

Hidden Text – What Lies Beneath – PDF Edition

This article delves into the hidden text world of PDFs.

This piece supplements a previous article

<https://www.thetimesusa.com/what-lies-beneath/>

looking at hidden text in Microsoft Office files, emails and email attachments.

Both this article and the previous article examine hidden text from the perspective of a search engine, specifically dtSearch. Large organizations like government agencies and 4 out of 5 of the Fortune 500's largest Aerospace and Defense companies rely on dtSearch enterprise and developer products to instantly search terabytes of Office files, emails, databases and web data. But even if you just want to search across your own PC, you can download a fully-functional 30-day evaluation version of dtSearch Desktop anytime at dtSearch.com.

PDF is actually a printer format. When you look at a PDF document inside a viewer like Adobe Reader, you are typically looking at the document as it would print. However, a search engine like dtSearch would review a PDF file not in associated application like Adobe Reader, but in its raw binary format. That binary format view can look quite different from the associated application view. In fact, if you were looking at a PDF in binary format, it would be hard to visually distinguish any words at all in the text.

By the same token, once a search engine parses a PDF in binary format, it can also see text that might escape scrutiny in a “normal” file view. Since PDF is a printer format, there could be text outside of the page boundary that might be hidden in a “normal” associated application view. That extra text, however, would be readily apparent to a search engine like dtSearch.

PDFs can also include metadata that would be not immediately available in an associated application view but that would be readily available to a search engine. And PDFs can have embedded objects such as embedded MS Office files that could be easy to overlook in an associated application view but that would be “plain as day” to a search engine like dtSearch.

And finally, PDFs can have “white on white” or “black on black” text. Such text would not be readily apparent in an associated application view, but would be completely apparent in binary format. In a recent high-profile criminal proceeding, certain PDF text was visually redacted with black rectangles prior to public release. But while visually blacked out, the underlying text itself (unbeknownst to those who publicly released it) was still fully accessible.

Article contributed
by [dtSearch®](#)



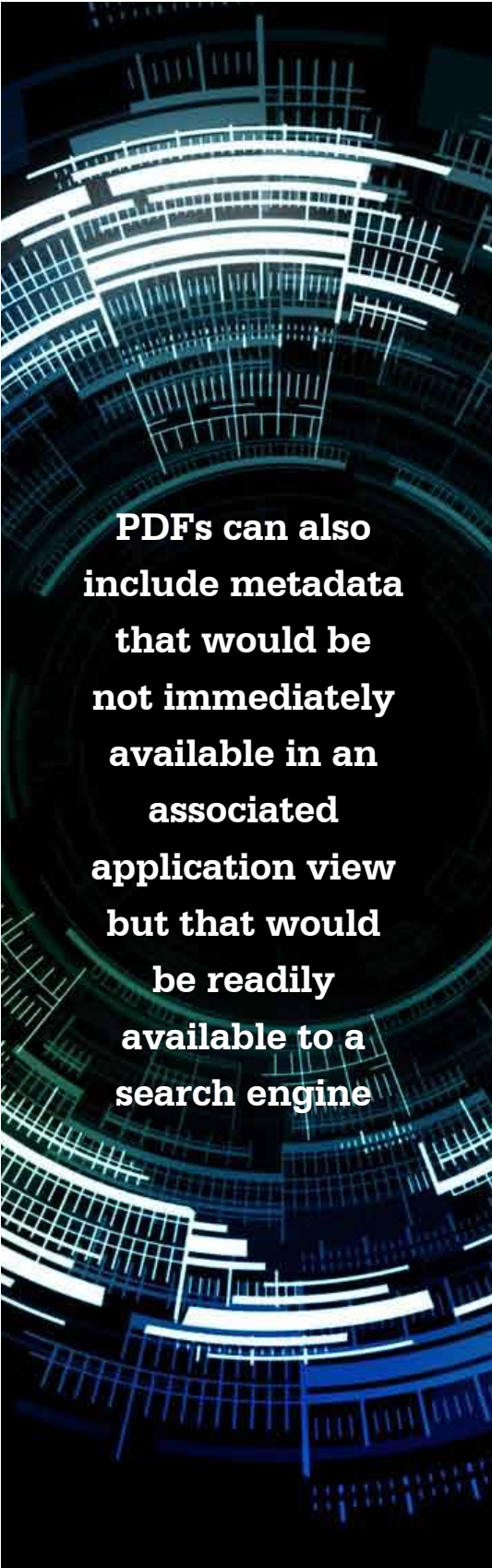
**This article
delves into the
hidden text
world of PDFs**

There is also a question of how a search engine recognizes files. What if someone gives a PDF file a Microsoft Office extension, like .docx instead of .pdf? While that might be confusing if you were looking at the file in a directory, a search engine like dtSearch would look at the binary file heading to determine the file type, not the filename extension. So that way, even if a PDF has a .docx extensions, dtSearch will still handle that PDF correctly.

Finally, coming 25 years after Adobe created the original PDF document format, Version 2.0 is a major new release of the PDF file type, and these files are just starting to get out there. dtSearch has also released a new version to make sure that PDF 2.0 files would be separately recognized, and treated accordingly.

From dtSearch.com, you can immediately download and try a fully-functional 30-day evaluation version to instantly search terabytes of your own data. And when you do try the software, check out the forensics-oriented section of the Features Map at dtSearch.com for more “deep dive” search tips on PDFs and other data.

Article contributed
by [dtSearch®](#)



PDFs can also include metadata that would be not immediately available in an associated application view but that would be readily available to a search engine