

Digging into Data – Unburied Treasure

Previously, I addressed how a search index like one generated from the program dtSearch® is like a data treasure map, supporting both individual and enterprise-wide concurrent searching with instant hit-highlighted search results (Article link). That way, multiple individuals can use the same treasure map at the same time and each arrive at whatever unique X marks the spot that individual is looking for.

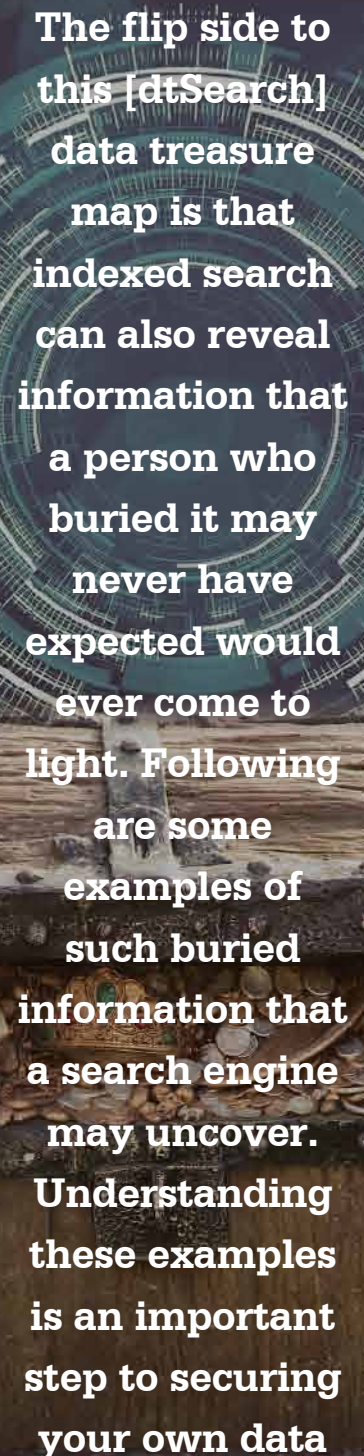
But the flip side to this data treasure map is that indexed search can also reveal information that a person who buried it may never have expected would ever come to light. Following are some examples of such buried information that a search engine may uncover. Understanding these examples is an important step to securing your own data.

(1) An end-user might save a file with a filename that doesn't match the file type. For example, an end-user might mislabel a Microsoft Excel file with a .PDF extension, or mislabel an email archive with a .DOCX extension. However, in parsing a binary format, a search engine like dtSearch will figure out the correct file format specification to apply by looking inside the binary file itself, rather than simply looking at the filename extension. So saving a file with a mismatched filename extension will not have any effect on the ability to uncover text in that file.

(2) An end-user might also nest a file inside another file. An example of that would be an email with a ZIP or RAR attachment and a PDF and Microsoft Word document inside and an Access database fully embedded in the Microsoft Word file. However, a search engine like dtSearch will work its way recursively through such nested attachments, making the inner file contents just as visible as the cover email.

(3) Many “Office”-type applications let you insert obscure metadata where that metadata won't appear by default when you look at a file in that application. In fact, you may have to click around extensively to find the metadata such that the likelihood of a casual viewer of the file in its native application finding the metadata is very low. However, to a search engine, that metadata is as readily accessible as any other text.

Article contributed
by [dtSearch®](#)



The flip side to this [dtSearch] data treasure map is that indexed search can also reveal information that a person who buried it may never have expected would ever come to light. Following are some examples of such buried information that a search engine may uncover. Understanding these examples is an important step to securing your own data

(4) In many applications, an end-user can also hide text by making it the same color as the background underneath the text. As a result, there can be white on white text or black on black text which inside the application itself may not be readily visible. However, to a search engine that text is easily apparent regardless of whether the text color in the application view matches the background color.

(5) An end-user can also slightly misspell a word. Typos in emails are everywhere – at least in my emails. But with fuzzy searching on, dtSearch will automatically sift through those slight typographical errors to find ProJectX when ProJectX is mistyped as ProPectX.

(6) Still another example is credit card numbers that may be buried in files. There is a feature inside of dtSearch that can check if digits presented together may be a valid credit card, even if there is no MasterCard, Visa or American Express insignia, for example.

(7) The final example relates to “image only” PDFs. Have you ever run across a PDF where you try to cut and paste text from it, but you can’t, because it is an image only? A search engine can’t find the text on such a PDF either, because it is an image only. However, a search engine like dtSearch can flag this type of file when it does its indexing, and let you know that you need to run it through an OCR program like Adobe Acrobat. At that point, the text of this file will be buried no more.

dtSearch enterprise and developer products instantly search terabytes of “Office” files, PDFs, emails along with attachments, databases and web-based data. The products can run “on premises” or on online platforms like Azure and AWS. Because dtSearch can instantly search terabytes of data, many customers are large enterprises like Fortune 100 companies and federal, state and international government agencies.

However, in addition to enterprise-level search, dtSearch also lets **you** search **your own** documents, emails and the like. Please go to [dtSearch.com](https://dtsearch.com), and download a fully-functional 30-day evaluation version to instantly search terabytes of your own data.

Article contributed
by [dtSearch](https://dtsearch.com)[®]

dtSearch
enterprise and
developer
products
instantly search
terabytes of
“Office” files,
PDFs, emails
along with
attachments,
databases and
web-based data