

Search in a Hybrid Work Environment: How It Works to Increase Productivity and Some Data Cautions

Setting up web-based search to enable secure, instant, concurrent "anywhere" search across terabytes of data is easy. Once set up, in-office and remote personnel can search a shared repository from any browser-ready device or computer, resulting in much less time hunting for critical data. In setting up search for a hybrid work environment, there are many items you need not worry about, although there are several data-related cautions to consider.

The first step in setting up search in a hybrid work environment is to download a search engine. The specifics of this article relate to dtSearch® which has 30-day evaluation versions available for download. But most of the concepts should be generally applicable to comparable search engines.

After downloading, tell the search engine what you want to index. The index is what makes instant concurrent browser-access search possible. Note that there is no "human" effort involved in indexing; all you need to do is point to the directories you want to index. In fact, no need to even tell the indexer what type of data you have. The indexer will automatically recognize popular "Office" formats, web-based file formats, PDFs, email archives, ZIP or RAR archives, etc. If you have an email with a ZIP attachment and in the ZIP attachment is an MS Word file with an embedded MS Access database inside, no problem.

Each index can hold a terabyte of text, and there are no limits on the number of indexes that the application can build and make available for simultaneous search. You can automatically set index updates to account for data additions, deletions or other modifications at any desired interval via the Windows Task Scheduler. Concurrent search can continue even while the index updates itself.

Turning to the web server, the search engine can run as an off-the-shelf application on a Windows IIS server located on-premises or in a cloud environment like Azure or AWS. The off-the-shelf application has customizable HTML5 forms which you can post "as is" or with edits to match your organization's unique vibe. In terms of security, the Windows IIS server's own security settings will apply.

Article contributed
by dtSearch®



Don't worry if your files themselves are not web-ready. The search results display can convert even non-web-ready Microsoft Word, Access, Excel, PowerPoint, OneNote, email, etc. files to HTML for display with highlighted hits.

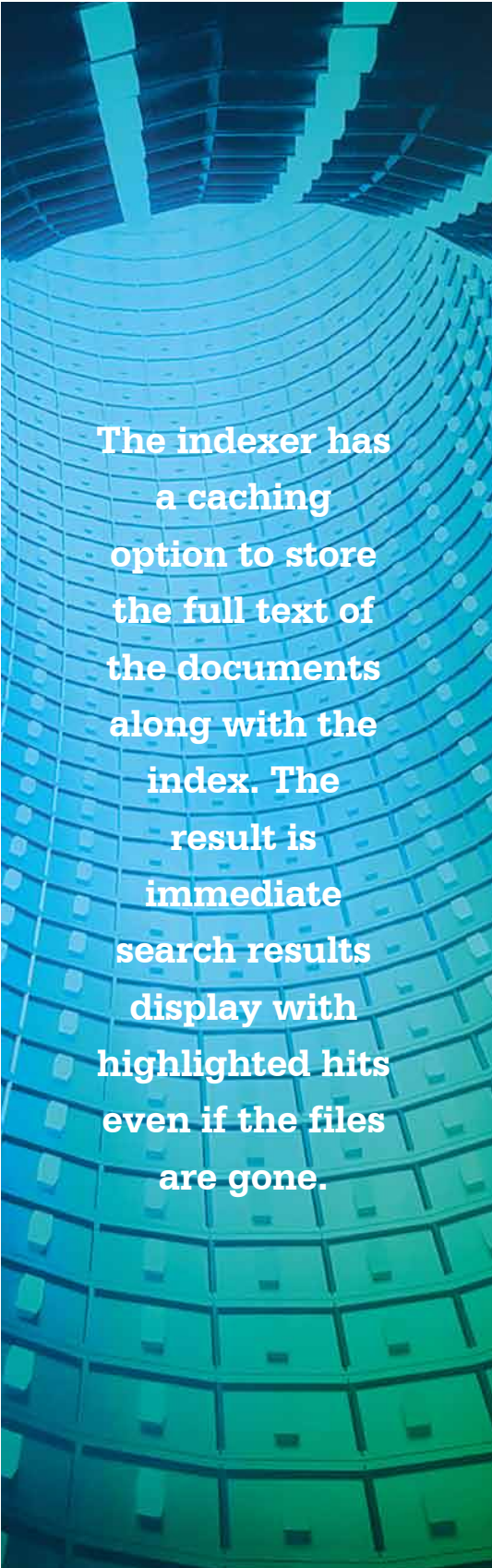
Alternatively, if you are a C++, Java or .NET Core developer, you can run the product line's SDK on any Windows, Linux or macOS server. Developers can also use the APIs to integrate full-text file search with metadata from a backend database like SQL, NoSQL or SharePoint. The SDK includes extensive document classification APIs for granular security settings tailored to each end-user. Any combination of database metadata, document metadata or full-text keywords can serve as the basis for setting the filtering parameters.

The search engine makes available over 25 different search features so end-users can immediately hone in on what they are looking for. Search itself operates in a stateless manner, with no limits on the number of concurrent search threads that can instantly proceed. Search results display can show a complete copy of retrieved files and other data with highlighted hits.

With secure web-based search, workers both in-office and out-of-office can instantly search the shared repository to find what they need. The following are some common concerns that should not derail setting up search for a hybrid work environment.

- ◆ First, don't worry if your files themselves are not web-ready. The search results display can convert even non-web-ready Microsoft Word, Access, Excel, PowerPoint, OneNote, email, etc. files to HTML for display with highlighted hits.
- ◆ Second, don't worry if the original files are not easily accessible via the web server. The indexer has a caching option to store the full text of the documents along with the index. The result is immediate search results display with highlighted hits even if the files are gone.
- ◆ Third, don't worry about "mismatched" file extensions. The search engine doesn't care if your Excel spreadsheets have .DOCX extensions or your PDFs have .ONE extensions. As a technical matter, the search engine's document filters which parse such formats look inside of the binary formats to determine the file type; the file extension is irrelevant here.
- ◆ Fourth, don't worry if the data may have slight typographical or OCR errors. Let's say someone types Case SugarSweetXorbo in an email instead of Case SugarSweetCarbo. The search engine's fuzzy searching can still find that misspelling.
- ◆ Fifth, don't worry if your files are in multiple languages. Search works with any Unicode-based text. If files or portions of files contain anything from English to Spanish to Chinese to Russian to Arabic to Swahili, that text will all be fully searchable.
- ◆ And finally, don't worry about a search engine like dtSearch getting a hold of your data. These applications do not send copies of your files, the indexes, search request information, etc. back to dtSearch servers. That is not how these applications work!

Article contributed
by [dtSearch®](#)



The indexer has a caching option to store the full text of the documents along with the index. The result is immediate search results display with highlighted hits even if the files are gone.

While the above are items not to worry about, there are some cautions to be aware of with regard to how a search engine "sees" data.

- ◆ **Caution #1:** Obscure Metadata. Sometimes Office files, PDFs and the like can have obscure metadata that is hard to see inside the file's native application. However, all metadata will be apparent to a search engine as it approaches files in their binary format, not their application view.
- ◆ **Caution #2:** Tracked Changes. If your files have tracked changes, the search environment can make these visible. For example, let's say you have a Microsoft Word document with changes tracked but not "accepted," these changes are still part of the binary format document and can appear in the online file display.
- ◆ **Caution #3:** "Invisible" Text. White on white text or black on black text is just like any other text to a search engine. Even if such text is hard to spot in a file's associated application, that text is fully available in a file's binary format and hence to a search engine.
- ◆ **Caution #4:** "Image-Only" PDFs. All of the above "cautions" relate to data that you might not expect a search engine to find but that it can find anyway. The reverse can also be true, particularly with regard to a certain type of PDFs. Normally, PDFs combine text with images. If you can highlight text in a PDF and copy it into another application, you have a normal PDF.

However, some PDFs may look normal but consist of an image only. If you try to copy and paste a selection of text from an "image only" PDF, that process will simply not work. While there is no external indicator that a particular file is an "image only" PDF, the indexer can flag such files for you during the indexing process. That way, you can run them through an OCR program like Adobe Acrobat to turn them into text-based PDFs that the search engine can fully search.

- ◆ **Caution #5:** Personal Information. A search engine can identify personal information like credit card numbers in files. In fact, one search option can specifically locate and flag valid credit card numbers anywhere in the indexed data. Accordingly, you may want to do a quick search pass for these prior to making search available.

With web-server-based search, your hybrid workforce no longer needs to spend inordinate amounts of time rummaging around for critical data. Instead, they can instantly find what they need to move ahead with their larger projects.

Article contributed
by [dtSearch®](#)

**Don't worry
about a search
engine like
dtSearch
getting a hold
of your data.
These
applications do
not send copies
of your files,
the indexes,
search request
information,
etc. back to
dtSearch
servers. That is
not how these
applications
work!**