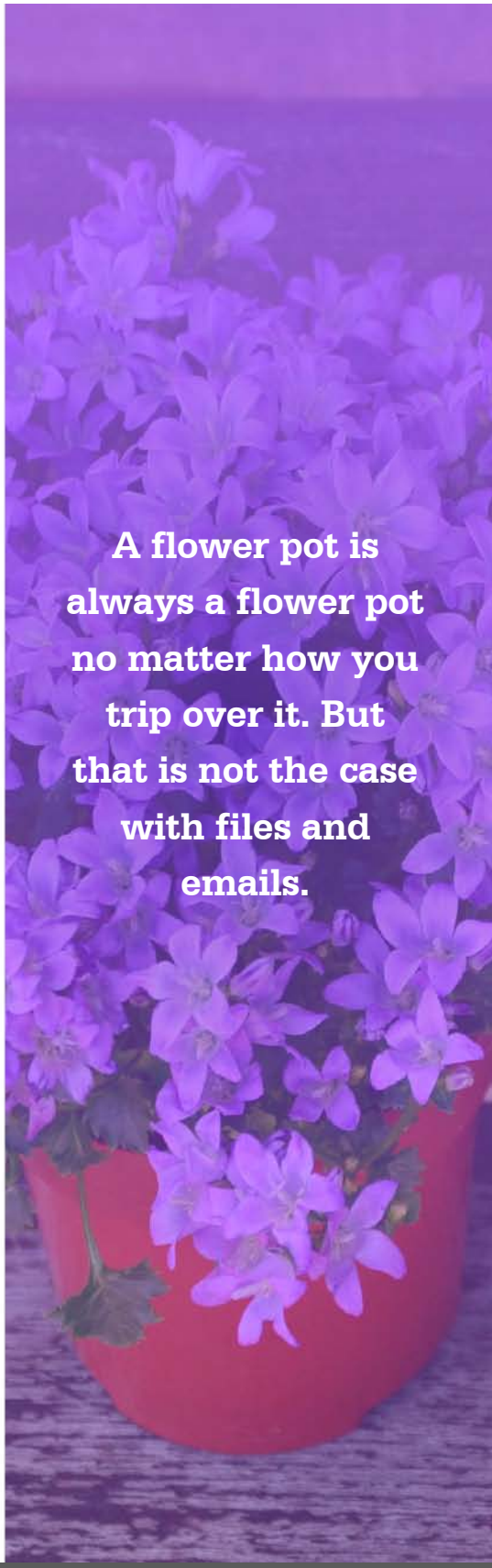
Cut Through the Data Clutter with a Search Engine, not a "Spring Clean"

Article contributed by dtSearch®

Spring is here, and with it the idea of spring cleaning to lighten clutter and more easily find our "go to" fair weather supplies. But while most spaces can benefit from a spring clean, enterprise data is not one of them. A search engine, however, can instantly cut through the clutter of terabytes of enterprise data

The nature of data. Physical objects have one mode of existence. A flower pot is always a flower pot no matter how you trip over it. But that is not the case with files and emails. Think how a word processing document looks in Microsoft Word, a spreadsheet in Excel, a database in Access, a presentation file in PowerPoint, a note in OneNote, a PDF in Adobe Acrobat or Reader, an email in Outlook or Exchange, etc.



However, a file or email sitting on your computer, your organization's network, or in the cloud looks nothing like the same file or email inside its associated application. In its "resting" state, a file looks like a jumble of binary codes. It can be difficult to make out a single word, much less read whole paragraphs. What you need is to instantly slice through the entire resting binary dataset, bypassing the need to retrieve each file in its associated application. That's where a search engine like dtSearch® comes in.

How a search engine works. The search engine builds an internal data structure called an index. The index stores each unique word and number and the location of each across all of the data. Each index can hold up to a terabyte of text, and there are no limits on the number of indexes that dtSearch can build and simultaneously cover in a search request.

Indexing is no work at all for the end-user; the search engine does it all. The end-user just needs to point to the folders and the like to include in the index and the software does everything else, figuring out the applicable format each item is in. A search engine like dtSearch can even automatically go through compressed archives like ZIP or RAR.

If a file is nested inside another file, the search engine will parse all that too. If you have an email with a ZIP file and inside the ZIP file is a Word file with an embedded Access database, the software will parse the entire nested structure. dtSearch will correctly parse a file even if a file has a different extension from the one it should have, like a Word document saved with a .PDF extension.

Note that dtSearch parses not only the full text of all files but also all metadata. Just like you can put a sticky note out of sight at the bottom of a flower pot, you can hide metadata in a file so that it is really hard to find in an associated application view. But that text will be "plain as day" to a search engine. Text that blends in with the background like black on black or white on white lettering is also just text for a search engine.

For data that constantly updates, dtSearch can update indexes automatically, including via the Windows Task Scheduler. The index updates need not re-index everything all over again. Instead, updates can just add new files, re-index modified files and remove data from deleted files. And updating an index does not affect continued concurrent searching, so dtSearch can update indexes as often as you like while everyone continues to search.

Searching. Indexed search can run from an individual PC, across a standard Windows network or from a website. The website can be hosted on a local server or a remote server on a platform like Azure or AWS. For web-based usage, dtSearch can operate in a completely stateless manner, so there are no limits on the number of concurrent search threads that can independently operate.

Using the index storing each unique word or number and its location in the data, a search engine like dtSearch can perform over 25 different search types across terabytes. You can enter an unstructured, natural language search request such as: *get me the March 2022 spring cleaning handbook*. Or you can enter a structured Boolean search request: *flower pot and spring cleaning and not (icicles or snow)*.

Article contributed by dtSearch®

For web-based usage, dtSearch can operate in a completely stateless manner, so there are no limits on the number of concurrent search threads that can independently operate.

You can also add proximity elements to that, like *flower pot* within 35 words of *azaleas* in either direction or *flower pot* but only within 12 words prior to *azaleas*. Additionally, you can supplement a search request with metadata search elements, such as subject contains *seasonal and not annual*. And you can do concept search to find synonyms of items.

Fuzzy search adjusts from 1 to 10 to sift through OCR or typographical errors. For example, if you are searching emails where typographical errors are common, a search for *flower pot* with a low level of fuzzy searching would also find *flomer pot*. dtSearch can also search for numbers and numeric ranges, as well as specific dates in any number of date formats along with date range searching. The product line can even identify credit cards that may be lurking in text. For forensics-type work, dtSearch can generate and search for hash values of all files in a data collection.

dtSearch supports all Unicode based languages, everything from European languages to double-byte Asian text to right-to-left Hebrew and Arabic text. dtSearch also has numerous options for relevancy ranking. Relevancy-ranking can be automatic based on the distribution and density of words in a data collection. Relevancy-ranking can also be user-defined such as positive and negative variable term weighting, including special positional and metadata-based term weighting. After a search, dtSearch can display all hits highlighted in a file along with the rest of the file text.

So ditch the spring cleaning when it comes to your enterprise data. And try a search engine instead.

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 with lots of data to search can download a fully-functional 30-day evaluation copy from dtSearch.com

Article contributed by dtSearch®

Updating an index does not affect continued concurrent searching, so dtSearch can update indexes as often as you like while everyone continues to search.