

The Data You Normally See Is Just the Tip of the Iceberg

When you look at a file in its associated application, like pulling up an email in Outlook or a PDF in Adobe Acrobat Reader, you think that you are seeing the whole thing. But what you are seeing is more like the tip of the iceberg. Today I want to delve into the larger submerged iceberg, as that is what a search engine like dtSearch sees.

A search engine like dtSearch® has to pre-process millions and sometimes billions of files to enable instant search either on an individual basis or on a multithreaded concurrent search basis across terabytes of data. The name for that pre-processing is indexing. Getting a search engine to index data is really easy; just point to the folders and the like to index, and the search engine will take it from there.

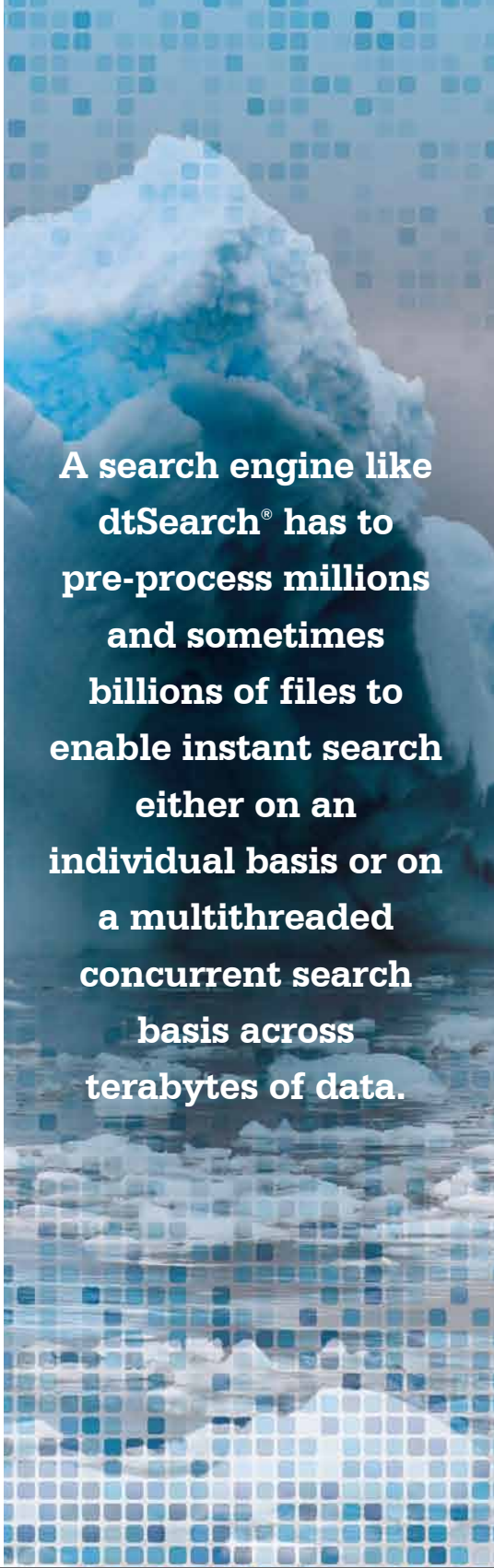
In indexing millions or billions of files, however, the search engine can't retrieve each file in its associated application. That would be way too slow. Instead, a search engine has to approach data in its binary format, as it sits there in the file system. You know when you retrieve a Word document in Microsoft Word, you expect it to be easily readable? That's the whole point of a word processor. But if you looked at a Word file in binary format as a search engine sees it, you'd have trouble discerning any sentences at all amid the sea of binary codes. Those binary codes represent Microsoft's internal instructions to Word for displaying the file.

To recognize all text and metadata in binary format, the search engine has to apply the right parsing specification. That is not easy to do, as parsing specifications can be hundreds of pages long. But the binary format also provides a much greater window into the file than an end-user would normally see, the equivalent of the larger iceberg under the waves.

Tip if the Iceberg, Example #1. Sometimes a file can hide inside the file system with a mismatched extension, like a Word document renamed filename.pdf or filename.dll. If you saw that in the file system, you probably wouldn't recognize that as a Word document. But a search engine recognizes the correct file type using the information inside the binary format itself, not the file extension. So a mismatched file extension will not affect a search engine's ability to correctly identify and apply the right parsing specification to a file.

Tip if the Iceberg, Example #2. An associated application view of a file – not just Word, but other “Office” formats like Excel, Access, PowerPoint, OneNote, email, PDF, etc. – can obscure certain metadata. Unless you know exactly where to click, you may not even know the metadata is there. But all metadata is fully accessible in the binary format. So a search for *secret Nebraska testing site* would retrieve that in metadata even if you wouldn't have spotted it in an associated application view of the same file.

Article contributed
by dtSearch®



A search engine like dtSearch® has to pre-process millions and sometimes billions of files to enable instant search either on an individual basis or on a multithreaded concurrent search basis across terabytes of data.

Tip if the Iceberg, Example #3. In addition to standalone files, data can assume a recursively embedded nested structure. For example, you can have an email with a ZIP or RAR attachment and inside that attachment is a PowerPoint and buried in the PowerPoint is an Excel spreadsheet. While even the PowerPoint display might obscure the full contents of the Excel spreadsheet, the whole spreadsheet would be readily apparent to the search engine when it automatically unpacks the email and its attachments using the binary formats.

Tip if the Iceberg, Example #4. Fuzzy searching is another way to locate data that might otherwise slip past undetected. Fuzzy searching works along with the other search options to find *misspellings*. So if I mistyped the word *misspelling* in an email as *misspelljng*, a search for misspelling with a small level of fuzziness on would still find that.

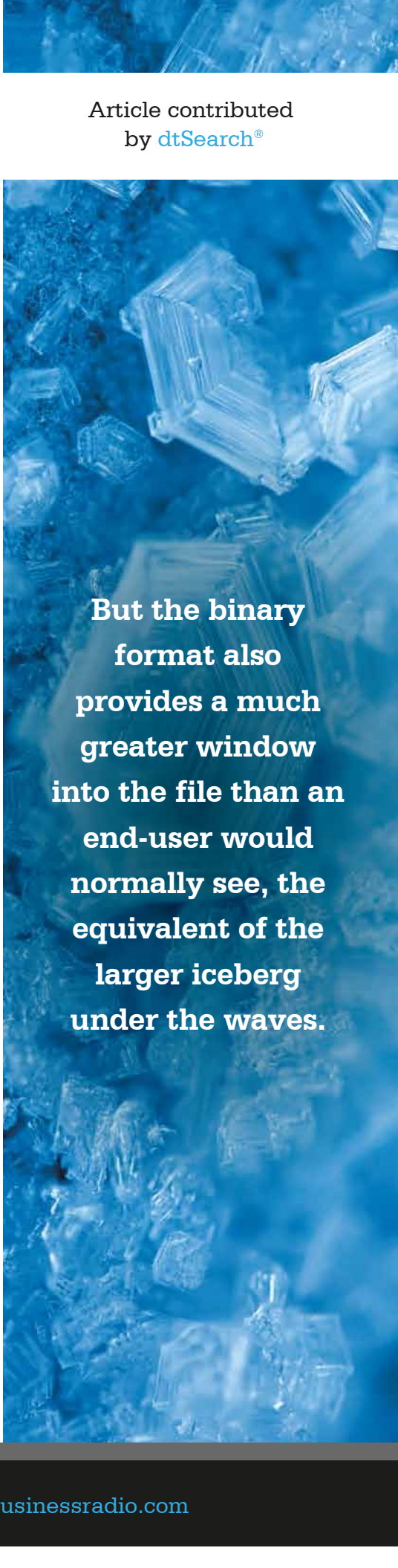
Tip if the Iceberg, Example #5. The classic example of hidden text is text that blends in with the background color inside a file's associated application, so white text against a white background, aqua text against an aqua background, etc. While this type of text is by design hard to spot in an associated application, it is as readily apparent as any other text inside the binary format

Tip if the Iceberg, Example #6. Most PDFs are text-based, letting you see words and copy and paste them. But sometimes what looks like words on a PDF page are just images of words, with no underlying text at all. When you try to copy and paste words out of such a PDF, you get nothing. A search engine can flag these "image only" PDFs letting you know that you need to run them through an OCR application like Adobe Acrobat to turn them into "searchable image" PDFs. Actually, this process is quite cool, because a "searchable image" PDF can save the whole underlying picture of the earlier "image only" PDF. But it also adds the OCR'ed words as an additional layer beneath the image. And it is these words that are now full-text searchable.

Tip if the Iceberg, Example #7. Along with "Office" files and emails, a search engine can also work with web-based formats stored as ordinary data in the file system. And it can work with cloud storage files visible through the Windows file system but stored remotely, like OneDrive or DropBox files as well as files synched through SharePoint. The search engine can index and search that type of data just like normal file data in the Windows file system. Depending on how you look at it, these cloud files can inhabit the tip of the iceberg or the larger submerged mass of the iceberg.

About dtSearch. dtSearch has enterprise and developer products that run "on premises" or on cloud platforms to instantly search terabytes of "Office" files, PDFs, emails along with nested attachments, databases and online data. Because dtSearch can instantly search terabytes with over 25 different search features, 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

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



**But the binary
format also
provides a much
greater window
into the file than an
end-user would
normally see, the
equivalent of the
larger iceberg
under the waves.**