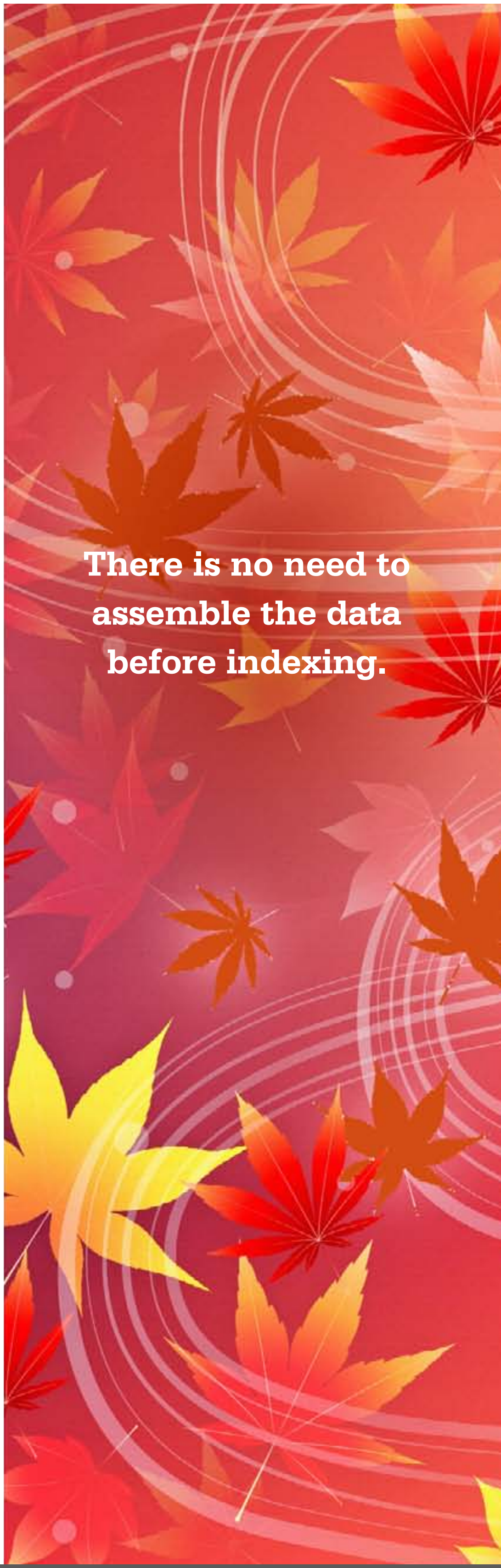
Article contributed
by dtSearch®

Gather Your Leaves, Not Your Content; Your Search Engine Is Fine With Scattered Data

Every fall, you might gather the leaves from your yard into a neat pile for pickup. From an enterprise search engine's perspective, however, you don't need to gather your data. The search engine can work just fine with scattered data.

How so?

A search engine instantly like dtSearch® searches terabytes after first indexing the data. A single index can hold up to a terabyte, with no limit on the number of terabyte indexes that the search engine can create and multiple end-users instantly concurrently search. There is no need to assemble the data before indexing. A single index can cover content from any number of data repositories. Just point to the folders and the like you want the indexer to cover. The indexer will figure out what's in each repository—PDFs; Microsoft Word, Access, Excel, PowerPoint or OneNote files; emails; web-based content, etc.—and take it from there.



There is no need to assemble the data before indexing.

What types of repositories are we talking about?

Some of the repositories can be ordinary files on local or network drives. Other repositories can consist of emails plus attachments on an email server. Still other content can reside on an Internet or Intranet site. dtSearch can even index and search Office 365 and remote SharePoint data so long as these present as folders in the Windows file system.

How does indexed search work with varied index repositories?

Searching across one or more indexes occurs in an integrated fashion, synthesizing search results regardless of where the data comes from. For example, suppose you do a natural language, unstructured search for *autumn leaves*. Vector-space relevancy-ranking will rank the search results across all of the indexed data. If *leaves* are prevalent across the indexed data, but *autumn* pops up in just a handful of places, then *autumn* will get a higher relevancy rank with denser mentions of *autumn* getting an even higher rank. Or you can add custom variable term weighting, giving *autumn* a positive weight of 7 across full-text content, *leaves* a positive weight of 3 across all content, and *green* a negative weight of 6 but only if it appears in certain metadata or positionally near the top or bottom of a file. The search engine will display all retrieved files with highlighted hits. And all this operates regardless of the location of the original data.

But can you take the specific location of the original file into account if you want to?

Yes. You can limit a search to certain folder paths. You can also limit a search to specific file types. The default, however, is integrated searching across everything. And dtSearch contains yet another option to facilitate searching scattered data.

What is that?

You can select caching at the time of indexing to store the full text of files inside the index itself. That way, when dtSearch goes to display a file with highlighted hits, it doesn't matter if the file is remote and slow to load, as it is already effectively right there. Caching makes for very snappy responsiveness for web-based or network-based document display after a search, particularly for large files like huge PDFs in a remote location that might be subject to a spotty connection.

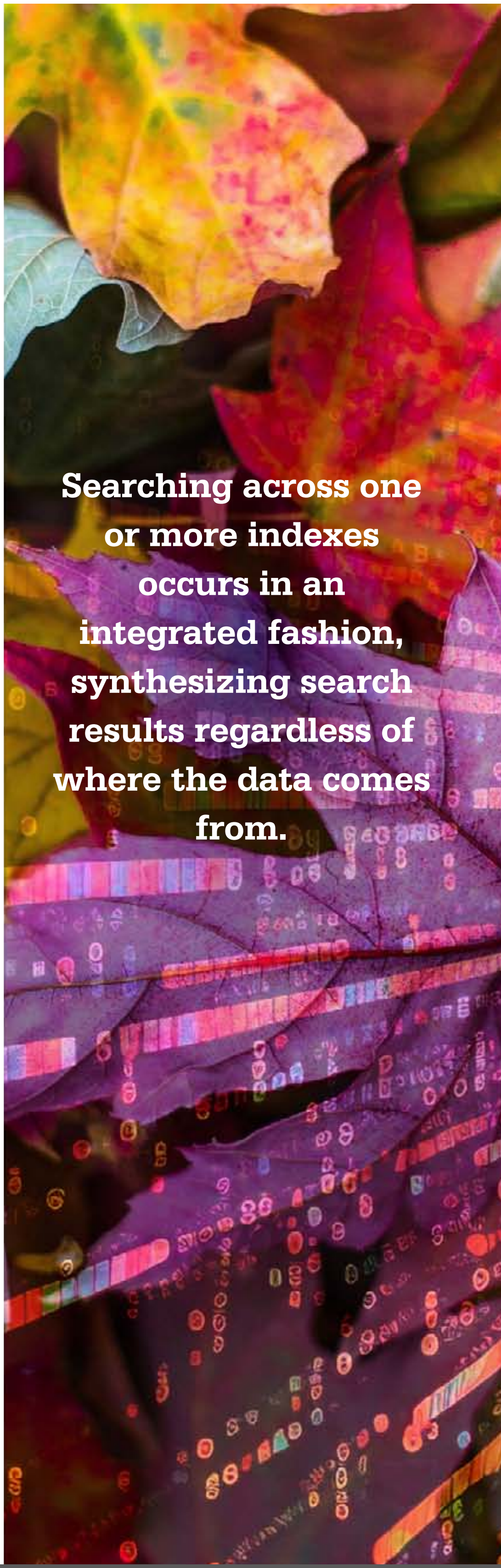
What happens when data changes?

You can set index updates as often as you want. The indexer can just re-index files that have been added, deleted or changed. Critically, the update process can proceed without affecting web-based or network-based concurrent searching. As a side point, while indexing is resource-intensive, searching is not. Online search can run statelessly, making multithreaded concurrent searching easy to scale.

What types of search options are available?

The easiest type of searching is unstructured natural language searching like the *autumn leaves* example earlier. Natural language search greatly benefits from the default vector-space relevancy-ranking bringing the most relevant files from whatever location to the top of the list. Precision searching supports over 25 different hit-highlighted search options. Boolean (and/or/not) and proximity operators can connect words and phrases: *autumn leaves and fall colors and (maple leaves w/27 oak leaves) and not pine needles*. You can also do a directed proximity search, like *maple leaves* but only appearing say within 13 words before *oak leaves*. Or you could require that *maple leaves* appear in certain file or email metadata.

Article contributed
by dtSearch®



Searching across one or more indexes occurs in an integrated fashion, synthesizing search results regardless of where the data comes from.

Reprinted with permission of US Daily Review, usdailyreview.com

Article contributed by dtSearch®

Are there other search options?

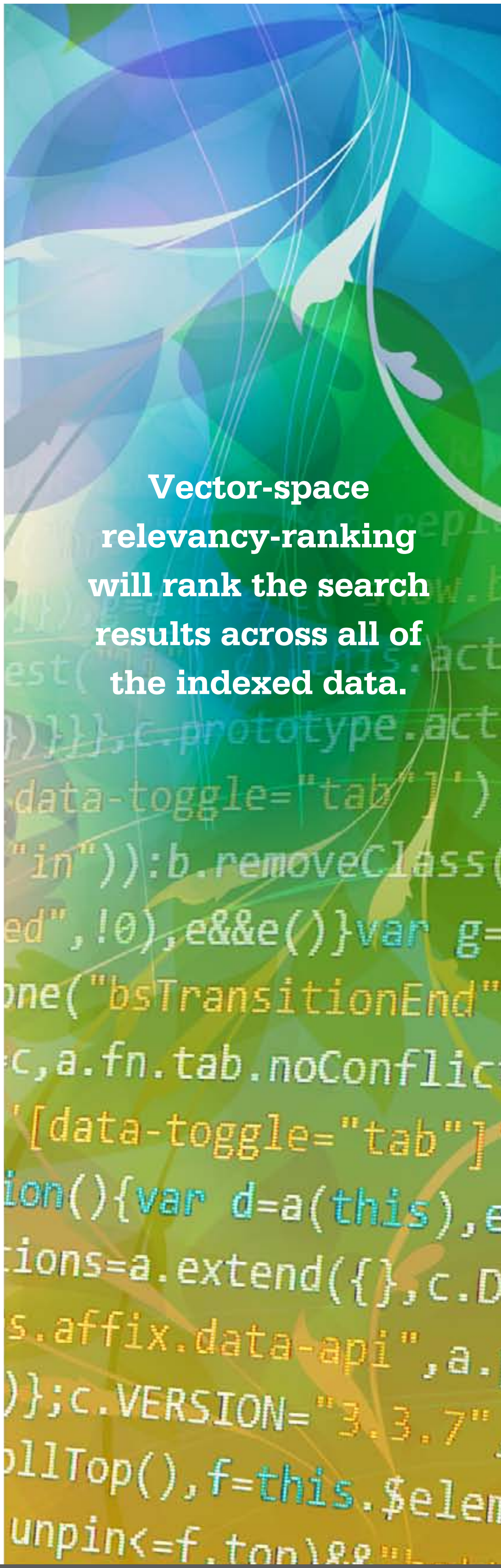
Concept searching would find synonyms like *fall* for *autumn*. Fuzzy searching adjusts from 1 to 10 to look for typographical deviations, which are common in emails and in OCR'ed PDFs. A fuzzy level of 1, for example, would find *autuwn* in a search *autumn*. dtSearch also supports searching for numbers and numeric ranges. And the software also supports date and date range searches even automatically extending to different date formats. For example, you could search for *date(September 1, 2023 to December 15, 2023)* and find *Nov 15, 2023* and *11/30/23*. dtSearch can even let you know if there are any credit card numbers in your data. And for those with multilingual content, Unicode support covers hundreds of international languages including left-to-right text and double-byte Asian text. You can also search for Unicode emojis. There are lots of these for fall.????

Anything else you want to add?

Enterprise search is nothing if not thorough, sifting through all leaves wherever they may fall. For example, dtSearch can still correctly index and search a file even if it contains a mismatched file extension, like a PDF saved with a .DOCX file extension. And dtSearch can process multilevel nested files, like an email with a ZIP or RAR attachment that includes an Excel spreadsheet with a Word document recursively embedded inside. Certain metadata can take extensive clicking around in a file's native application before you even see that it is there, but the search engine can immediately find it. Lastly, text can hide in an Office file if it is the same color as its background, like tan text against a tan background. But it is right there for the search engine.

So don't wait for the next season. Please go to dtSearch.com to try out individual text search or enterprise-wide concurrent search.

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



Vector-space relevancy-ranking will rank the search results across all of the indexed data.

Reprinted with permission of US Daily Review, usdailyreview.com