

Already Given Up on Your New Year's Resolution To Organize Your Data? Try a Search Engine Instead

Already given up on your New Year's resolution to organize your data? Try a search engine instead to instantly find anything across terabytes on your PC, your network and beyond, no resolution required.

A search engine like dtSearch® uses an index that it builds to instantly search through terabytes. To get the search engine started, all you have to do is point to the Windows folders, email repositories and the like to index, and the search engine does everything else. No need to even tell the search engine what types of data it is working with. The search engine can on its own figure out whether something is a PDF; a Microsoft Word, Access, Excel, PowerPoint, OneNote file; part of a compressed data format like ZIP or RAR; an email format; or even a web-based data format.

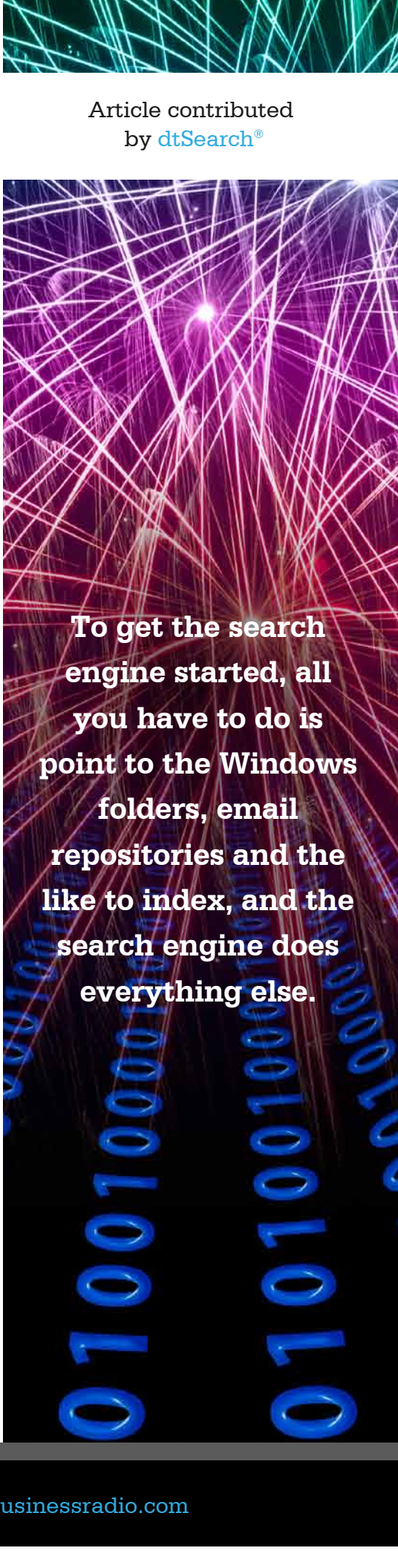
For efficiency, the search engine has to review each item in its binary format, as opposed to retrieving it its native application like you would normally view a Word file in Microsoft Word. If you took a look at a binary format directly, you'd see a maze of binary coding, making it hard in many cases to discern individual words at all. The search engine needs to apply the right parsing specification to filter out the text.

The parsing specification for Microsoft Word, for example, is hundreds of pages long and very different from the parsing specification for PDF. Even different versions of the same basic file type can have significantly different parsing specifications. When PDF 2.0 came out, the format was sufficiently different from the original PDF that dtSearch had to update its entire product line to handle the new format correctly.

Note that the search engine does not determine what format each file is from the file extension. It is all too easy to save a Microsoft Word document with a .PDF extension or a PDF with an Excel spreadsheet extension. This can happen accidentally, or even on purpose if someone wants to obscure something. A better way for a search engine to determine the file format is to look inside the binary file itself.

After applying the correct parsing specification, the search engine is ready to identify a file's full text and metadata and start indexing. The index is just an internal tool for the search engine to use. A single dtSearch index can hold up to a terabyte of text across a variety of data sources. The search engine can build and seamlessly search any number of terabyte indexes, with integrated relevancy-ranking across everything. In a multi-user search environment, different search threads can operate independently, with no built-in limit on the number of concurrent instant search threads that can run.

Article contributed
by [dtSearch®](#)



To get the search engine started, all you have to do is point to the Windows folders, email repositories and the like to index, and the search engine does everything else.

With over 25 different search features, a search engine like dtSearch can outdo even an epically organized human. The search engine can handle simple, unstructured “any words” or “all words” search requests. Or the search engine can process precision Boolean and/or/not and proximity search requests connecting a long series of words and phrases. For example, you could look for (*ProjectXYZ* or *ProjectQRT*) along with *North Florida* appearing within 12 before *South Beach Development* in a file that doesn't also mention *ProjectABC* in certain metadata.

Concept search extends a search request to synonyms. Adjustable fuzzy searching sifts through minor typographical and OCR errors, like *Florida* mistyped as *Floriba* in an email. After a search runs, the search engine can display a full copy of retrieved items with search term highlights for convenient browsing. For a different view of returned data, the search engine can also instantly re-sort search results by a new measure of relevancy or other criteria.

Along with words, indexed search supports looking for specific numbers or number ranges as well as dates or date ranges – even automatically extending across different formats like *1/13/23* and *January 13, 2023*. The search engine can generate and search for hash values as well as finding certain number patterns in text, like flagging any credit card numbers that may appear in indexed data.

With Unicode support, searching covers not only English and other European languages but also hundreds of other international languages. Search even covers right-to-left languages like Arabic and Hebrew and double-byte Chinese, Japanese and Korean text. A single file or email can include multiple international languages with the search engine supporting the entire progression.

Best of all, a search engine goes deep, finding data that even the most industrious human reviewer can miss. Files can have tricky multi-level nested elements. For example, you can have an email with a ZIP or RAR attachment including a PowerPoint with an Excel spreadsheet embedded inside. While a human reviewer may struggle to review all of this in the appropriate application, the search engine can readily sift through the entire nested structure from the binary format itself.

Likewise, obscure metadata that can be hard to spot in a file's native application is “front and center” in binary format. The same applies to text that blends in with the background color inside a file's native application. Black writing against a black background or white writing against a white background is easily visible in binary format.

In sum, don't kick yourself for abandoning your New Year's Resolution to organize your data. Just download 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 with lots of data to search can download a fully-functional 30-day evaluation copy from dtSearch.com

Article contributed
by [dtSearch®](http://dtSearch.com)

Note that the search engine does not determine what format each file is from the file extension.