Enterprise Search — Ghostly Data Edition

Article contributed by dtSearch®

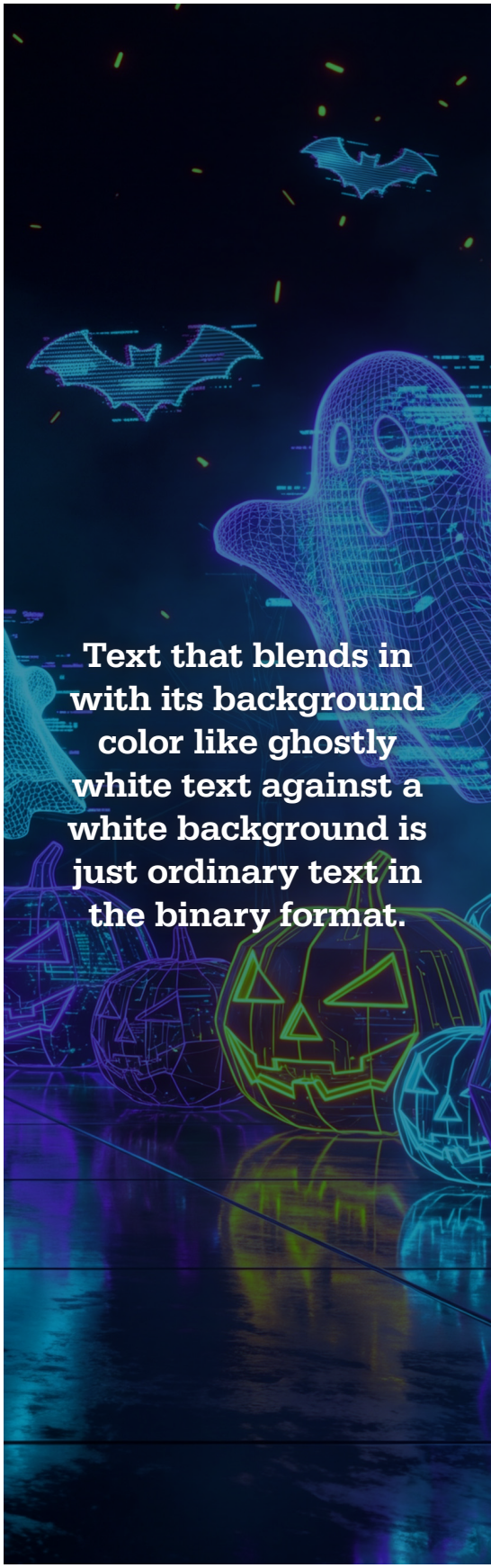
With Halloween coming, today's topic is ghostly data as it relates to enterprise search like dtSearch®. 👻 Tracked changes, for example, can leave a ghostly residue. You may think that certain text is gone. But in a file with tracked changes on but not definitively accepted, the removed text can still haunt the file. 👻 A supposedly deleted file can also linger as ghostly data in the form of a recursively embedded attachment. You deleted that Excel spreadsheet as a standalone file. But it can live on undead in a ZIP or RAR email attachment including a Word document which itself embeds a copy of the supposedly deleted spreadsheet.

Any other ghostly data examples?

👻 Ghostly white text against a white background, ghostly black text against a black background or ghostly orange text against an orange background is another spooky example. While such text may not be visible in a file's associated application, it doesn't mean it isn't there. 👻 Obscure metadata is another ghostly data example. You can click around extensively in an associated application view of a file and still miss certain metadata. 👻 A file with a mismatched file extension like a Word document saved with a .PDF extension or a PDF saved with a .DOCX extension can also become ghostly data. Such a file is easy to miss when reviewing specific types of files, but it can still haunt your archives.

So how does enterprise search find these ghosts?

Enterprise search like dtSearch instantly sifts through terabytes of data using multiple concurrent search threads. But it does so only after indexing the data. And here's the key. When enterprise search builds its indexes, it utilizes the binary format of each file. While the indexer has to ascertain the file type of each file to correctly parse it, the indexer can use the binary format alone for this step. That way, it doesn't matter if a file has a mismatched file extension.



Reprinted with permission of USA Business Radio, usabusinessradio.com

Does the binary format access help with the other ghostly data examples as well?

Text that blends in with its background color like ghostly white text against a white background is just ordinary text in the binary format. The binary format will further include text marked for deletion absent the definitive acceptance of tracked changes. Regardless of how hard it is to spot in a file's associated application, all metadata is right there in the binary format. The binary format also enables parsing of recursively nested files like the email with a ZIP or RAR Word document that itself embeds an Excel spreadsheet.

What does it take to get the software to build an index across these binary formats?

Just point to the folders to cover and the indexer will take it from there. So long as files appear as part of the Windows folder system, it doesn't matter if the files themselves are remote like DropBox files, SharePoint attachments or Office 365 / OneDrive files. A single dtSearch index can hold up to a terabyte of local and remote email and other file text, and there are no limits on the number of indexes the software can build and instantly concurrently search. As data evolves, index updates reflecting enterprise data changes can proceed automatically without affecting ongoing instant concurrent searching.

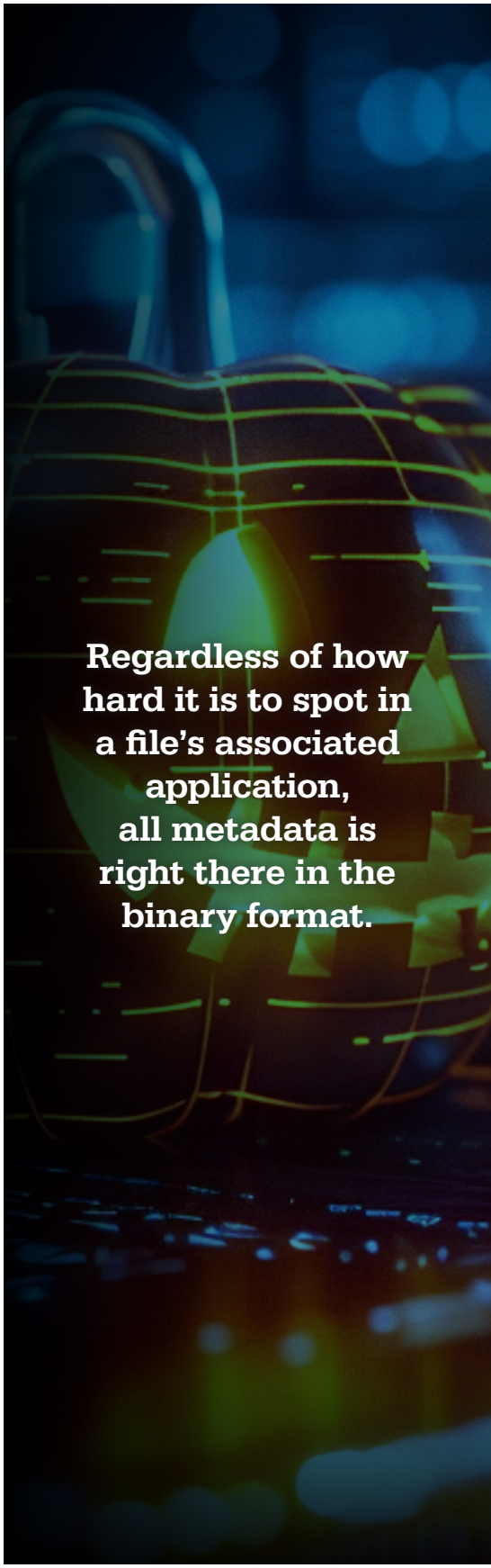
What about international languages?

For international languages, the binary format covers a file's Unicode encodings. A single email or other file can cycle through English, other European languages with different alphabets, right-to-left Middle Eastern text, double-byte Asian text and other languages, and the indexer will pick up all of that. The binary format even includes Unicode emojis like a jack-o'-lantern 🎃 for Halloween.

And general search options?

dtSearch has over 25 different instant concurrent full-text and metadata search features to sift through spooky and non-spooky enterprise data. Enter a simple "all words" or "any words" natural language search. Or engage precision searching with phrase, Boolean (and/or/not) and proximity requirements. Add on

Article contributed by dtSearch®



Reprinted with permission of USA Business Radio, usabusinessradio.com

concept searching to extend a search for *ghostly* to *spectral*. Fuzzy searching adjusts from 0 to 10 to sift through possible typographical or OCR deviations like *Hallomeen* for *Halloween*.

And other search options?

Searching can also cover numbers or numeric ranges as well as dates and date ranges, either across all text or just in specific metadata. A search from *10/20/26 to 11/5/26* would find *10/31/26* as well as *October 31, 2026* and *Oct 31 2026*. dtSearch can also generate hash values for all indexed files and then search for specific hashes. Despite best efforts to purge enterprise data of credit card numbers, some undead instances may remain. dtSearch can flag any credit card numbers haunting enterprise data. And dtSearch can literally search for ghosts in the form of a Unicode ghost emoji 👻.

How does dtSearch present search results?

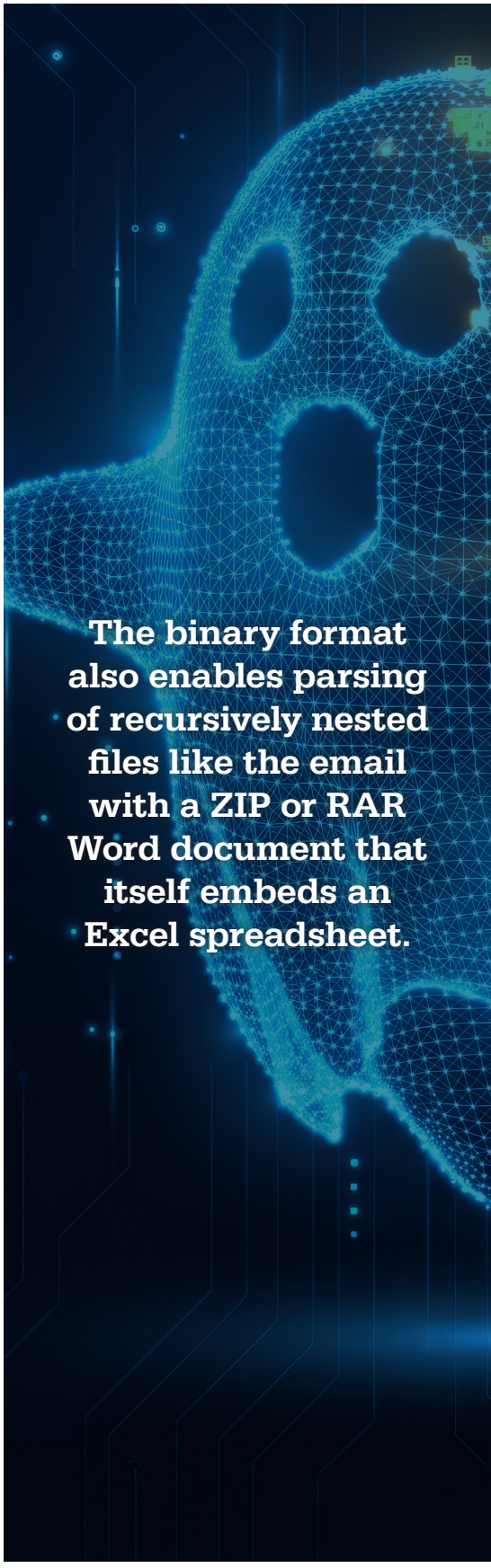
Default relevancy ranking prioritizes files based on the rarity and density of search terms in indexed data. Take an "any words" search for *chocolate lollipops licorice*. If *chocolate* and *lollipops* are common in indexed data but *licorice* rare, then *licorice* mentions will get a higher relevancy rank. Files with the densest *licorice* occurrences will come out on top. Or use custom variable term weighting like assigning *chocolate* a positive weight of 4, *lollipops* a positive weight of 6 and *licorice* a negative weight of 9.

Weighting can apply to search terms anywhere or prioritizing appearances in certain metadata or towards the top of bottom of files. For a different view of search results, re-sort via a new metric like file date or file location. Search results show full copies of retrieved files with highlighted hits for convenient browsing. Hit highlights can be all one color, or each search term can have its own highlight color.

Final thoughts?

Halloween is a great time to pay attention to ghosts in your enterprise data. dtSearch.com has fully-functional 30-day evaluation downloads to instantly and concurrently search across terabytes of haunted and non-haunted enterprise data.

Article contributed by dtSearch®



Reprinted with permission of USA Business Radio, usabusinessradio.com