Hidden Data Makes the News

Article contributed by dtSearch®

Enterprise search finds not only text openly in files, but also on occasion content that someone might have tried to hide. For most people, the notion of hidden content represents more of an abstraction than something they might expect to encounter. But abstraction no more, as some retrievable redactions in government disclosures recently made the news. Before I get into the details, a quick overview of how enterprise search works will help frame the issue.

How does enterprise search work?

Enterprise search allows one or more concurrent users to instantly search terabytes. It does so, however, only after initially indexing the data. The data can include both local and remote files like Office365 and DropBox files as well as SharePoint attachments. In indexing such remote content, dtSearch's off-the-shelf Windows product simply needs to see the remote files as part of the Windows folder system. Indexing further works with not only standalone files but also recursively nested files, like an email with a ZIP or RAR attachment holding an Excel spreadsheet which itself embeds a Word document.

How easy is it to index data?

Indexing couldn't be easier. With dtSearch, just point to the folders, email archives and the like to cover and the indexer will take it from there. A single index can hold up to a terabyte of text and there are no limits on the number of indexes that the indexer can create and end-users instantly concurrently search. Data evolves with file additions, deletions and replacements. Fortunately, dtSearch can update its indexes without affecting continued individual or concurrent searching.

So what does indexing actually do?

Indexing stores each unique word and number and the place of each in the data. In parsing files for indexing, the software will *not* individually retrieve each file in its associated application. Instead, the indexer will

The network option has some advantages for those in-office, while a web-based and potentially cloud-based approach has benefits for those operating remotely. Whether network-based, intranet-local server or cloud-based, enterprise search offers instant concurrent searching only after first indexing the data.

go straight to each file's binary format. Direct binary access is of course speedier than individually opening each file in its associated application. More importantly, direct binary access makes indexing a lot more thorough, including coverage of hidden text.

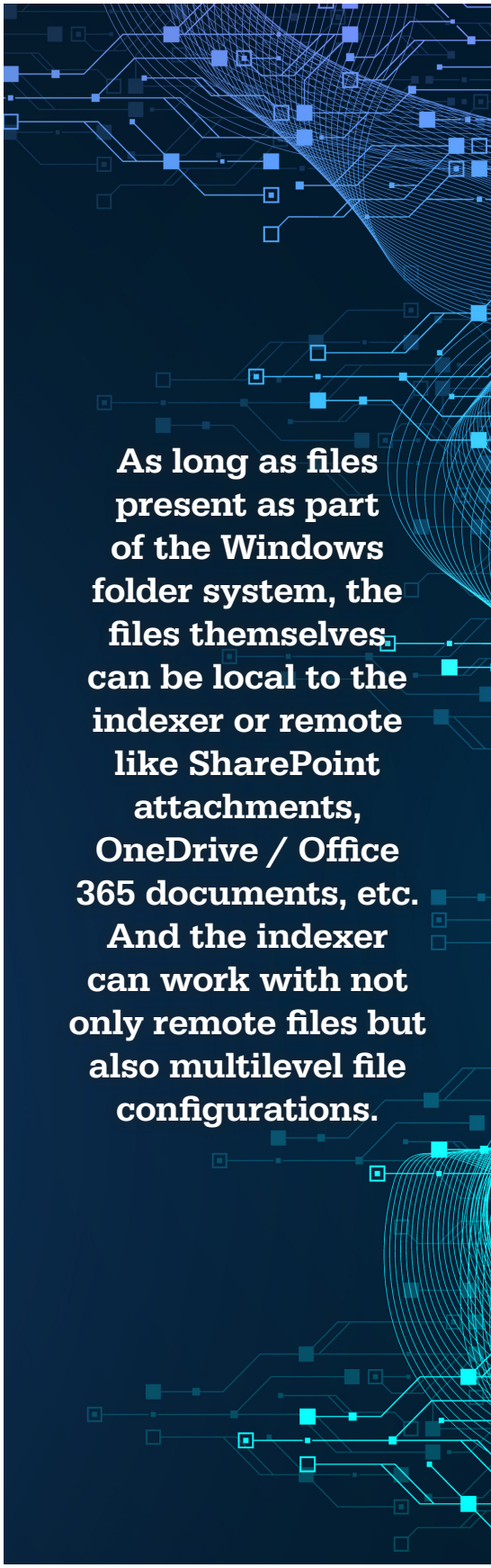
How so?

One way to hide content, either intentionally or unintentionally, is to save a file with a mismatched file extension, like saving a PDF not with a .PDF extension but say with an email file extension. Another way to hide information is to save it as obscure metadata. You can click around extensively in a file's associated application display and still miss certain metadata. The associated application display also masks text that blends in with its background color like white text against a white background or black text against a black background. If a file has pending but not fully accepted tracked changes, the default associated application display can show the text minus the deletions while the deletions remain as part of the overall file. Similarly, some redaction applications put a black rectangle over text to delete, making it look like the text is gone when it is merely hidden underneath the black rectangle.

But enterprise search can uncover such hidden text?

Yes. Let's start with mismatched file extensions. While the indexer needs to accurately identify each file's exact file type to apply the correct parsing specification, the indexer can use the information inside the binary format itself for this determination. That way, it won't matter if you have a PDF with an email file extension, an Access database with an Excel extension, or a OneNote file with a PowerPoint extension. Also, all metadata, no matter how hard to find in an associated application view, is readily apparent in a file's binary format. Text that blends in with its background color in an associated application view is on the same footing as any other text in a file's binary format. Pending tracked changes are immediately available in a file's binary format as well.

Article contributed by dtSearch®



What about the retrievable redactions?

Some government disclosures included portions that a redaction program had blacked out. If the files were printed on a printer, the blacked-out portions would have remained invisible. But the files were produced as PDFs, with the original text remaining underneath the black rectangles. Although not visible from inside most PDF viewer displays, an end-user could still copy and paste the redacted text into another file. Because enterprise search accessed the binary formats of these PDFs, the blacked-out text was further available for searching.

Can you quickly go over the search options that dtSearch provides?

dtSearch has over 25 different search features ranging from free-form natural language search options to highly structured phrase, Boolean (and/or/not) and proximity expression handling. Enter a query across all file text or limit certain search elements to specific metadata. Fuzzy searching can sift through typographical or OCR errors. Concept searching can extend a search request to synonyms. Searches can also include individual numbers or numeric ranges as well as individual dates or date ranges extending to common date variants. dtSearch can even flag credit card numbers across indexed data.

And other search features?

For multilingual text, dtSearch leverages Unicode to support hundreds of international languages, including European languages with varying alphabets, double-byte Asian characters and right-to-left Middle Eastern text. dtSearch has numerous options for relevancy-ranking and other search result sorting—as well as options for instant re-sorting by a completely different metric. And dtSearch can display the full text of retrieved files with highlighted hits for easy search results navigation.

Final thoughts?

dtSearch.com has fully-functional 30-day evaluation downloads to start your organization now on instant concurrent searching across terabytes. Or just download the government files and see what you can find.

Article contributed by dtSearch®

dtSearch supports Unicode spanning hundreds of international languages. A single file can cycle through multiple languages, including not only different European languages but also right-to-left Hebrew and Arabic and double-byte Chinese, Japanese and Korean. dtSearch and Unicode will follow that whole progression.