This Halloween, Know Your Undead Data

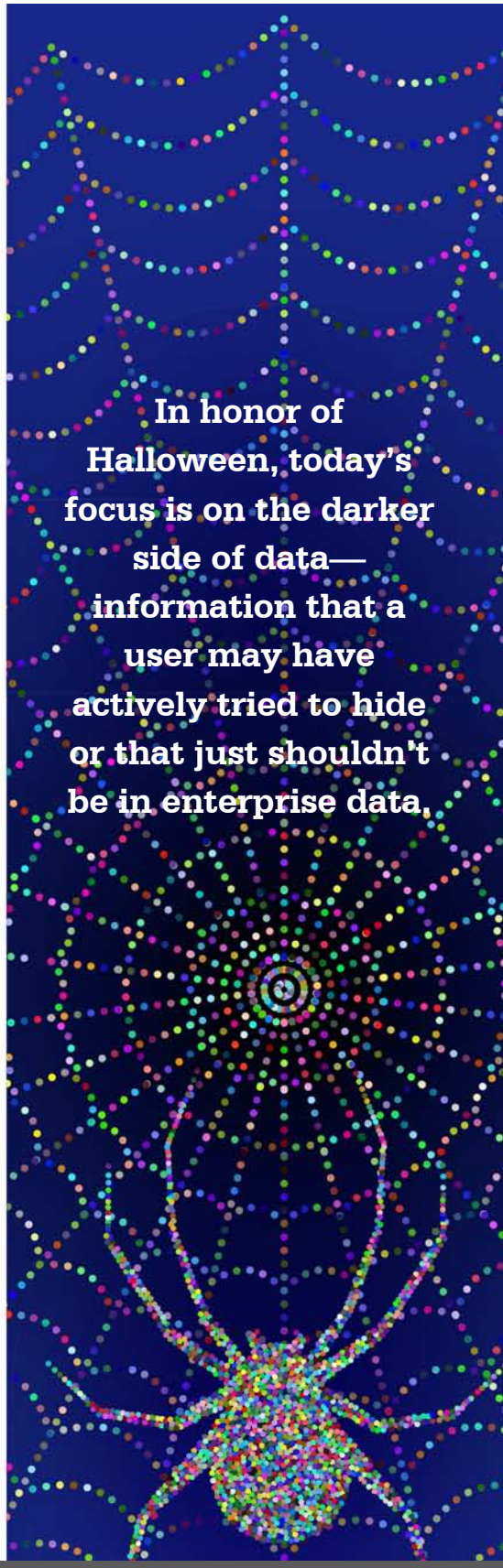
Article contributed by dtSearch®

Normally, enterprise search focuses on the positive side of data. Enterprise search lets multiple end-users simultaneously locate content an organization wants them to find to move forward productively with their day. In honor of Halloween, today's focus is on the darker side of data—information that a user may have actively tried to hide or that just shouldn't be in enterprise data. Let's call this undead data.

Undead data retrieval operates the same way as any other data retrieval. Following is a quick overview of how enterprise search works generally before getting into some specific undead data examples. Enterprise search like dtSearch® can instantly search terabytes only after it first indexes the data. Indexing pre-digests each unique word and number and its location in the data. A single dtSearch index can hold up to a terabyte of text, and there are no limits on the number of indexes dtSearch can build and end-users can instantly concurrently search.

Indexing is easy. Just point to the folders, emails, etc. that you want to cover, and dtSearch will do the rest. For efficiency, the indexer approaches each file in its binary format rather than pulling up each in its native application. Parsing a binary format requires identification of its exact right file type. PDF, Microsoft Word, Access, Excel, PowerPoint, OneNote, web-based files, email files and other file formats all have radically different parsing specifications.

Fortunately, the indexer can automatically identify the file type through the binary format. Indexing can even work seamlessly with online content like SharePoint attachments and Office365 files so long as these present as part of the Windows folder system. After indexing, end-users can concurrently search using over 25 different full-text and metadata search features. Enterprise search displays a complete copy of retrieved items with highlighted hits for easy navigation.



Reprinted with permission of Daily Telegraph USA, **dailytelegraphusa.com**

While indexing takes a lot of resources, searching does not, allowing multiple concurrent search threads to proceed at once. Online search can run statelessly, making multithreaded queries easy to scale. And updating an index to accommodate files that have been added, deleted or changed will not affect concurrent searching.

The above represents a quick overview of how enterprise search generally works. Now for the undead data examples.

Undead Data #1: The first undead data example is a file with a filename that obscures the actual file type, like a PDF saved with a .DOCX extension or a Word document with a .PDF extension. This file extension mismatch can impede general access to the file. But because dtSearch uses information inside each binary file for file type identification rather than relying on the file extension, such a file will be fully undead from the perspective of enterprise search.

Undead Data #2: An email with a ZIP or RAR attachment can contain a bunch of files, including an Excel spreadsheet and a Word document recursively embedded in the spreadsheet. That Word document may easily slip through the cracks of normal everyday operations. However, enterprise search can go through compressed and other multilevel attachments down to the innermost nested files, changing their status to undead.

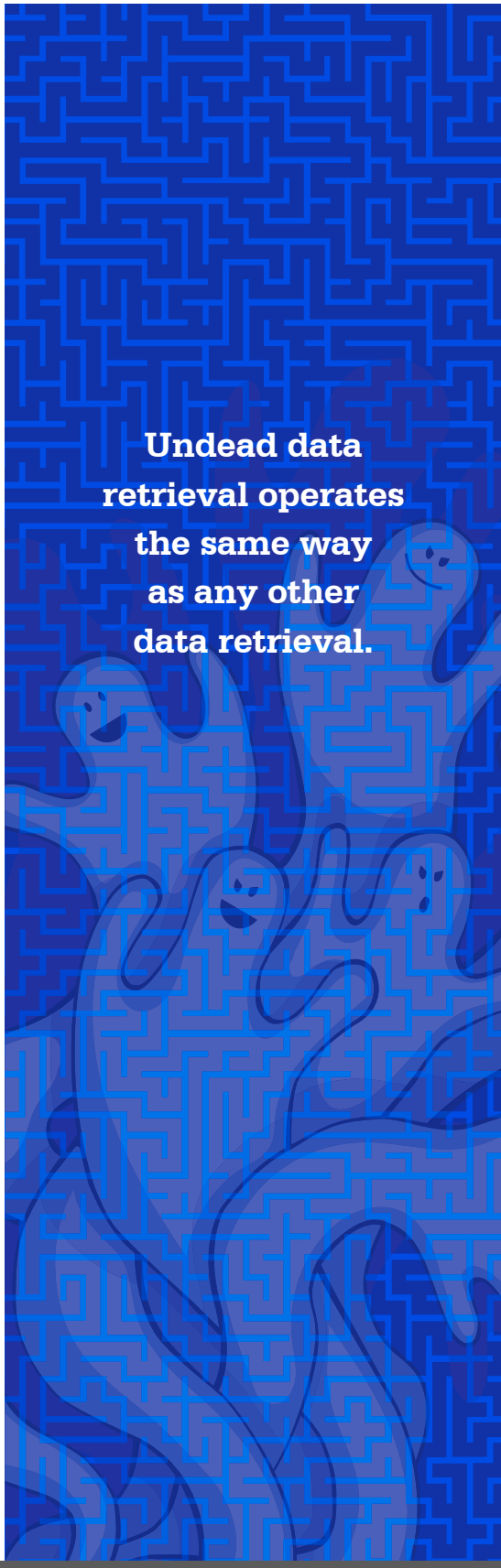
Undead Data #3: Sometimes a file can have text that merges with the background color, like black writing against a black background. Such text is all but invisible looking at a file from inside a file's native application. But in binary format, all text is on the same level regardless of contrast colors, making camouflaged text fully apparent and undead for enterprise search.

Undead Data #4: A file can also have tracked changes that may not by default present themselves in a native application view of the file. However, if not fully removed, such tracked changes can remain in the binary format and are hence undead for enterprise search.

Undead Data #5: Files can have obscure metadata that an end-user might never happen upon in the file's native application. But all metadata, no matter how hard to spot in

Article contributed by dtSearch®

Undead data retrieval operates the same way as any other data retrieval.



a file's native application, is fully present in the binary format and thus undead for enterprise search.

Undead Data #6: Enterprise data often includes image-only PDFs. These may look at the folder level just like ordinary PDFs. But when you are in a PDF viewer and try to copy and paste text from it, no actual text will copy out. Because such files consist of an image, enterprise search will only be able to index the filename and the metadata, not the full text. But dtSearch can flag such files during the indexing process, so you know to run them through an OCR program like Adobe Acrobat Reader and then bring them back to the dtSearch indexer. In this way, the full-text contents go from dead to undead.

Undead Data #7: Sometimes an OCR engine can mis-OCR a word or an end-user can mistype a word say in an email such as *Hallomeen* for *Halloween*. Fuzzy search adjusts from 1 to 10 to sift through such typographical errors, bringing them back from the dead.

Undead Data #8: The last example is information like credit card numbers that should be dead and gone, but can sometimes creep into open enterprise data. dtSearch has an option to take any sequence of digits that may represent a credit card number and run them through an internal credit card checking algorithm. When these do represent a credit card number, dtSearch can flag it, letting you stay on top of undead credit cards that may slip into general enterprise data.

In sum, use a search engine like dtSearch to know your data—living or undead.

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 can download a fully-functional 30-day evaluation copy from dtSearch.com to enable instant concurrent enterprise search across terabytes of data.

Article contributed by dtSearch®