

# How Exactly Enterprise Search Instantly Finds That Needle in the Haystack

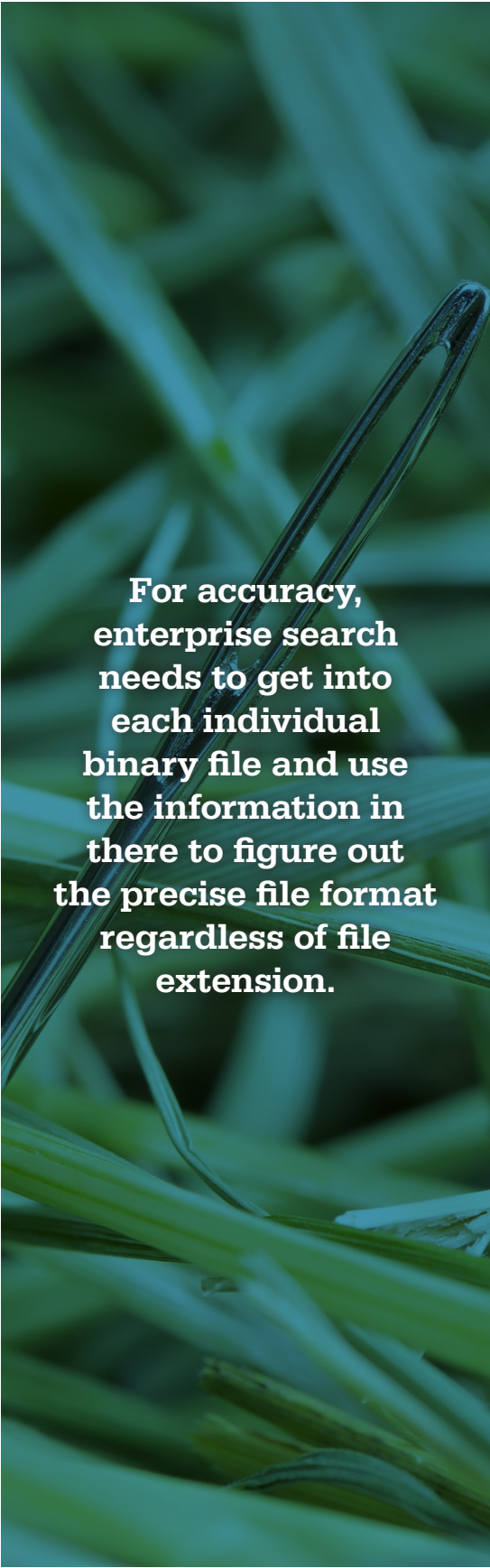
Article contributed  
by [dtSearch®](#)

Enterprise search can instantly find a needle in an unimaginably large data haystack. Even better, with concurrent searching, any number of end-users can immediately locate very different needles across the same large data haystack. But that easy-breezy needle retrieval can happen only after enterprise search undertakes a huge amount of pre-processing work. Today's goal is to show how all that fits together.

Enterprise data typically consists of a random mix of file types—"Office" files, PDFs, compression formats, emails, etc. And it isn't necessarily just standalone files. You can have recursively embedded files. If you looked at such data in its binary format, you'd probably see what looks like a mishmash of binary codes, making it hard to discern any individual words. Enterprise search like [dtSearch®](#) has to parse the content by applying each binary format's file specification.

**File format recognition.** Individual file specifications can be intricate, with some hundreds of pages long. Enterprise search has to accurately identify the exact right file type of each item to match that with the exact right parsing specification. If you were trying to figure out a file type, you'd probably start with the file extension. But that isn't sufficient for enterprise search as it is all too easy to have a PDF saved with a OneNote extension or a PowerPoint saved with an Access database extension. For accuracy, enterprise search needs to get into each individual binary file and use the information in there to figure out the precise file format regardless of file extension.

**Remote and other data.** The same binary format access also enables enterprise search to handle nested components like an email with a ZIP or RAR attachment holding a Word document with an



For accuracy,  
enterprise search  
needs to get into  
each individual  
binary file and use  
the information in  
there to figure out  
the precise file format  
regardless of file  
extension.

Excel spreadsheet embedded inside. dtSearch can automatically index not only local files but also remote files like Office 365, SharePoint and DropBox so long as these appear as part of the Windows file structure.

**Processing the data.** After figuring out the file format and applying the right parsing specification, enterprise search saves the resulting information to an internal summary called an index. From the end-user's perspective, getting enterprise search to build an index is simple. All the end-user needs to do is tell enterprise search the overarching folders, email archives and like to index, and the software will take it from there. From the perspective of enterprise search, however, indexing is a ton of work, with the final index storing each unique word and number across the data and the location of each word and number in the data.

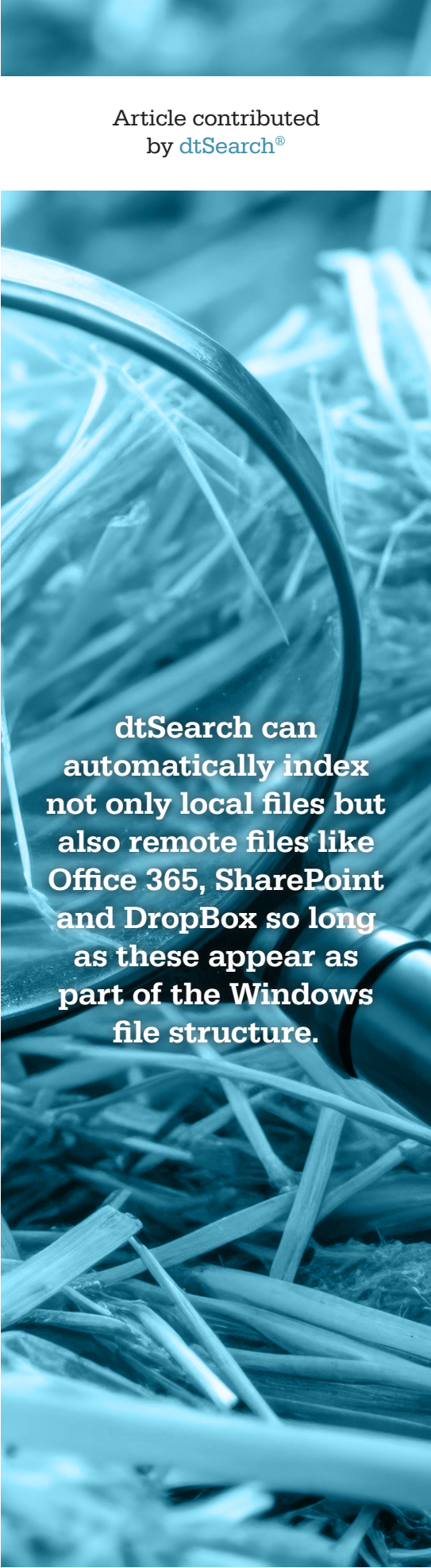
**Depth of data access.** Expect indexing to delve deeply into the data. When you pull up a file in its associated application like a Word document in Microsoft Word, you might not see all the text. For example, it is easy to miss white text against a white background. Certain redaction programs can also show redacted text under a black rectangle, making it look like the text is gone when it is still there, just hidden under the rectangle.

Indexing will find all of that. Indexing can also pick up track changes that have not been fully accepted. And indexing can pick up very obscure metadata that might be hard to spot inside a file's associated application. Indexing can further flag PDFs that are image-only and need OCR from a product like Adobe Acrobat for full-text processing.

**On-premises and cloud operation.** In dtSearch, a single 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 that a search can span. After indexing, concurrent enterprise search can run from a classic Windows network, from a local web server, or from the cloud such as Azure or AWS. While indexing takes a lot of system resources, search threads are resource-light, making them easy to scale.

**Data growth.** Just as haystacks can grow with new hay going in and old hay going out, enterprise data

Article contributed  
by dtSearch®



dtSearch can  
automatically index  
not only local files but  
also remote files like  
Office 365, SharePoint  
and DropBox so long  
as these appear as  
part of the Windows  
file structure.



can grow with file modifications, new files and file deletions. Index updates can accommodate all that without the need to re-index everything “from scratch.” And updating an index can proceed without interfering with concurrent searching, so there is no reason not to keep indexes current.

**Search options.** dtSearch has over 25 different individual and concurrent search options. Less experienced users can enter basic “all words,” “any words,” or “exact phrase” searches. More experienced users can take advantage of intricate Boolean (and/or/not) and proximity search formulations. Search across all data or limit some search elements to certain metadata.

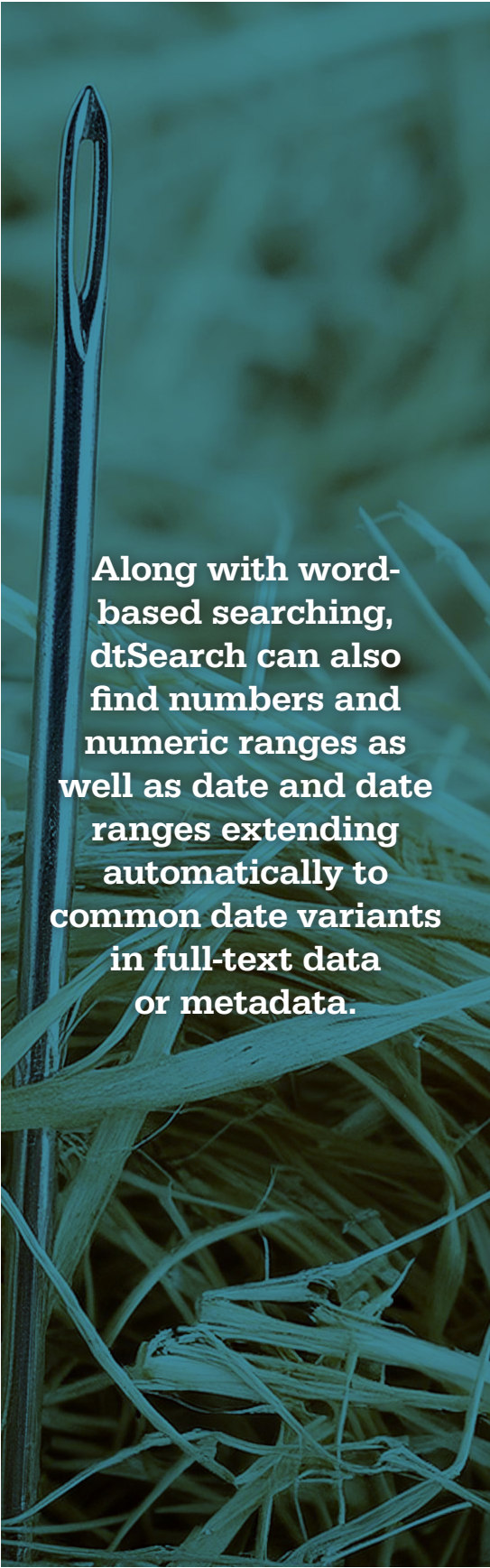
Along with word-based searching, dtSearch can also find numbers and numeric ranges as well as date and date ranges extending automatically to common date variants in full-text data or metadata. Fuzzy searching adjusts from 1 to 10 to sift through typographical and OCR errors. Concept searching extends a search request to synonyms. dtSearch can even identify any credit card numbers in data.

**Multilingual text.** For international languages, current data formats use Unicode to store text. Unicode supports hundreds of international languages that enterprise search like dtSearch can work with. A single email or other file can proceed through any number of European, double-byte Asian and right-to-left Middle Eastern alphabets and languages and Unicode and dtSearch will follow all of that.

**Relevancy ranking and display.** dtSearch further has multiple options for relevancy ranking. For a new view of search results, users can instantly re-sort by a different metric like file date or file location. Whatever the sorting, users can browse full copies of retrieved files with highlighted hits.

**Summing it all up.** Enabling multiple end-users to simultaneously and instantly query terabytes is easy with enterprise search. Just check off the folders and the like to index and the software will take it from there. dtSearch.com has fully-functional 30-day evaluation enterprise search downloads to get you started.

Article contributed  
by [dtSearch®](#)



Along with word-based searching, dtSearch can also find numbers and numeric ranges as well as date and date ranges extending automatically to common date variants in full-text data or metadata.