Celebrate The Earth By Replacing Your Paper Stacks With A Search Engine

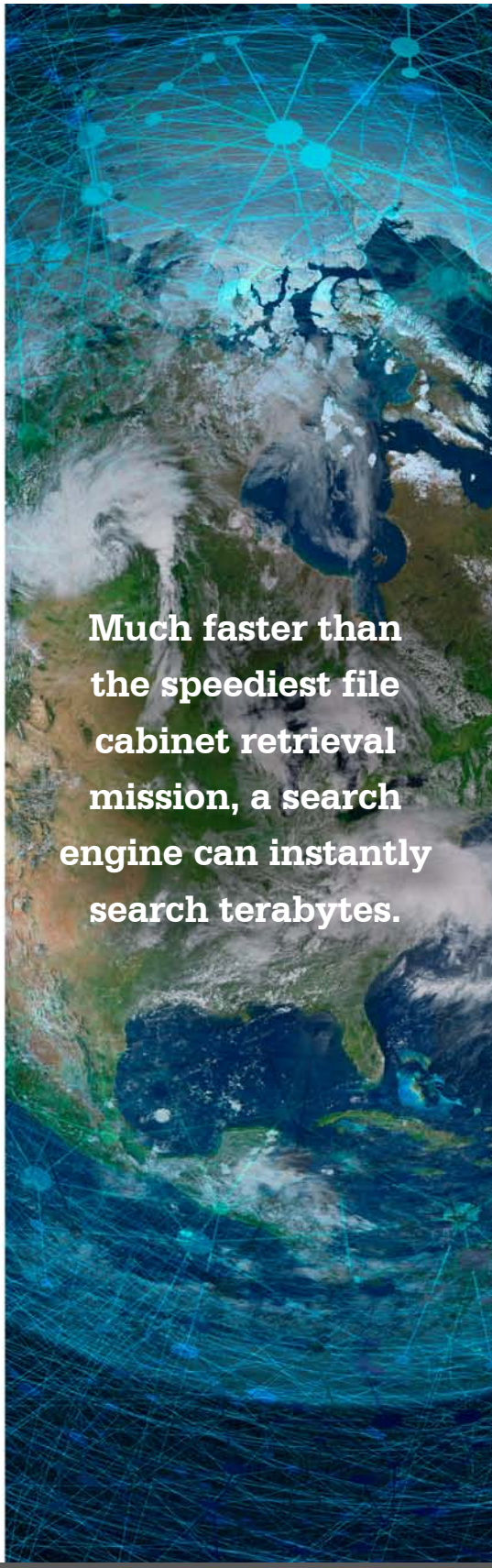
Article contributed by dtSearch®

Got mounds of paper your organization has amassed from the beginning of time? Recycle and replace them with a more eco and efficient solution in the form of a search engine. The first step is to scan and OCR the paper into “searchable image” PDF.

You may have a concern that crucial data could disappear in the paper-to-PDF process. When you scan and OCR into “searchable image” PDF using an application like Adobe Acrobat, you can retain a full copy of the original page, including all text and images. The process preserves a picture of anything on the paper, from a drawing of a springtime meadow to a scrawled note with someone’s initials. A search engine like dtSearch® can then display the page image in Adobe Reader with highlighted “hits” superimposed on the page image.

Much faster than the speediest file cabinet retrieval mission, a search engine can instantly search terabytes. First, however, a search engine needs to index the data. An index is not like a classic back-of-the-book index. Instead, an index is an internal tool cataloguing each unique word and number and the location of each across all text and metadata. While indexing is a lot of work for the search engine, the end-user just needs to point to the folders and the like to cover and the search engine does everything else.

After indexing, the search engine can perform over 25 different types of instant searches, displaying retrieved files with highlighted hits. Multiple people can instantly concurrently search across one or more indexes, with search running over a network, from a local web server, or from the cloud such as on Azure or AWS. No one in the office will ever have to rummage through those old file cabinets again!



Looking ahead to the post-paper world, in addition to PDFs an index can also cover other content like Microsoft Word, Access, Excel, PowerPoint, OneNote, web-ready data, and even emails plus attachments. The search engine on its own figures out the data type of each item, regardless of whether a document has a "mismatched" file extension, like a PDF with a .DOCX extension. (The search engine looks inside the file to determine the correct file type, not at the file extension.)

Multilevel file structures are also not an issue. You can have an email with a ZIP or RAR attachment containing an Excel spreadsheet with an Access database embedded inside the spreadsheet and the search engine will unwrap all of that. Black on black or white on white or purple on purple writing is just text for a search engine. Additionally, the search engine can find metadata content, no matter how obscure in the original file.

Finally, 3 "pro tips" for the post-paper world.

- When you have a large collection of PDFs, often the collection will include files that look like normal PDFs but are really "image only." (Have you ever tried to copy-and-paste text from a PDF but were unable to do so because there was no underlying text there, only an image?) A search engine can flag these "image only" PDFs so you can run them through an OCR program like Adobe Acrobat to make them full-text searchable.

- When OCR'ing old documents, you can end up with minor OCR errors, like *mistaqes* for *mistakes*. Fuzzy searching at a low level can sift through these OCR errors. You can apply fuzzy searching on top of nearly all of the 25+ other text query options. Fuzzy searching is also a good idea for formats like emails that are prone to typos.

- Beyond locating words and phrases in any number of configurations, a search engine can also perform numeric searches, including identifying credit card numbers in indexed data. That way, you can ensure that the data that you post for shared office searching is "clean."

Here's to life beyond paper!

Article contributed by dtSearch®