6 Ways You Can't Hide Text in Files, Emails and Other Data From a Search Engine

Article contributed by dtSearch®

Before getting into the methods that *won't work* for hiding text in files, emails and other data, a quick overview of how a search engine operates is in order. A search engine like dtSearch® instantly searches terabytes after indexing. An index is simply an internal tool holding each unique word and number across all indexed data, and the location or locations of each in the data. Indexing is no effort for the end-user; just point to the folders and other data sources to index and the search engine does everything else.

A single index can hold up to a terabyte of text and there are no limits on the number of terabyte indexes that the search engine can cover in a search request. In a shared work environment like a network or an online environment, multiple people can instantly concurrently search across the same terabyte indexes, each seeing search results with highlighted hits. Updating indexes to reflect data modifications can proceed without affecting instant concurrent searching.

(1) The first way you can't hide text from a search engine is saving it in a file with a misleading file extension, like saving a Microsoft Excel spreadsheet with a PDF extension or a OneNote file with a .docx file extension. It is true that a search engine needs to correctly identify the file format of each item it is indexing to correctly parse that item. Excel has a very different parsing specification from PDF, and OneNote has a very different specification from Word. But a search engine can figure out the file format by looking inside the binary version of each item. The filename extension is completely superfluous for this purpose.

(2) The second way you can't hide text from a search engine is through a text color that blends in with the background, like white text against a white background or black text against a black background. When you are looking at a file in its native application, that technique may very well work to obscure the text. However, when a search engine parses the file, text that in a file's native application blends in with the background is just as readily apparent as any other text.

(3) The third way you can't hide text is to put the text in obscure metadata. Almost all files have metadata. There is obvious metadata like a filename or an email subject line. And then there is less conspicuous metadata that may take a huge amount of clicking around in a file's native application to see. But regardless of how obscure metadata may be in a file's native application, inside the binary format that a search engine works with, the metadata is just as accessible as any other text.

But regardless of how obscure metadata may be in a file's native application, inside the binary format that a search engine works with, the metadata is just as accessible as any other text.

(4) The fourth way you can't hide text is to save it as an image-only PDF. Let me first explain what an image-only PDF is. When you retrieve a standard PDF file in a viewer like Adobe Reader, you can of course see the text. And you can also copy and paste selections from the text into a different application like a Word file. But sometimes if you try copying and pasting text from a PDF, you end up with no useful output because you are working with a picture only. The name for that is an image-only PDF.

While a search engine can't find text inside an image-only PDF, the search engine can flag such PDFs during the indexing process to indicate that these are image-only PDFs. Once a search engine flags these files, an OCR application like Adobe Acrobat can turn them into full-text searchable PDFs. At that point, the text is readily available to a search engine.

(5) The fifth way you can't hide text is through minor misspellings like typing Project 2024 as Promect 2024 in an email. A search engine's adjustable fuzzy searching can sift through that. A low level of fuzzy searching would find Promect 2024 in a search for Project 2024. A higher level of fuzzy searching would find not only Promect 2024 but also Promext 2024 in a search for Project 2024. In addition to sifting through misspellings, fuzzy searching is also helpful for sifting through potential OCR errors in PDF files. The bottom line is don't expect minor typos to thwart a search engine.

(6) The sixth way you can't hide text from a search engine is to set up some type of complex container structure, like inserting an Access database inside a Word file. A search engine can go deep in parsing nested data structures. If you have an email with a ZIP or RAR attachment with a Word document embedding an Access database, the search engine can parse the entire structure right down to the inner layer.

Now that we have the ways you can't hide text from a search engine, a quick bonus note on what you can find with a search engine. Indexed search includes over 25 different search options. These cover not only word-based search requests but also number-based queries and even mixed number and word expressions. For example, a search engine can find all dates in a date range even if the dates appear in different formats like March 15, 2020 and 5/17/22. A search engine can also identify any valid credit card numbers that may appear in files, embedded files, metadata and the like.

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 with lots of data to search can download a fully-functional 30-day evaluation copy from dtSearch.com

Article contributed by dtSearch®