Hidden Text – What Lies Beneath

Article contributed by dtSearch®

This article addresses hidden text from the perspective of a search engine like dtSearch. When you view a file or email in the program associated with that file, like when you view a spreadsheet in Microsoft Excel, or read an email in Outlook, or look at a PDF file in Adobe Reader, you are viewing that file through the lens of its associated application. But that is not how a search engine like dtSearch would "read" that same document.

Search engines approach files in their raw binary format. That binary format view can look quite different from the associated application view. For example, if you were to look at the average Microsoft Word file in its binary format, it would be tricky to pick out any text at all visually. By contrast, retrieving that file in Microsoft Word would make reading the document quite straightforward. But while the associated application view may appear more comprehensive to the unaided eye, as the search engine parses a file in binary format, the search engine can also "see" text that might escape scrutiny in an ordinary Microsoft Excel, Word, Access, Outlook, PowerPoint, OneNote, PDF, Outlook, etc. file view.

For example, Microsoft Word might show a document in its final form, but the document itself might still have tracked changes inside that aren't immediately visible unless you think to check a "track changes" view. A search engine like dtSearch, however, would immediately see all of those track changes when it reads the document in binary format, even if they would not be displayed by default in the Microsoft Word file view. That same document might have document properties metadata that is not immediately available in the ordinary file view, but would be readily searchable by a search engine. There could also be "white on white" text in a document. That text would be completely invisible with an ordinary file view, but would be plain as day for a search engine.

Most modern document formats can also have embedded objects. Of course, an email file can have a ZIP attachment with a PDF inside. And you would realize that if you pulled up the email in Outlook. But there are also less visible embedded file manifestations. For example, you could have a Microsoft Word document with an Excel spreadsheet embedded in it. When you view the file in Word, you might see just a few lines of that spreadsheet. But the whole spreadsheet is still there as an embedded object, and the whole thing would be visible in binary format. In fact, dtSearch can even navigate among multiple levels of embedded objects which might be totally obscure in an ordinary file view. For example, you could have email with a ZIP attachment containing a PowerPoint with an embedded Excel spreadsheet.

In closing, enterprises with extremely large data sets like government agencies and 4 out of 5 of the Fortune 500's largest Aerospace and Defense companies use dtSearch enterprise and developer products to instantly search terabytes of Microsoft Office documents, email files, databases and other online data. But even if you just want to search the data on your own PC, you can download a fully-functional 30-day evaluation version of dtSearch Desktop anytime at dtsearch.com. And please check out the Features Map page at dtSearch.com for more "deep dive" search tips, like searching for any credit card in a collection of data, generating and searching for unique file hash values, etc.

Reprinted with permission of The Times USA, thetimesusa.com | Original article link (4/29/2020)