Misconceptions About Enterprise Search

Article contributed by dtSearch®

Misconceptions about enterprise search abound. This article will attempt to resolve some common ones—and get you on your way to instantly searching terabytes.

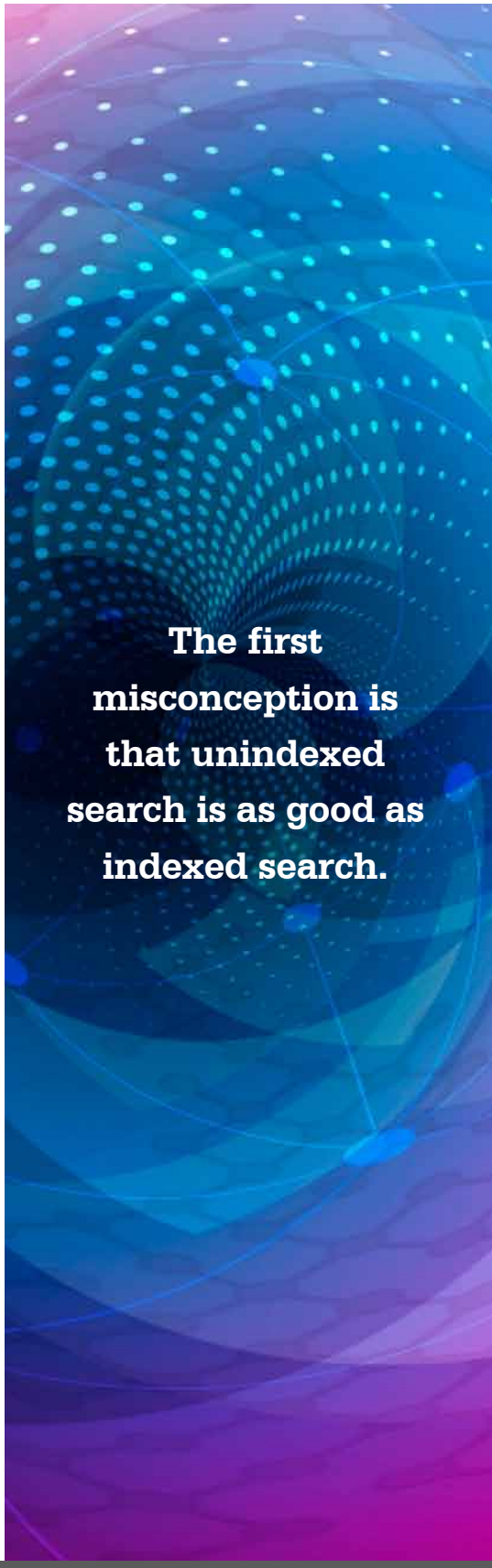
The first misconception is that unindexed search is as good as indexed search. For example, the application dtSearch® offers both indexed and unindexed search. However, indexed search is far and away the gold standard. Indexed search is instant, even across terabytes of content and even in a multithreaded concurrent search environment such as with a network installation, on a local web server, or in the cloud.

Beyond the speed of indexed search, it also enables more search options. Most of the 25+ dtSearch search options cover both indexed and unindexed search. But indexed search has some extra search options as well, like the ability to flag credit card numbers that may appear in indexed data. The indexer can run a series of numbers which might represent a credit card number through a validation algorithm to determine if it is actually a credit card.

The next enterprise search misconception is that building an index is somehow hard. In reality, it couldn't be easier. All you need to do is point to the folders, email archives, and the like to index, and the search engine does everything else, reviewing each file in its binary format. From the binary format, the search engine determines the applicable file type. After figuring out the file type, the search engine uses the file format specification for that file type to recognize all full-text and metadata.

Beyond storing each unique word and number in the data, the index also stores information on the location of each word and number. A single index can hold up to a terabyte of text. There are no limits on the number of terabyte indexes that the search engine can create, and end-users can instantly concurrently search.

For changing datasets, the search engine can use the Windows Task Scheduler to update indexes as often as you like. To update an index, the search engine need only re-index files that have been added, deleted or modified since the previous index build. Updating an index does not block out individual or concurrent searching, so all searching can continue unaffected during the update.



Reprinted with permission of The Times USA, thetimesusa.com

The next misconception is that a search engine will incorrectly handle files with a mismatched file extension, like a PDF saved with an .DOCX file extension. It is true that a search engine needs to correctly identify the file type of every file to determine the relevant parsing specification to apply. But a search engine can figure out the applicable file type from the binary file itself, without reference to the file extension at all. In fact, the file extension is extraneous to this process.

The next misconception is that a mistype will thwart a search engine. Say you mistype Mississippi in an email, maybe adding or deleting an extra S or mistyping a P as a Q. But fuzzy searching adjusts from 1 to 10 to accommodate text deviations. Even a low level of fuzzy searching would pick up any of these Mississippi mistypes. Fuzzy searching works alongside other search types, like Boolean and/or/not searching and proximity searching, so it is easy to just keep fuzzy on at a low-level while searching.

Why not leave fuzzy searching on at a high level? While a higher level of fuzzy searching will pick up the largest numbers of typographical and OCR deviations, it also finds false hits. At some point, Mississippi with a high enough level of fuzziness is also going to pick up Missouri, so it's a trade-off.

The next misconception is that text that is obscure in an associated application display will be equally unapparent to a search engine. If you look at a standard file—PDF, Word, Excel, Access, PowerPoint, OneNote, etc.—in its native or associated application, white text against a white background, black text against a black background and the like can be very hard to spot. But in binary format, black on black or white on white is just as apparent as regular black on white writing.

Likewise, certain metadata is easy to miss in an associated application in that it can take a whole lot of clicking around before you even realize it is there. But all metadata is equally apparent in the binary format of a file. Similarly, a file can have a recursively embedded document inside of it where only a few lines of the embedded document may be visible by default. But the whole embedded file is easily accessible in a binary format view. A search engine can also handle a multilevel nested file structure, like an email with a ZIP or RAR attachment containing a Word document with an Excel spreadsheet embedded inside.

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®