When Text Seems To Do a Disappearing Act (But Doesn't)

Article contributed by dtSearch®

If you look at some recent tech forum discussions, you'll see that when some people thought that they were deleting their posts from a certain social media site, they were making them disappear from the user interface, without actually deleting them from the site. Leaving aside social media, your own content may seem to disappear but actually remain fully present from the perspective of a search engine like dtSearch (dtSearch.com).

The first example is black on black text, red on red text, white on white text, etc. This type of text can hide in a spreadsheet, database file, presentation file, word processing document, etc. when you view the file in its associated application. By associated application, I mean Microsoft Word for a Word document, Microsoft Access for an Access database, a PDF viewer like Adobe Reader for PDF files and the like.

A search engine, however, doesn't work with documents, emails, etc. when they are pulled up inside their associated application. Rather, a search engine approaches files in their binary format when they are just sitting there on a PC, network or web server. In binary format, black on black or red on red text looks the same as any other text and is fully readable by a search engine.

As a related example, certain redaction or editing programs can make text seemingly disappear from your screen as part of the redaction or editing process. But the text is still there. You can frequently copy and paste the seemingly invisible text. And as long as it remains present in a file's binary format, the invisible text will be readily available to a search engine.

For yet another example, certain metadata can be very hard to find in an associated application. You may need to click and click in just the right way before running across it. But when a search engine views a file in its binary format, the metadata along with the full text will be in plain view.

Multilayer embedded files can also be a source of "disappearing" text. For example, you can have a Microsoft Excel document which embeds a Microsoft Word document. Inside the Excel file, you may see only a fraction of the Word document. But the file's binary format can reveal the entire embedded structure to a search engine looking for embedded files. So you could have a ZIP or RAR attachment to an email with a Word document embedded inside the Excel file and everything will be full-text searchable.

A search engine, however, doesn't work with documents, emails, etc. when they are pulled up inside their associated application. Rather, a search engine approaches files in their binary format when they are just sitting there on a PC, network or web server.

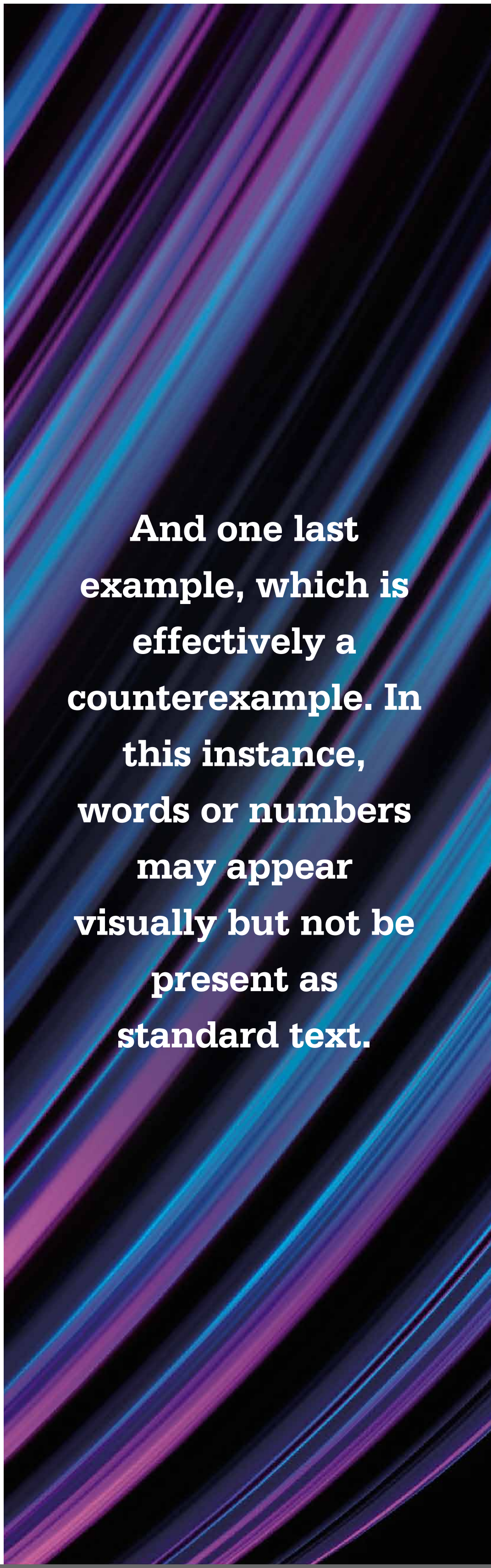
Reprinted with permission of USA Daily Times, usadailytimes.com

Article contributed
by dtSearch®

Another item that may affect file visibility is saving a file with a different extension than its associated application would expect. For example, you can save a PDF file with a Microsoft Word extension or an Access database with an Excel extension. But while this can sometimes trick the associated application, a search engine like dtSearch will not rely on the file extension to figure out the type of document and the parsing specification to apply. Rather, the search engine will look inside the actual binary format to figure out what type of file it is. In sum, you can't hide a file from a search engine just by giving it a non-conforming extension.

And one last example, which is effectively a counterexample. In this instance, words or numbers may appear visually but not be present as standard text. The most common example of this is "image only" PDFs. There are two basic types of PDFs. The first is the normal type of PDF you would typically obtain if you print a document to PDF. The second looks like a normal PDF, but when you try to copy and paste some text from it, the text does not respond. In that case, you may have an image-only PDF that requires OCR to turn it into machine-readable Unicode. Such "image only" PDFs may be hard to spot in a collection of general PDFs. dtSearch has an option to flag image-only PDFs as you index them, and then you can send them through an OCR program like Adobe Acrobat to turn them into regular PDFs.

So how do you run a search engine like dtSearch on your own data? dtSearch has enterprise and developer products that run "on premises" or on cloud platforms to instantly search (with 25+ search options) terabytes of "Office" files, PDFs, emails along with nested attachments, databases and online data. Because dtSearch instantly search terabytes, 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



And one last example, which is effectively a counterexample. In this instance, words or numbers may appear visually but not be present as standard text.

Reprinted with permission of USA Daily Times, usadailytimes.com