Keeping Up With Exponential Data Growth

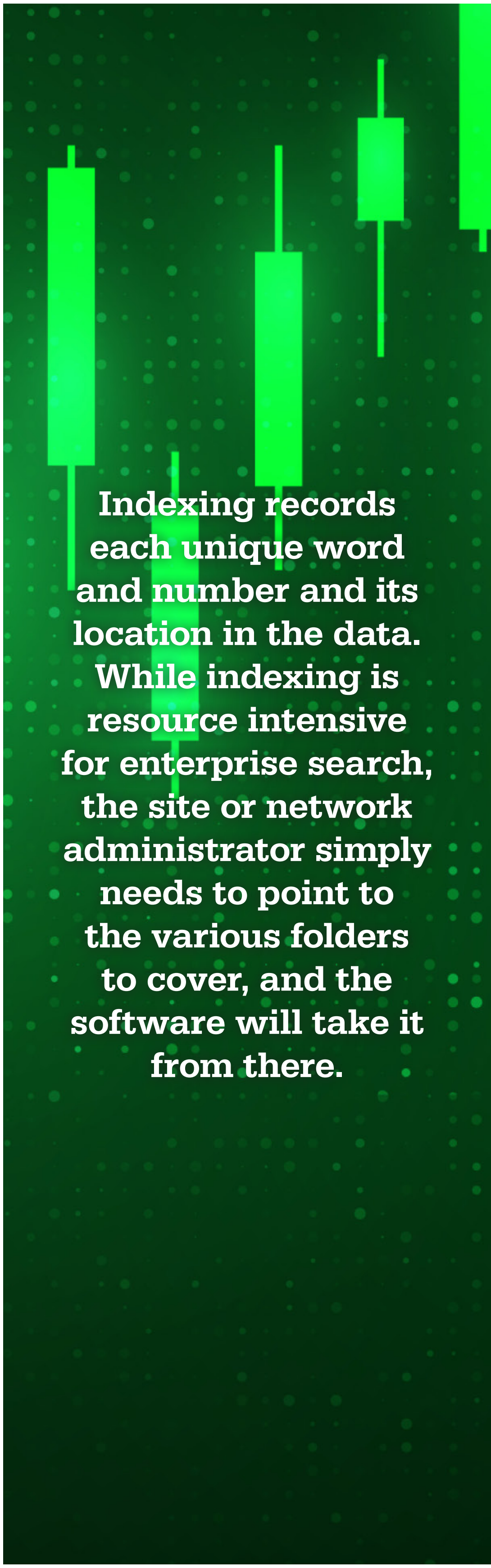
Article contributed
by dtSearch®

Data growth is exponential. To keep up, enterprise search needs to operate as a dynamic product line, not a static one. But first, a quick overview of how enterprise search works. Enterprise search can run from a Windows network, from a local web server or from the cloud like Azure of AWS. However the software runs, before offering instant concurrent searching across terabytes, enterprise search must index the data.

Indexing records each unique word and number and its location in the data. While indexing is resource intensive for enterprise search, the site or network administrator simply needs to point to the various folders to cover, and the software will take it from there. With dtSearch®, for example, the files themselves can be local or remote like SharePoint, Office365, DropBox and the like. The indexer only needs to see them as part of the Windows folder system.

And it is not just standalone local and remote files that the indexer can work with. The data can include an email with a ZIP or RAR attachment with a Word document inside and an Excel spreadsheet embedded within that, and the indexer will cover all of that content. A single dtSearch index can hold up to a terabyte of text. The software has no built-in limits on the number of terabyte indexes that it can generate and end-users simultaneously search.

For efficiency in working with large volumes of data, the indexer processes files in their binary formats rather than retrieving individual files in their associated applications. (The latter would be way too slow.) To correctly parse each binary format, the indexer needs to first figure out the exact right file type. Of course file extension is not a reliable indicator of file type as it is all too easy to save a file with a mismatched file extension or even no file extension.



Indexing records each unique word and number and its location in the data. While indexing is resource intensive for enterprise search, the site or network administrator simply needs to point to the various folders to cover, and the software will take it from there.

Reprinted with permission of USA Business Radio, usdailyradio.com

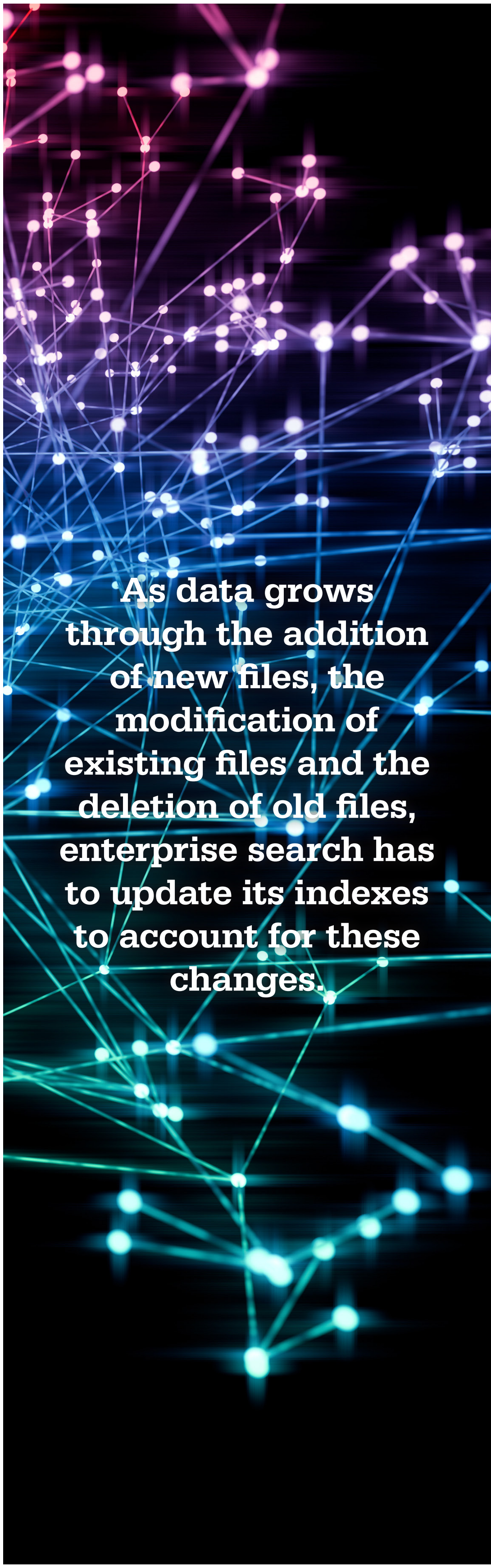
The more reliable approach to determining file type is to use information inside the binary format itself. Once the indexer has the correct file type, the indexer uses its so-called document filters to parse each binary format to identify all words and numbers and their locations. In parsing files, the document filters have to recognize not only the main text of files but also all metadata, including obscure metadata that can be very hard to spot in a file's associated application.

And this brings me to the first way in which enterprise search has to be a dynamic proposition, not a static one. Every time Microsoft adjusts its file formats—Office or Office365 Word, Excel, PowerPoint, Access, OneNote, Outlook/Exchange, etc.—enterprise search needs to adjust its document filters to ensure accurate parsing of the updated formats. The same adjustment requirements also apply to non-Microsoft formats like PDF (which added PDF 2.0 format a few years ago), as well as more structured database and other web-ready file formats as these evolve.

And there is another way in which enterprise search has to operate dynamically not statically. As data grows through the addition of new files, the modification of existing files and the deletion of old files, enterprise search has to update its indexes to account for these changes. dtSearch can automate index updates via the Windows Task Scheduler to make it easy to keep indexes current. And index updates can occur without interrupting on-going instant concurrent searching, so there is no down-time for the end-user.

In parsing files, the document filters work with Unicode. Itself an evolving standard, Unicode governs the representation of hundreds of international languages in a wide variety of file formats. Unicode encodings enable enterprise search to work with not only English but also any other covered languages. These include dozens of European languages many with varying alphabets; double-byte Chinese, Japanese and Korean; and right-to-left Hebrew and Arabic. A single file or email can encompass any number of Unicode encodings and enterprise search will pick up all of that.

Article contributed
by dtSearch®

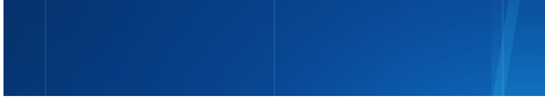


As data grows through the addition of new files, the modification of existing files and the deletion of old files, enterprise search has to update its indexes to account for these changes.

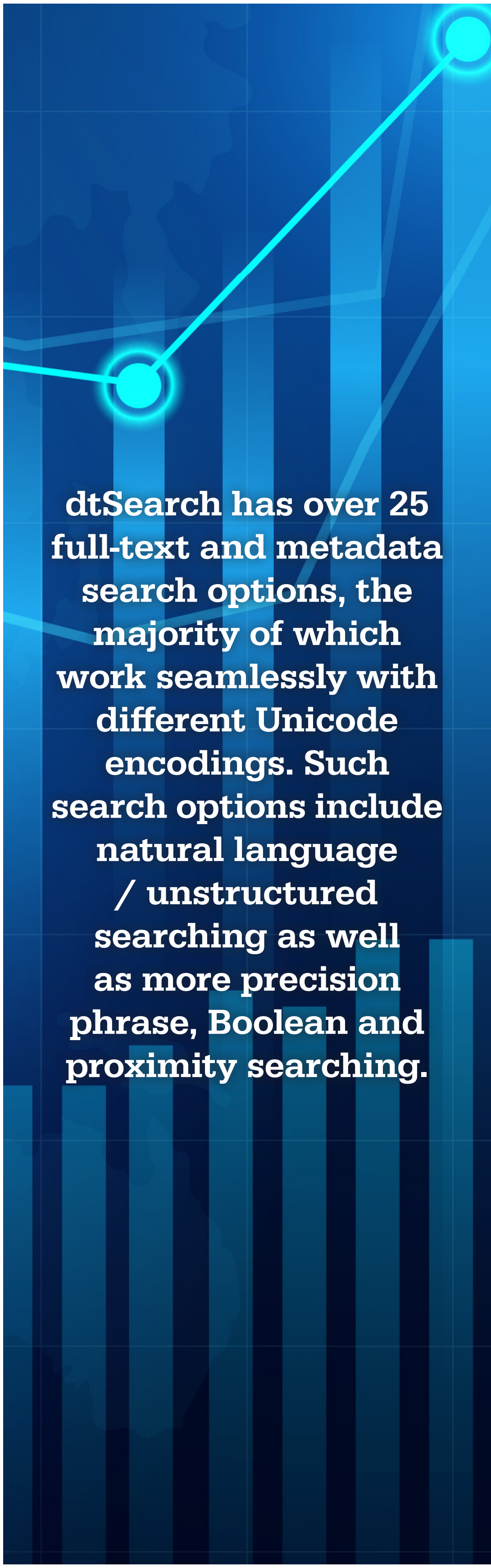
dtSearch has over 25 full-text and metadata search options, the majority of which work seamlessly with different Unicode encodings. Such search options include natural language / unstructured searching as well as more precision phrase, Boolean and proximity searching. Even fuzzy searching which adjusts from 0 to 10 to accommodate various levels of OCR and typographical errors works with multilingual text.

Built-in relevancy-ranking looks for hit density and rarity across indexed data so it too works regardless of language. Alternatively, end-users can apply their own positive or negative variable term weighting encompassing all indexed text or selectively just to text in certain metadata or positionally near the top or bottom of a file. Or an end-user can start out with some type of relevancy-ranking and then instantly re-sort using a different metric like filename, file date or file location for a fresh view on search results. Whatever the sorting, dtSearch can display files with highlighted hits for easy review.

dtSearch.com has fully-functional 30-day evaluation downloads so your organization can get started now on instant concurrent searching through terabytes of its own data. And of course dtSearch is a dynamic product line, not a static one. dtSearch.com has a link to sign up for new product line version notices. Expect a new product line version notice in the next few weeks.



Article contributed
by dtSearch®



dtSearch has over 25 full-text and metadata search options, the majority of which work seamlessly with different Unicode encodings. Such search options include natural language / unstructured searching as well as more precision phrase, Boolean and proximity searching.

Reprinted with permission of USA Business Radio, usdailyradio.com