

For the Dark Days of Winter, Let's Focus on "Dark Enterprise Data"

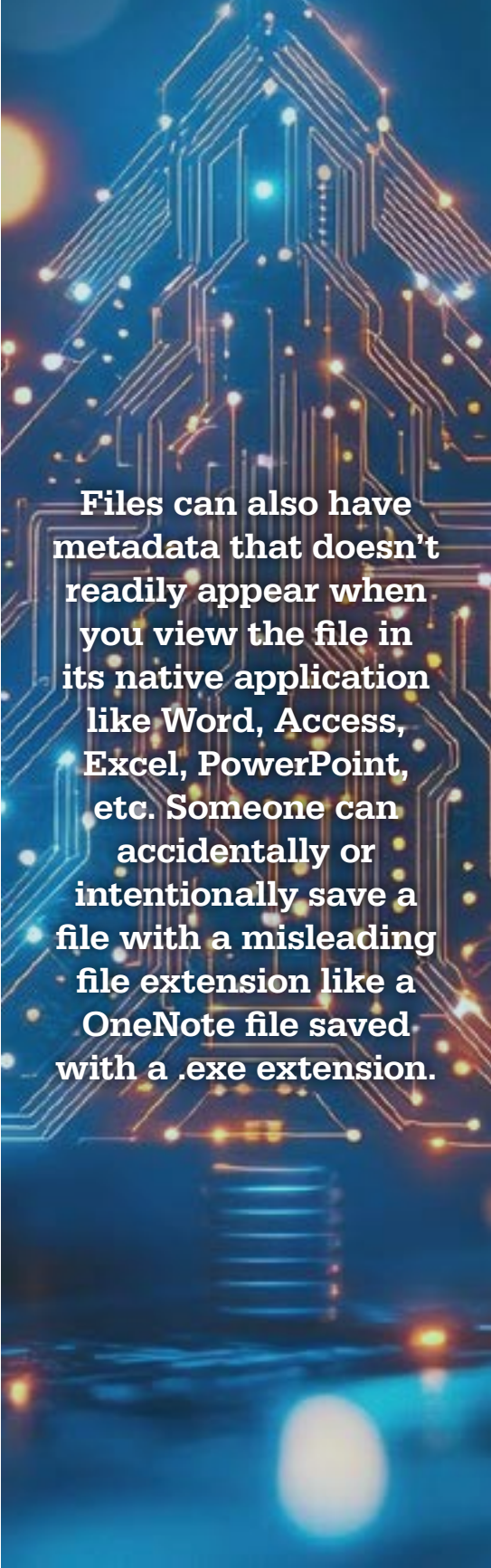
With a polar vortex covering most of the country, the continuing dark days of winter seem like a good time to focus on "dark enterprise data," or murky hard-to-spot content hiding in everyday data. For example, recursively embedded file formats are a Microsoft Office feature. But viewing a Word document in Microsoft Word, you might see just a few rows of an embedded Excel spreadsheet, missing most of the incorporated content.

Emails can have container attachments like ZIP or RAR which can themselves include containers like a PDF Portfolio. Files can also have metadata that doesn't readily appear when you view the file in its native application like Word, Access, Excel, PowerPoint, etc. Someone can accidentally or intentionally save a file with a misleading file extension like a OneNote file saved with a .exe extension.

Or someone can add text to a file that blends in with its background color like slate gray lettering against a slate gray background. Some redaction programs add a black rectangle over text. But while visually blacked out, the text itself can remain underneath the black rectangle. Tracked changes can also show files without deleted text. But if the file editor doesn't fully accept the changes, the deletions can remain as part of the total record.

Also, have you ever looked at what seemed like an ordinary PDF inside a PDF viewer like Adobe Acrobat Reader but when you tried to copy and paste some text, nothing copied out? That was likely an image-only PDF. While it may visually appear like a regular PDF, an image-only PDF has no underlying digital text to copy and paste or otherwise work with. Enterprise search and "dark enterprise data." Enterprise search like dtSearch® bypasses retrieving files in their native

Article contributed
by dtSearch®



Files can also have metadata that doesn't readily appear when you view the file in its native application like Word, Access, Excel, PowerPoint, etc. Someone can accidentally or intentionally save a file with a misleading file extension like a OneNote file saved with a .exe extension.

applications and instead looks to binary formats. To correctly parse a file, enterprise search needs to pinpoint the file type. Fortunately, the binary format lets enterprise search correctly detect the file type regardless of the file extension.

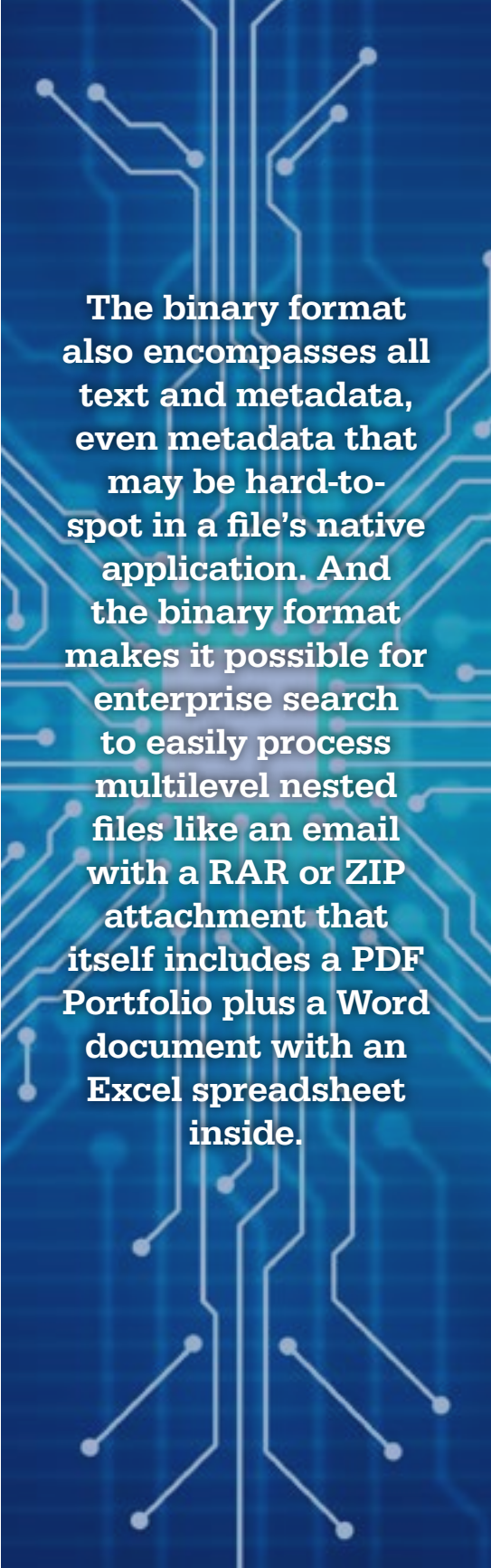
The binary format also encompasses all text and metadata, even metadata that may be hard-to-spot in a file's native application. And the binary format makes it possible for enterprise search to easily process multilevel nested files like an email with a RAR or ZIP attachment that itself includes a PDF Portfolio plus a Word document with an Excel spreadsheet inside.

Text that blends in with the background color like slate gray text against a slate gray background is on the same footing as any other text in binary format. Visually redacted text that may remain in hiding in a native application view of a file is also fully searchable. The binary format even lets enterprise search detect whether a PDF is "image only." dtSearch can flag all such files so you know to run them through an OCR program like Adobe Acrobat.

How enterprise search works. Enterprise search first has to index the data before providing instant searching. Getting dtSearch to index the data is simple. Just check off the folders, email archives, etc. to index and the indexer will take it from there. Using binary formats, the indexer compiles all unique words across the data and records the words' locations in the data. In addition to working with local files, the indexer can also automatically work with files that, while appearing as part of the Windows folder system, are actually remote like SharePoint attachments, OneDrive / Office 365 files, DropBox files and the like.

A single dtSearch index can hold up to a terabyte of text and metadata, and there are no limits on the number of indexes dtSearch can build and multiple end-users concurrently instantly search. Indexed search can proceed in a classic Windows network environment, through an "on premises" intranet or internet server, or via a cloud server such as Azure or AWS. Whatever the environment, index updates do not stop instant multithreaded searching so it is easy to keep indexes current.

Article contributed
by dtSearch®



The binary format also encompasses all text and metadata, even metadata that may be hard-to-spot in a file's native application. And the binary format makes it possible for enterprise search to easily process multilevel nested files like an email with a RAR or ZIP attachment that itself includes a PDF Portfolio plus a Word document with an Excel spreadsheet inside.

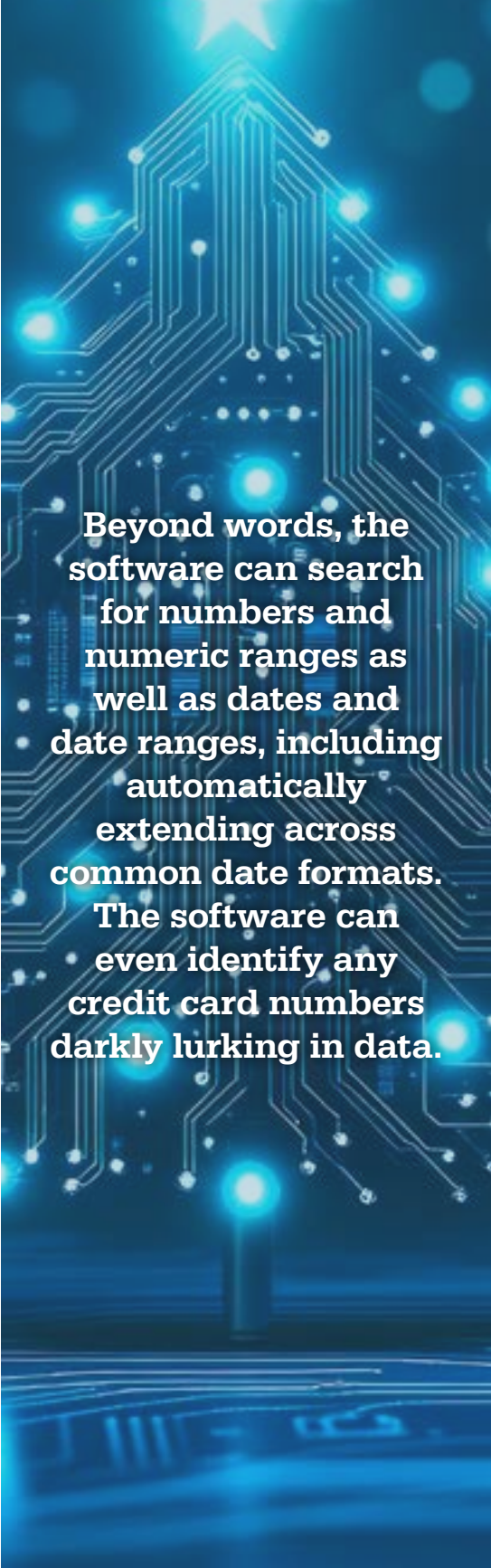
Article contributed
by dtSearch®

Search options. dtSearch, for example, has over 25 different search features. Options range from free-form natural language “all words” or “any words” searches to intricate Boolean (and/or/not), phrase and proximity searches. Query all text, or limit search elements to specific metadata. Concept searching finds words with similar meanings. Fuzzy searching adjusts from 1 to 10 to sift through typographical and OCR errors.

Beyond words, the software can search for numbers and numeric ranges as well as dates and date ranges, including automatically extending across common date formats. The software can even identify any credit card numbers darkly lurking in data. Binary formats include Unicode encodings underlying hundreds of international languages. A single email or other file can span English, other European languages, right-to-left Hebrew or Arabic, and double-byte Asian text, and all will be searchable.

dtSearch defaults to relevancy ranking weighing retrieved files according to the density and rarity of search terms across indexed data. Or end-users can apply their own positive or negative variable term weighting with optional adjustment for search term appearance in certain metadata or at the top or bottom of a file. For a fresh perspective on search results, the software can instantly re-sort by a completely different criterion like file date or file location. Whatever the sorting, search results show the full text of retrieved files with highlighted hits for convenient review.

So let dtSearch enable instant concurrent search across all the terabytes of your enterprise data, dark or light, online and offline. Visit dtSearch.com for fully-functional 30-day evaluation downloads.



Beyond words, the software can search for numbers and numeric ranges as well as dates and date ranges, including automatically extending across common date formats. The software can even identify any credit card numbers darkly lurking in data.