

All Text—Even Invisible—Factors Into Enterprise Search

Article contributed
by [dtSearch®](#)

Ripped from the news: watch out for invisible text. Text you can't see can hurt you. A professor gave the class an essay assignment. Many students just copied and pasted the assignment directly into AI. But as it turned out, the visible text was not the key; it was the invisible text that was critical.

Mixed into the visible text were additional words that blended in with the background color. While the blended words weren't visible to the students, when the students copied and pasted the assignment into AI, the AI read the invisible words as part of the question. The invisible verbiage told the AI to add a specific nonsense word to its answer. When a student's submission included the nonsense word, the professor could see that the student had fed the assignment into AI.

Now you may not be a student. But the impact of invisible text isn't limited to academia. All text visible or otherwise factors into enterprise search. Enterprise search like dtSearch® enables instant concurrent multithreaded searching across terabytes. But it does so only after initially indexing that data. The indexing process recognizes all unique words and numbers and tabulates the location of these across the data. A single dtSearch index can hold up to a terabyte of text and there are no limits on the number of indexes that the software can create and make available for instant simultaneous searching.

While indexing is highly resource intensive, concurrent searching is very resource-light, enabling instant multithreaded searching. dtSearch can further update its indexes to reflect new files, file edits or file deletions without disrupting ongoing instant multithreaded searching. And while indexing will always be more time consuming than searching, a 64-bit multithreaded indexer can speed up the process a lot.



**Ripped from
the news: watch out
for invisible text.
Text you can't see
can hurt you.**

Getting dtSearch started on indexing is easy. Just point to the email archives and other folders to index and the indexer will automatically take it from there. So long as the files appear in the Windows folder system, the files themselves can be local or remote, including One Drive/ Office 365 files, DropBox files, SharePoint attachments, etc.

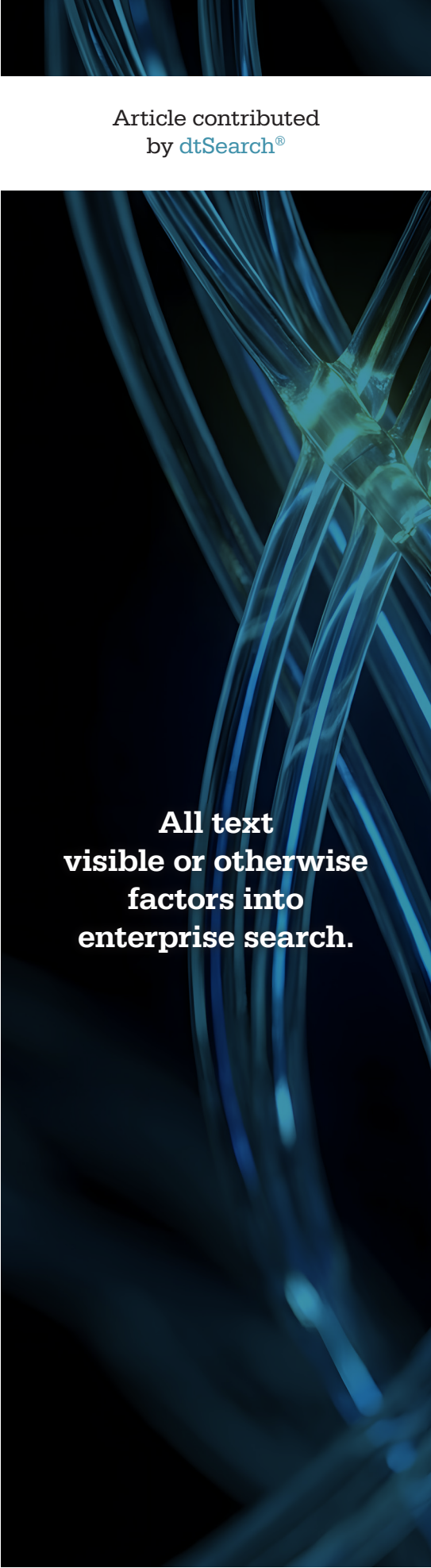
To review a file, you'd typically pull it up in an associated application like a PDF in Adobe Reader, a spreadsheet in Microsoft Excel, a database in Access, a word processing document in Word, a presentation in PowerPoint, a note file in OneNote, web-ready content in a browser, an email in an email program, etc. But for the indexer, pulling up each file in its associated application would be way too slow. Instead of pulling up each file in its associated application, dtSearch heads to each file's binary format to parse it.

This binary format approach affects the scope of enterprise search, including with respect to hidden text. To correctly parse a binary format, the indexer needs to figure out its exact right file type. Of course, it's easy to save a file with an extension that does not match the file type, like saving a Word file with a .PDF extension or saving a PDF with a .DOCX extension. But the binary format has all of the file type information inside of it, enabling file type recognition regardless of file extension.

The binary format further lets the enterprise search indexer correctly parse not only standalone files but also nested files. An email can have a ZIP or RAR attachment containing an Excel spreadsheet with a recursively embedded Word document and the binary format will make all that available to the indexer for parsing. There is a lot of obscure metadata that is easy to miss viewing a file in its associated application. But all metadata is readily apparent in a file's binary format, and easily accessible to the indexer.

The binary format further covers visible text that you'd easily spot viewing the file in its associated application as well as camouflaged text like the professor used, blended into the background color. In fact, all text regardless of contrast is on the same footing in the binary format. Also in the binary format are tracked

Article contributed
by [dtSearch®](#)



**All text
visible or otherwise
factors into
enterprise search.**

changes if not fully accepted and deleted. Similarly, a black rectangle can block out text, making it appear redacted. But the text itself can remain in the binary format for the indexer to access.

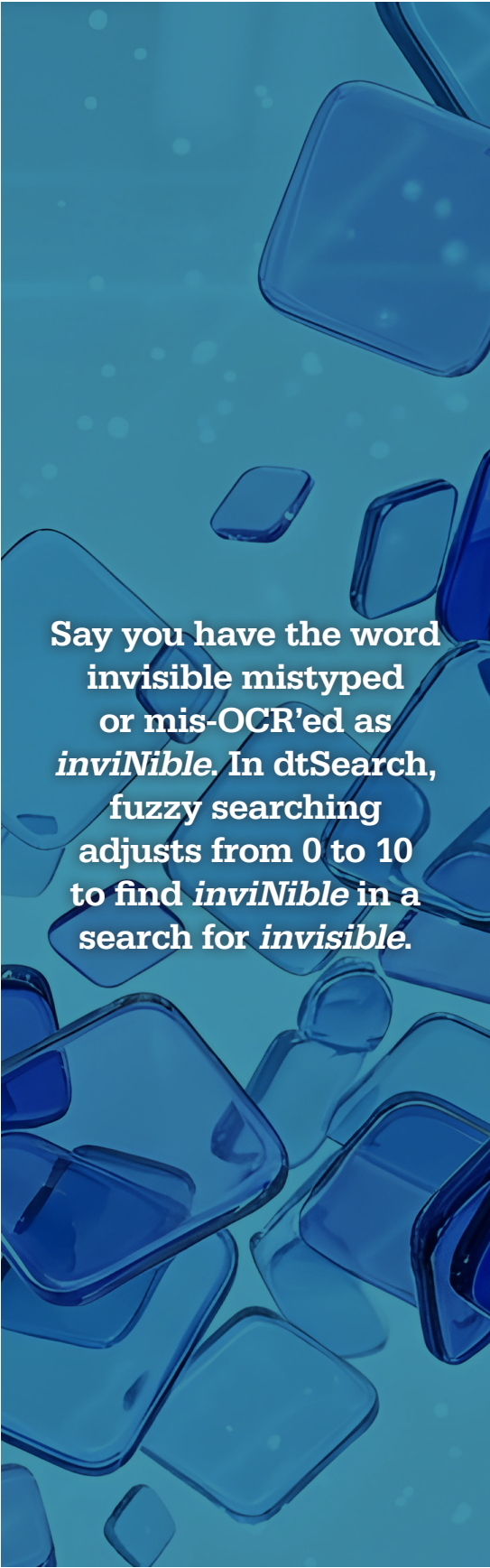
For international language coverage, the binary format encompasses all of a file's Unicode encodings. A single email or other file can go from one language to the next including English, other European languages, right-to-left Middle Eastern languages and even double-byte character Chinese, Japanese and Korean. The binary format will include all of that encoding information for the indexer to pick up.

After indexing, dtSearch has over 25 different full-text and metadata instant concurrent search types, including natural language, Boolean and/or/not, proximity, numeric range, regular expression, credit card number, hash value and other search options. Some of these 25+ concurrent search options are specifically configured to sift through common data deviations. Say you have the word *invisible* mistyped or mis-OCR'ed as *inviNible*. In dtSearch, fuzzy searching adjusts from 0 to 10 to find *inviNible* in a search for *invisible*.

Another example of sifting through common data deviations would be date and date range searching. A date range search for *date(9/30/26 to 2/15/27)* would pick up not only *1/10/27* but also *January 10 2027* and *Jan 10 2027*. Similarly, concept searching can extend to synonyms from a built-in thesaurus, like *unseeable* for *invisible*. dtSearch also supports custom synonym rings. If you are working on the *Tom Jones* case, while *Tom* and *Jones* may not normally be synonyms, you can make them synonyms for purposes of your case.

Download a fully-functional 30-day evaluation copy from dtSearch.com to enable instant concurrent searching across terabytes of visible, invisible and even *inviNible* enterprise data.

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



**Say you have the word
invisible mistyped
or mis-OCR'ed as
inviNible. In dtSearch,
fuzzy searching
adjusts from 0 to 10
to find *inviNible* in a
search for *invisible*.**