

# Tame Enterprise Data

The more data an organization has, the harder it can be to sift through in a traditional manner. First, you'd have to pull up potentially millions of files individually, each in its associated application. You'd need to retrieve Word documents in Microsoft Word, Excel spreadsheets in Excel, Access databases in Access, OneNote files in OneNote, PowerPoints in PowerPoint, PDFs in Adobe Acrobat Reader, emails in an email reader, etc.

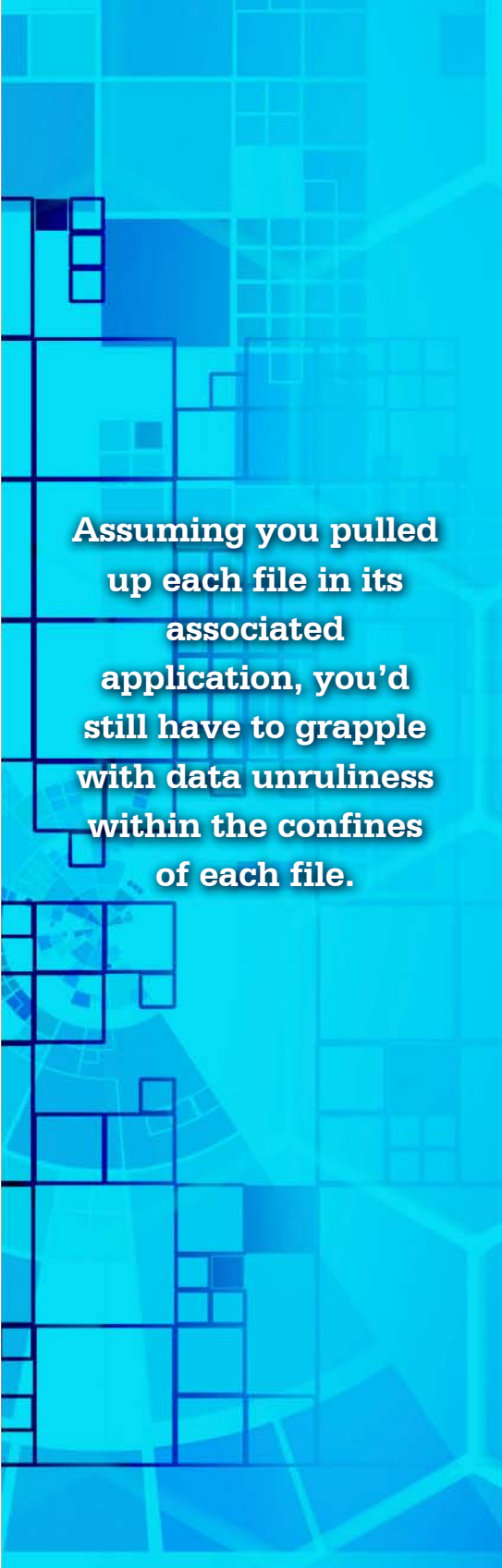
## **What makes the hard way to tame data even**

**harder.** If that wasn't enough, some files might have misleading file extensions such as a PDF ending in a .DOCX extension. And then there are nested formats, like an email with a ZIP or RAR attachment with an Excel spreadsheet with a Word document recursively embedded inside. Assuming you pulled up each file in its associated application, you'd still have to grapple with data unruliness within the confines of each file.

A file can have obscure metadata requiring a huge amount of clicking around in its associated application before you even realize it is there. And then you can have text that is easy to miss because it blends in with a background color in its associated application like black text against a black background or white text against a white background. And it is not just text that someone tried to hide. In some applications, you can have text a person tried to redact but that continued to exist under a black rectangle.

**The easy way to tame enterprise data.** The key to taming large volumes of data is enterprise search like dtSearch. Enterprise search heads directly to files' binary formats, bypassing the need to pull up each file in its associated application. The software further uses the binary formats themselves to determine the file type, so a PDF with a .DOCX extension will not

Article contributed  
by dtSearch®



**Assuming you pulled up each file in its associated application, you'd still have to grapple with data unruliness within the confines of each file.**


affect the correct handling of the file. No matter how difficult it might be to locate certain metadata in an associated application file view, that metadata is right there for enterprise search. And text that blends in with its background color in its associated application is just straight-up text to enterprise search.

**How the easy way works.** Enterprise search instantly searches terabytes after first indexing the data. While indexing sounds like a lot of work, just point to the folders and the like to index, and the software will take it from there. With dtSearch, a single index can hold up to a terabyte of text and there are no limits on the number of terabyte indexes the software can create and instantly search. The indexing process can even work with cloud-based files. For example, Office365 files or SharePoint attachments that appear as part of the Windows folder system operate just like local files for indexing.

**All together now.** While reviewing files in each file's associated application tends to be an individual process, the index structure permits concurrent access through a classic Windows network or through an Internet or Intranet portal. For online search, enterprise search can operate statelessly, with no built-in limits on the number of concurrent search threads. To accommodate changing data, enterprise search can automatically update its indexes at specified intervals without affecting continuing concurrent search.

**Search as you like it.** Enterprise search has over 25 different search features for instantly querying its terabyte indexes. One end-user can enter plain unstructured text to search: *get me everything on the telecommunications satellite launch into low orbit*. Another might enter a more precision-oriented search request linking words or phrases with Boolean (and/or/not) or proximity connectors: *satellite* within

Article contributed  
by dtSearch®



**The indexing  
process can even  
work with  
cloud-based files.**



37 words of *low orbit*, with subject metadata including both *telecommunications* and *launch*, and a full-text date range of 2/15/24 to 3/14/25 in the first few words of the file. Note that this date range would automatically cover not only 5/16/24 but also other common date formats like *May 16, 2024*. Concept searching can extend a search to find *blastoff* for *launch*. Fuzzy searching adjusts from 1 to 10 to sift through typographical or OCR errors like *telecommuMications* for *telecommunications*. The software can even flag any credit card numbers that may be inside of the data.

**Getting to what's relevant.** By default, the software will relevancy-rank retrieved files by search term density and rarity. Take an “any words” search request for *space*, *launch* or *orbit*. If *space* is prevalent across indexed files with *launch* and *orbit* less frequent, then files with *launch* or *orbit* would receive a higher relevancy ranking. And files with the densest mentions of these would receive the highest weighting.

**Even more ranking options.** End-users can add their own variable term weighting, like giving *launch* a positive weight of 5, *orbit* a positive weight of 9 in subject metadata, *space* a positive weight of 7 if it occurs near the top or bottom of a file, and *moon* a negative weight of 3. For a different search results view, end-users can instantly re-sort by some completely unrelated criterion, like file date or file location. Whatever the sorting, enterprise search can show a full copy of retrieved files with highlighted hits for easy browsing.

**Our multilingual world.** Enterprise search works with Unicode which supports literally hundreds of international languages. A single file or email can switch from English, to another European language, to a right-to-left language like Hebrew or Arabic, to double-byte Asian text, and then back to English and enterprise search will follow the entire Unicode progression.

Article contributed  
by dtSearch®

End-users can add their own variable term weighting, like giving *launch* a positive weight of 5, *orbit* a positive weight of 9 in subject metadata, *space* a positive weight of 7 if it occurs near the top or bottom of a file, and *moon* a negative weight of 3.