Working Remotely or “In Office” — What You Need to Know About the Duality of Files

Article contributed by dtSearch®

Think about your files. You probably picture the Microsoft Word document you were just editing as it appears in your Microsoft Word application. Or you may think about a PDF as it appears in a viewer like Adobe Reader, a presentation in PowerPoint, a spreadsheet in Excel, or as an email as it appears in Outlook.

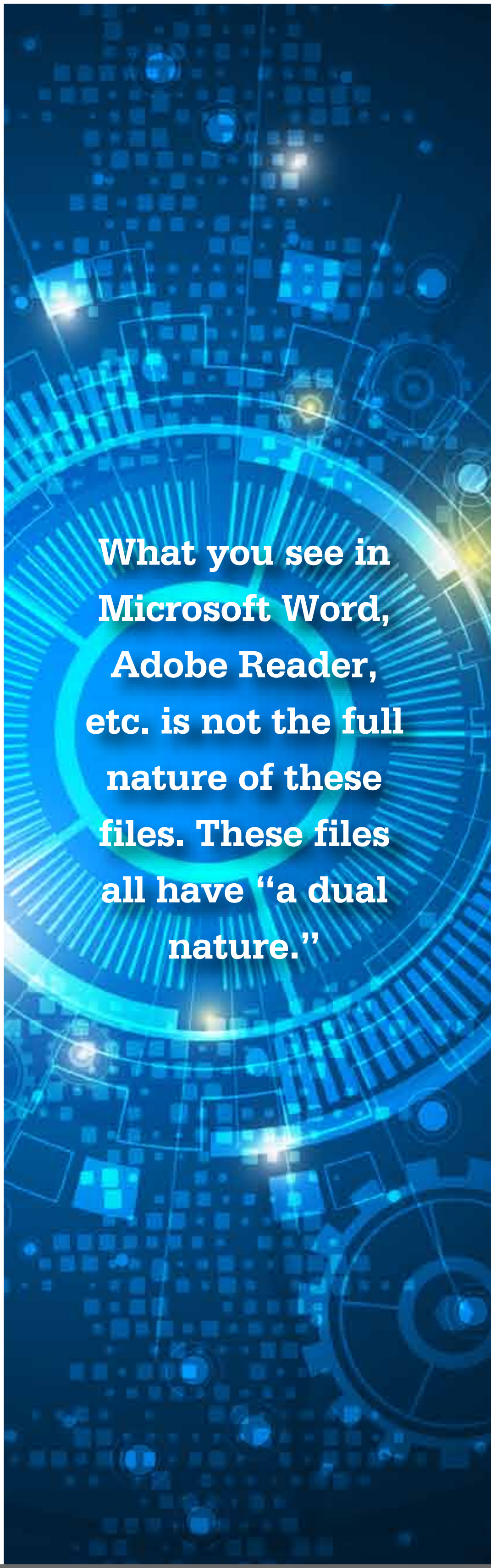
What you see in Microsoft Word, Adobe Reader, etc. is not the full nature of these files. These files all have “a dual nature.” In fact, these native applications views are more like the tip of the iceberg when it comes to a file’s alternate binary format existence. A file’s binary format is the relevant mode when it is just sitting on your hard drive, network or online portal.

The binary format typically looks nothing like what you see inside an associated application. For example, inside of Microsoft Word, a document is typically easy to read in terms of complete sentences and paragraphs. In binary format, it may be hard to pick out even a single word. You may just see random letters floating in a sea of gibberish-looking codes.

While a binary format may look like a sea of gibberish to the naked eye, to a search engine, a binary format is more like a crystal ball. Inside the crystal ball is not just what you can see in an associated application view, but so much more.

How does a search engine parse a binary format? The first step to parse a binary format is to identify the correct binary format specification to apply. The binary specification for “interpreting” a OneNote document is very different from the binary specification for “interpreting” a PDF. The PDF is very different from the binary specification for “interpreting” an email. And these specifications can be beyond complex — approaching hundreds of pages of technical documentation.

One way to identify the correct binary specification to apply would be to look at the filename extension. If a filename ends in .DOCX the Microsoft Word specification would apply and if it ends in .PDF — the PDF file specification would apply. But what if someone saves their PDF files with a .DOCX filename extension and their OneNote files with a .PDF filename extension?



What you see in Microsoft Word, Adobe Reader, etc. is not the full nature of these files. These files all have “a dual nature.”

Reprinted with permission of ReadWrite, readwrite.com | Original article link (7/2/2020)

The more accurate way to identify the relevant specification to apply to a binary file is to look inside the binary file itself. Looking inside the binary file itself — you can determine the format type, rather than looking at the filename extension. With the correct format type — no matter what extension someone tacks onto a Microsoft Word document — the correct parsing mechanism can still apply.

Practical Tips

1) When you use a search engine like dtSearch: the filename extension does not affect the ability to find a file. A lot of times, you can have metadata relatively hidden in an associated application view. This means that the data will not pop up by default; you'd have to do some considerable clicking around to find the information. However, to a search engine, all text and data are on the same footing.

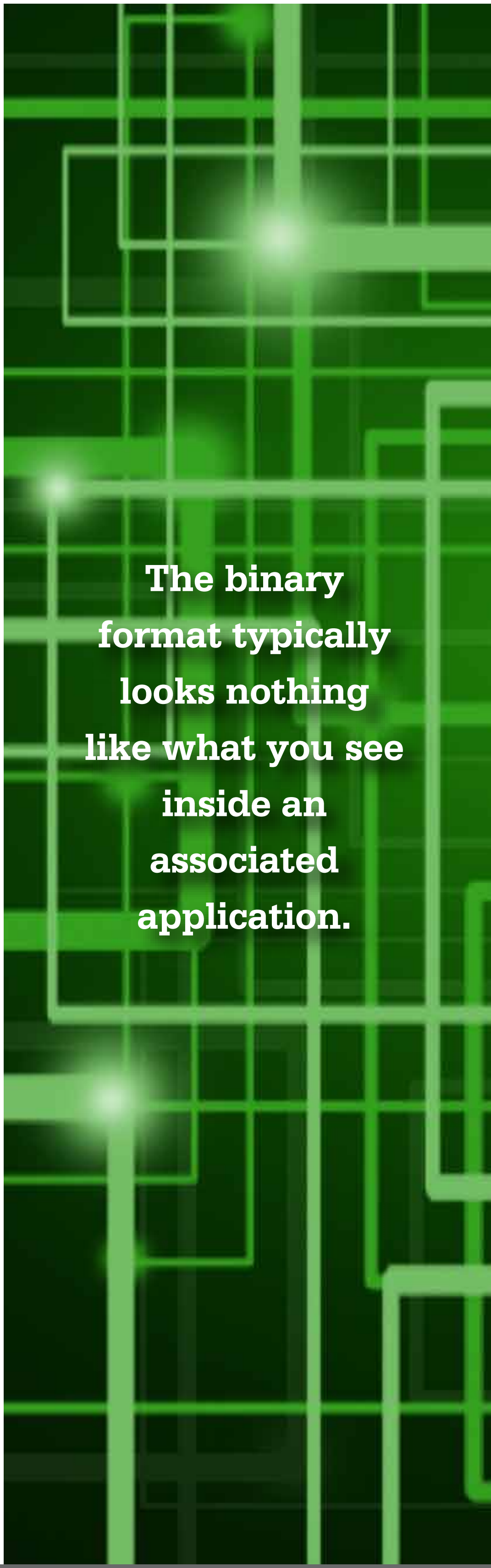
2) The second practical tip relating to the dual nature of files and a search engine then is that there is no metadata too obscure for the search engine to easily find.

3) The third practical tip relates to "black on black" or "white on white" or "red on red" text. These types of text will typically be completely invisible in an associated application view. However, it is just as apparent as any other text to a search engine. Therefore, the third tip relating to the dual nature of files and a search engine is that the visual contrast between words and background inside of an application does not matter to a search engine.

4) The last suggestion here is "file specific," and relates to a subset of files that I will call "image only" PDFs." Sometimes you'll run across a PDF where you try to cut and paste the text from it, but you can't, because it is a picture of text only, and does not actually include a digital version of the text. By the same token, as an image only, a search engine is not going to see the text there either — the search engine only "sees" the image (along with any metadata).

Keep in mind that a search engine can identify "image only" PDFs specifically. The search engine then flags the image to indicate that the file requires optical character recognition or (OCR). Remember that OCR is a separate application — such as an app like Adobe Acrobat can perform. Once optical character recognition (OCR) happens — you can then cut and paste the text at will and the text will be "all there" for a search engine to find.

Article contributed
by dtSearch®



The binary format typically looks nothing like what you see inside an associated application.

Reprinted with permission of ReadWrite, readwrite.com | Original article link (7/2/2020)