies and government agencies. But anyone can download a fully-functional 30-day evaluation copy from dtSearch.com to enable instant concurrent enterprise search across terabytes of data.

Article contributed
by dtSearch®