Enterprise Search: Myth vs Reality

Article contributed by dtSearch®

When you think of a search engine, you probably associate to Google or Bing. Those are great for navigating the public web. But they are not going to let you locate an email exchange from nine years ago or find a footnote reference in millions of office files. For that, you need a different product category, enterprise search.

With enterprise search, one or more concurrent search threads can instantly search terabytes of organizational data, including over 25 different full-text and metadata search options and the display of retrieved items with highlighted hits. Sounds cut-and-dried, does it not? But scratch a bit deeper, and you'll find some myths about enterprise search that are quite at odds with its reality. While some myths are relatively inconsequential, others can have effects that you need to be aware of in terms of the reach of enterprise search.

Myth 1: Searching is resource intensive. In reality, searching – even concurrent searching – uses negligible resources. And online search can run in a completely stateless manner, making it very easy to scale. The step that precedes instant search is resource intensive. To instantly search terabytes, enterprise search first has to index the data. But while the initial indexing consumes system resources, it does not require human intervention. All you need to do is point to the folders, email archives, online data repositories, etc. to index, and enterprise search will take it from there. Further, updating of an index to reflect new, modified or deleted files can occur at regular intervals on a schedule with concurrent searching continuing unaffected.

Myth 2: Enterprise search approaches data in a similar way as you. You probably use the Microsoft Word application to view a Word document, a PowerPoint to display a PowerPoint file, OneNote to view a OneNote file, Access to display an Access database, Excel to see a spreadsheet, a viewer like Adobe Acrobat Reader to see a PDF, an email program to display emails, etc. Enterprise search does none of that, heading straight to the binary formats of files. This binary format access applies both to classic office files and cloud files like Office 365 and certain SharePoint files that appear in the standard Windows folder system but are actually remote.

Myth 3: Misapplied file extensions, such as .DOCX for a PDF, can throw off enterprise search. Underlying this myth is the correct assumption that enterprise search has to definitively identify the file format before parsing a file. A single binary file

Myth 1: Searching is resource intensive.

Myth 2: Enterprise search approaches data in a similar way as you.

Myth 3: Misapplied file extensions, such as .DOCX for a PDF, can throw off enterprise search.

format specification can be hundreds of pages long, and applying the wrong one would not be pretty. But what this myth misses is that enterprise search can look inside of a binary format to determine the applicable file type; the file extension is not relevant.

Myth 4: A nested file configuration, like a ZIP or RAR attachment to an email including a Word document with an Excel spreadsheet embedded inside, can obscure some contents. Just as enterprise search uses the binary format for its initial file format identification, it can also use the binary format to identify nested file situations. You may not see the full text of a nested Excel spreadsheet from within Microsoft Word, but the whole thing will be available to enterprise search in binary format.

Myth 5: If you don't see text in a file, enterprise search won't see it either. Because enterprise search approaches files in their binary format, it has a much more comprehensive view of files than you would through a standard file view. For example, black text against a black background or white text against a white background may look invisible inside a standard file view. In binary format, however, such text is on the same level as any other text. "Hidden" metadata that may take a huge amount of clicking around before you even discover that it is there in a standard file view is immediately apparent in binary format. If a file has track changes that remain in it, even if you may not see these by default in a standard file view, such changes will remain accessible in the binary format and hence to enterprise search.

There is a counterpoint involving text that you can see but enterprise search can't, and that is "image only" PDFs containing an image of text. (You know when you try to copy and paste text from a PDF but nothing copies? That is likely an "image-only" PDF.) Enterprise search can flag these for you following indexing, letting you know that you need to apply an OCR application like Adobe Acrobat to digitize the text. You can then send these back to enterprise search with available text to work with.

Myth 6: Enterprise search offers text retrieval, which is word-based. In fact, in addition to operations like "all words," "any words," word and phrase Boolean (and/or/not) and proximity searching, enterprise search can also extend to numbers. Numeric-oriented search covers searching the full-text plus metadata (or metadata only) for specific numbers, numeric ranges, dates and date ranges (even automatically extending across different date formats), hash values, and even certain numeric sequences. For example, enterprise search can identify credit card numbers that may be in the data. After a search, just as with an ordinary word and phrase search, enterprise search can display a full copy of retrieved files with highlighted hits.

Article contributed by dtSearch®

Myth 4: A nested file configuration, like a ZIP or RAR attachment to an email including a Word document with an Excel spreadsheet embedded inside, can obscure some contents.

Myth 5: If you don't see text in a file, enterprise search won't see it either.

Myth 6: Enterprise search offers text retrieval, which is word-based.